Umeå University
Department of Psychology
Dissertation (Thesis/D-uppsats) Spring Term 2005

Format Dependence in Calibration and the Relationship between Overconfidence and Other Cognitive Measures

Stina Jonsson

Supervisor: Anna-Carin Olsson and Patrik Hansson

Format Dependence in Calibration and the Relationship between Overconfidence and Other Cognitive Measures

Stina Jonsson

Using a between-group design, the study further confirms the format dependence between the two probability assessment formats, *interval production* and *interval evaluation* found in earlier research (Winman et al., 2004) where *interval production* generated more overconfidence. In line with the hypotheses, working memory capacity (WMC) and fluid intelligence (Gf), correlated negatively with overconfidence in *interval production* but not in *interval evaluation.* According to *The Naïve Sampling Model* (NSM) proposed by Juslin et al. (2004), this is due to a psychological difference between assessing confidence as a probability (*interval evaluation*) and expressing confidence by producing intervals. Contrary to the hypothesis, Gf and not WM, had the higher correlation with overconfidence. A hierarchical multiple regression analysis disclosed the model to be significant, but the individual β weights were not. Episodic memory (EM) did not correlate at all with overconfidence.

In our everyday life we continuously gather information about our surroundings. For instance, we tend to get a rough idea about the distances between cities by travelling, and through media and education etc. When people are asked to estimate an unknown quantity, for example the distance between two cities, they will claim to be 100% sure that the distance lies between x and y km, where in fact people are quite often wrong. On the other hand, if you give them an interval, say the distance lies between 40 and 55 km, they can rather accurately estimate how sure they are that this statement is true (for example, they might say that they are 80% certain that the statement is true).

Calibration refers to the agreement between the actual state of affairs of an event or events and individuals' assessed belief of the event or events. If a person is asked 100 questions about unknown quantities (For example, *does the population of Singapore exceed 5 million?*) and after each such question is asked to assess how confident (.00-1.00) they are that their answer is correct, the person might have a mean confidence level of say .75, indicating that that the person is on average 75% confident, but only have given about 50 % correct answers. This phenomenon is called the *overconfidence phenomenon* and occurs when the mean confidence judgement is higher than the relative frequency of correct answers. Conversely, underconfidence occurs when the mean confidence judgement is lower than the relative frequency of correct answers. Many studies show that people are systematically overconfident about the accuracy of their knowledge and their judgement (e.g., Klayman, Soll, González-Vallejo, & Barlas, 1999; Koriat, Lichtenstein, & Fischhoff, 1980; Lichtenstein, Fischhoff, &Phillips, 1982). To produce confidence judgements that are *realistic* or *calibrated*, the mean confidence

level should, in the long run, equal the relative frequency of correct alternatives chosen (Lichtenstein et al., 1982; Yates, 1990). A person that is .75 confident should have 75 correct answers out of 100. Early research indicated that overconfidence varied considerably depending on the difficulty of the judgement task (Lichtenstein et al., 1982; Yates, 1990). Results indicated that people were very overconfident, that they answered correctly, when asked hard questions. Conversely, people displayed almost no overconfidence or even underconfidence when judging easy tasks. This effect is called the *hard-easy effect*. However, in a meta-analysis, Juslin, Winman, and Olsson (2000) argued that when non-representative selection of items and regression effect are controlled for, there is little evidence for this effect.

Format dependence

Format dependence in confidence judgements has been proven to be a robust phenomenon (e.g. Klayman et al., 1999; Winman, Hansson, & Juslin, 2004) Subjective probability distribution is a probability distribution that represents an individual's belief about the range and likelihood of values for a quantity. There are several ways to obtain a person's subjective probability distribution for unknown quantities (e.g. the population of Singapore). See Figure 1 for examples of the four assessment formats described below.

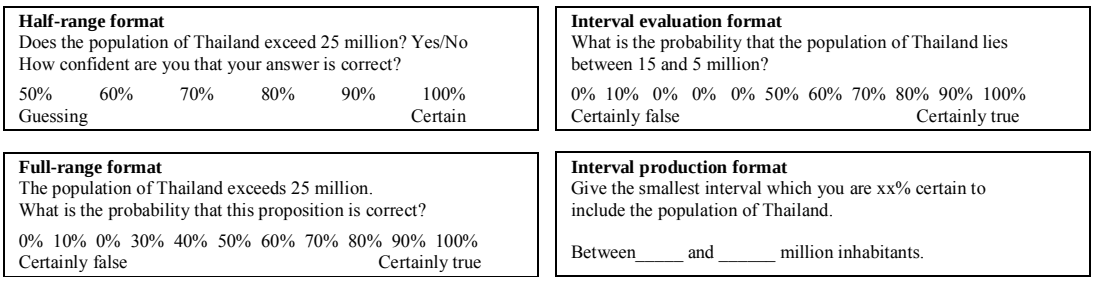

Figure 1. Format dependence. Examples of four assessment formats.

The four formats should formally produce the same result when the same stimulus material is used. This is however not the case. With the *half-range format*, the person first decides whether or not a statement is true (*the population of Singapore exceeds 5 millions; (a) yes, (b) no*) and then makes an assessment of the probability that this answer is correct on a scale between .5 (guessing) and 1 (certain). With the *full-range format*, the person directly estimates the probability that the statement is true on a scale between 0 (certainly false) and 1 (certainly true). These two formats produce very little or no overconfidence when statistical effects and biases in task selection are controlled for (Juslin et al., 2000).

With *interval production*, the person is given a probability and is asked to produce an interval around their best guess for the quantity reflecting the probability level. This format often yields extreme overconfidence (Klayman et al., 1999; Winman et al., 2004). For example, 90% confidence intervals should include approximately 90%

of the true values but often include as little as 45% of the true values (Klayman et al., 1999). The two first formats however, only involve assessing one fractile of the subjective probability distribution (e.g., What is the probability that the population of Singapore *exceeds* 5 million?) whereas *interval production* involves assessing two fractiles. To control for this possible confounding, Winman et al. (2004) studied *interval evaluation* which, like *interval production*, involves the assessment of two fractiles of the subjective probability distribution. In *interval evaluation* people are asked to assess the probability that the target quantity falls within a given interval. Winman et al. (2004) demonstrated that the *interval production* format generates significantly more overconfidence than the *interval evaluation* format. The effect is even seen when individuals produce intervals and later evaluate the same interval they have produced themselves.

How does overconfidence arise?

A cognitive bias (Kahneman, Slovic, & Tversky, 1982) has long been the dominating explanation for the overconfidence phenomenon. It suggests that the information processes that the judgements are based upon are biased. The overconfidence earlier found in the half-range format was explained by the tendency to retrieve information from memory and focus on information that supports rather than contradicts our hypothesis (Korivat et al., 1980). Under the same paradigm, the *interval production* format results in overconfidence due to the anchoring-and-adjustment heuristic by which people begin with a starting value (e.g. their best guess about the quantity) from which they insufficiently adjust their estimates (Tversky, & Kahneman, 1974). In recent years, these theories have failed to receive support in several studies (Block & Harper, 1991; Juslin, Wennerholm, & Olsson, 1999) and there has been an attempt to explain these biases in human judgement, not as a consequence of biased cognitive processes and an individual's motives, but as a result of biases in the input of information (Fiedler, 2000; Fiedler & Juslin, in press).

The Naïve Sampling Model (NSM) proposed by Juslin et al. (2004) maintains that there is a psychological difference between assessing confidence as a probability (*interval evaluation*) and expressing confidence by producing intervals. People are seen as fundamentally intuitive statisticians (Peterson & Beach, 1967). But people also have a tendency to treat sample properties as direct estimators of population properties (Fiedler, 2000; Fiedler & Juslin in press). Consequently, man is seen as a *naïve* intuitive statistician. A main assumption of the NSM, contrary to the cognitive bias paradigm, is that the cognitive processes that operate on the available information are not biased, but are basically in accordance with normative principles of reasoning and logic. Studies indicate that people are aware that predictions based on smaller samples are *not as reliable as predictions based on a lager sample. People also make inferences based on a sample that vary to a large extent, and produce very unsure predictions compared to predictions made from samples that only vary a little. (Kareev, Arnon, & Horwitz-Zeliger, 2002).*

Overconfidence still arises, mainly due to two types of naivety. First of all, the assumption that samples they come across are random and representative of the environment, which is not often the case (Fiedler, 2000). For example, if the distribution of the population figures of Asia that the person *knows* systematically deviates from the distribution of all population figures of *all* Asian countries, the result is overconfidence in both *interval production* and *interval evaluation*. Media coverage, personal interests, travelling etc. may be the cause of the discrepancy between the objective and the subjective distributions (See Hansson, Juslin, & Winman 2005 for a more detailed description). Secondly, people are naïve in that they treat samples as unbiased estimators of the population. In *interval production*, the sample dispersion is used as an estimator of population dispersion. This however, is a biased estimator because statistically the average dispersion of the sample will always be less than the dispersion of the population even in the long-run. The sample always varies less than the population. People generally fail to correct for this bias (by multiplying the sample dispersion with n/n-1 to get an unbiased estimator) (Kareev et al., 2002) which results in too narrow intervals and consequently overconfidence. Theoretically, overconfidence should increase with smaller samples and should almost disappear with very large or infinite samples. In *interval evaluation*, the sample proportion is used to estimate the population proportion. In contrast, the sample proportion is an unbiased estimator because the average sample proportion will eventually equal the population proportion. This explains why *interval evaluation* does not lead to extreme overconfidence. The NSM predicts that working memory capacity should correlate negatively with overconfidence if the sample size used for these inferences is constrained by working memory capacity.

Working memory and fluid intelligence

Working memory (WM) temporarily holds limited information accessible so that cognition can take place. Miller (1956) suggested that the human working memory capacity (WMC) was limited to 7 ± 2 items. He also claimed that the efficiency of storage capacity could be increased by the use of intelligent grouping or "chunking". Cowan (2001) defining the term "chunk" as two or more concepts that have strong associations to one another and much weaker associations to other chunks, presented in his article, considerable evidence suggesting that WMC is limited to three to five chunks. Applying this theory of a WMC of 4 ± 1, we find that when the dispersion of a population is statistically estimated from a sample with sample size n=4, without correcting for the bias (1/(n-1)), the result strongly resembles the result attained when overconfidence is actually measured in subjects producing intervals. To test if long-term memory affects overconfidence, Hansson (2005) demonstrated that overconfidence can not be cured even with extensive training. The amount of items learned increased significantly with training, but this had little effect on overconfidence in *interval production*. This indicates that long-term memory does not affect overconfidence, which is in line with NSM predictions.

In a series of experiments, Kareev et al. (2002) showed that people with less WMC (digit-span of less than 6) predicted less variability after being exposed to a sample of a population, than people with a WMC exceeding a digit-span of 6. This indicates that WMC is connected to overconfidence since overconfidence is highly dependent on the perceived variability of the sample. WMC have also been found to correlate negatively with subadditivity (when the sum of subjective probability judgement exceeds 100%), a phenomenon closely related to overconfidence (Dougherty, Hunter, 2003).

Many researchers have maintained that working memory and intelligence are identical or very closely related constructs (see Ackerman et al., 2005). In a meta-analysis, Ackerman et al. (2005) investigated WMC and its relation to general intelligence (g) and general fluid intelligence (Gf). They found a correlation between Gf and WMC of about $r = .48$ which would imply less than 25% shared variance. As a direct reply to this article, Oberauer, Schulze, Wilhem, and Süβ (2005) reanalysed the data and claimed that Ackerman, Beier, and Boyle (2005) underestimated the shared variance due to several methodological shortcomings and biases. They found an estimated correlation of $r = .85$ which point to a shared variance of 72%. Kane, Hambrick, and Conway (2005) also reanalysed the data (particularly considering the outcomes of latent-variable studies) and found a strong correlation (median .72) which would indicate a shared variance of about 50%. It is obvious, that there is no easy answer to the question about the relation of WMC and Gf. Both WMC and Gf will be measured in this study and are expected to correlate negatively with overconfidence. Episodic memory (EM) will also be measured in this study and is not expected to correlate with overconfidence to any extensive degree. If this holds true, it would give further strength to the proposition that long-term memory is not related to overconfidence.

Aims of the study

The aim of this study is in part to replicate a study by Winman et al. (2004) where the format dependence effect for *interval production* and *interval evaluation* was tested for the first time, and also try to further verify the strong effect found. If it can be proven that overconfidence is largely due to assessment formats, this knowledge could be used in applied settings to reduce overconfidence. The study also investigates the association between overconfidence and other cognitive measures which has not been done extensively in earlier studies. This will be done under controlled learning, similar to what was done by Hansson et al. (2005) where a strong negative correlation between working memory capacity and overconfidence was found.

The hypotheses of this study are: (a) the *interval production* format will yield more overconfidence than *interval evaluation* (format dependence), (b) working memory capacity (WMC) will be negatively correlated to overconfidence in *interval production* but not in *interval evaluation* (c) fluid intelligence (Gf) should correlate negatively with overconfidence but to a lesser extent than WMC (d) episodic memory

is expected to correlate negatively with overconfidence, to a lesser extent than WMC and Gf or not correlate at all.

Method

Participants

122 volunteers studying at Umeå University (n = 68) and Uppsala University (n = 54) in Sweden (68 women and 54 men, mean age = 24.8, *sd* = 3.96), participated in the experiment. The participants were guaranteed a basic amount of money (250 SEK) and to increase motivation they were told that they could receive a bonus (maximum 100 SEK) for performing well on the test. The bonus was based on the correlation between the point estimates given in the test phase (on the quarterly income of the 116 companies that appeared in the training phase) and the true values (See *Material and apparatus* and *Design and procedure* for further explanation). The subjects were divided into one interval production group and one interval evaluation group (60 in one group and 61 in the other).

Material and apparatus

The experiment was carried out on a PC. Each participant was tested individually in a laboratory session lasting between 3 and 4 h. Background information about participants' age and gender was collected.

Confidence judgement in interval evaluation and interval production. One hundred and thirty-six population figures for countries from five continents (Asia, Europe, North America, South America and Australia/Oceania) listed in the United Nations database (2002) were used as criterion. One hundred and thirty-six real company names were randomly assigned to a country in the database and thereby also classified as belonging to one of five fictitious regions (A-E) corresponding to the five continents. A new and independent such random assignment was preformed for each participant. The target variable was described as the quarterly income of the company (in millions). Over- and underconfidence (calibration) was measured by subtracting the in-range proportion from the mean subjective probability. Overconfidence was defined by a positive score, underconfidence by a negative score, and a zero score is a perfect calibration.

Working memory capacity (WMC) was tested by a digit-span test which in turn consisted of three subtests (a) a passive repeat back test (random numbers between 0 and 9) where the participants were required to repeat the numbers in the same order as they had been presented by typing the numbers on the dashboard, (b) a test requiring the participants to repeat numbers back in ascending order after viewing all the numbers in the digit-span, and (c) a test requiring the participants to repeat back numbers in ascending order after viewing the numbers in the digit-span one at a time (one number per second). Each digit-span was presented for 5 seconds and the

participants had 15 seconds to type the digit-span. All participants were exposed to identical digit-spans. If the participants correctly produced at least three out of four digit-spans (on each level) they proceeded to the next level (adding one number to the digit-span). WMC test (c) was more difficult than test (b) which in turn was more difficult then test (a). Working memory capacity was determined by the length of the longest sequence of digits of at least three out of four could be recalled correctly. The different WMC tests (a, b, and c) were *Z*-transformed to give equal weight and one measure of WMC was created.

Episodic memory (EM) was tested by first having the participants learn word-pairs during a training phase. Each word-pair was presented on the computer screen for two seconds. During a test phase the participants were asked to recall the second word of the word-pairs after they were given the first word as a recall cue. Half of the word-pairs were associated (ex. Silver-Gold) and the other half consisted of word-pairs that were not associated (ex. Bench-Mother). Episodic memory capacity was measured by how many of the words they could correctly recall after the participants had been given the cue recall.

Fluid intelligence (Gf) was assessed using Raven's Advanced Progressive Matrices (APM; Raven, Court, & Raven, 1988). The test is a widely used, nonverbal test of analytic intelligence. It includes 12 training items (no time limit) and another 36 which the participants have 40 minutes to complete all or as many items as possible. It requires participants to select the correct shape (among 8 alternatives) to complete geometric patterns in a matrix (3x3) where the bottom right shape is missing. A participant's score is the total number of correct solutions.

Design and procedure

In the training phase participants were randomly exposed to 116 of the 136 different companies, two times (232 in all). The remaining 20 company income figures were only presented in the test phase and were therefore *new items* for the participants. This was done to ensure that any judgement about these 20 items most have taken place as an inference. They were required to guess the quarterly income and received feedback on the correct values. The test phase for the *interval production* condition preceded as follows: An initial point estimate for a randomly selected company was followed by a requirement to produce a .5, .8 or 1.0 (randomly selected) probability interval that the participants believed would include the true value with the stated probability. The following wording was used: *Produce the (smallest) interval within which you are 80% (probability .8) certain to include the company VERSATA Inc's (which belongs to Region B) income: Between ___ and ___ million.*

The test phase for the *interval evaluation* condition proceeded as follows: An initial point estimate about the quarterly income figure for a randomly selected company was followed by a requirement to assess the probability that a presented interval included the true value. The interval was centred on a value randomly drawn from the database of income figures (i.e. all of the 136 companies). The interval width equalled the point estimate provided by the participant. They were then

required to assess the probability that the true value lay within the interval in a full-range probability scale. The *Interval Evaluation* task had the following format: *The company VERSATA Inc (which belongs to Region B) has an income that lies between X and Y million/quarter. What is the probability that the statement above is correct? 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%.*

The participants in both groups then performed the three working memory tests, after which they had a 30 – 60 minute break. After the break the participants learned 24 word-pairs constituting the episodic memory training phase. Thereafter they completed the *Raven's*, and ended the experiment by performing the *episodic memory* test. See test schedule in Figure 2.

PART ONE
- (1) Company income *training* phase (2 x 116 items)
- (2) Company income *test* phase (116 items + 20 new items)
- (3) Working memory
  - (a) WM – see all at once – repeated back in identical order
  - (b) WM – see all at once – repeat back in ascending order
  - (c) WM – see one at a time – repeated back in ascending order

BREAK
- 30 – 60 minutes

PART TWO
- (4) Episodic memory *training* phase
- (5) Raven's Advanced Progressive Matrices
- (6) Episodic memory *test* phase

Figure 2. Test schedule.

Resultat

Significance level was set to $\alpha = .05$. The proportion correct recalled target values (indicated by point estimates) was .14 (*sd* = .13), indicating that the participants could accurately recall approximately 16 of the 116 quarterly company incomes. Proportion correct recalled target values did not correlate significantly with overconfidence in neither *interval production* ($r = -.175$, $p = .179$) nor in *interval evaluation* ($r = .044$, $p = .738$) for *not correct items*. The proportions correct in-range values (where participants produced an interval that did include the true value) were .43 for the .50 probability interval (95% CI = ±.02, *sd* = .50), .51 for the .80 probability interval (95% CI = ±.02, *sd* = .50) and for the 1.00 probability interval .66 (95% CI ≈ ±.02, *sd* = .48). The average probability given in *interval evaluation* was .36 and the proportion correct was .27 (95% CI = ±.01, *sd* = .45).

Figure 3 shows the total format dependence between the two conditions for *not correct items (*all items, excluding items participants could correctly render indicated by correct point estimate) $t(119) = 7.048$, $p = .000$. This also holds for *new items* (the 20 items the subjects had not seen in the training phase) $t(119) = 2.902$, $p = .004$. Seeing as both *not correct items* and *new items* displayed a significant difference, only the result from *not correct items* (will now be called overconfidence) will be

presented since *new items* only consists of 20 items which makes it a less reliable measure. Note that *new items* are included in *not correct items*.

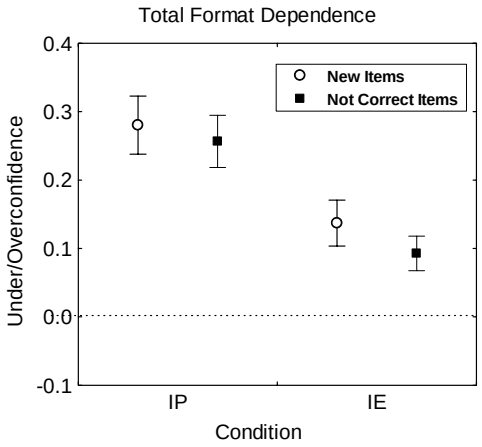

Figure 3. Total format dependence between *interval production* (IP) and *interval evaluation* (IE). Bars represent 95% CI for the mean overconfidence for *new items* and *not correct items*. The dotted horizontal line represents perfect calibration.

Table 1 presents Pearson's correlation coefficients between under/overconfidence, working memory capacity, Raven's (fluid intelligence, Gf), and episodic memory (EM) in the two conditions. As can be seen working memory capacity (WMC) was significantly negatively correlated with overconfidence in *interval production* whereas in *interval evaluation* there was a small negative but non-significant correlation between overconfidence and WMC. However, the difference between the correlation coefficients fail to reach statistical significance (p = .1429, one-tailed). Fluid intelligence (Gf), correlated significantly with overconfidence in *interval production* but not in *interval evaluation*. Episodic memory did not significantly correlate with overconfidence in either *interval production,* or in *interval evaluation*.

Table 1

Pearson r Correlations Between Under/overconfidence, Working Memory, Raven's (Fluid Intelligence), and Episodic Memory in Interval Production (Panel A) and Interval Evaluation (Panel B) (* = $p < .05$,** = $p < .01$).

A

Interval Production			
	Under/-overconfidence	Episodic memory	Raven Intelligence
Working memory	-.303*	.443**	.397**
Raven's intelligence	-.323*	.296*	
Episodic memory	-.153		

B

Interval Evaluation			
	Under/-overconfidence	Episodic memory	Raven Intelligence
Working memory	-.109	.245	.376**
Raven's intelligence	-.217	.127	
Episodic memory	.049		

A hierarchical multiple regression analysis with overconfidence in the *interval production* condition as the dependent variable and estimated working memory capacity(WMC), fluid intelligence (Gf), and episodic memory (EM) as independent variables was carried through. A model, only with WMC, had a R^2 = .096. When Raven's (Gf) is added to the model, the amount of explained variance increases to R^2 = .143 (change in *F*-value n.s. p = .081). When EM was added there was only a minimal change, R^2 = .144 (change in *F*-value n.s. p = .907). All models were significant including this model ($F(3,56) = 3.128$, p = . 033). Table 2 summarizes the result of the individual β weights, where none proved to be significant.

Table 2

β Weights, and T-values with Significance for the Linear Multiple Regression Model with Overconfidence in the Interval Production Condition as Dependent Variable and WM, Raven, and EM as Independent Variables. Zero-order and Semipartial Correlation Are Also Shown.

				Correlations	
	β	*t*	*p*	Zero-order	Semipartial
Working memory	-.221	-1.519	.134	-.309	-.188
Raven's intelligence	-.240	-1.762	.084	-.323	-.218
Episodic memory	.016	.117	.907	-.153	-.014

Because WMC and Gf intercorrelate to a relatively large degree, multicollinearity might be the cause of the non-significant β weights. To estimate the proportion of total variance in the dependent variable explained uniquely by the three independent variables, semipartial correlations were calculated and are presented in Table 2 along with zero-order correlations. This model was constructed according to the hypothesis (WM, Raven, EM) but seeing as Raven's and not WM displayed the highest correlation with overconfidence a model with the order Raven's, WM and EM, was tested which only marginally differed from the original model.

Table 3

β Weights, and T-values with Significance for the Linear Multiple Regression Model with Overconfidence in the Interval Production Condition as Dependent Variable and WM, Raven, EM and Interaction between WM and Raven as Independent Variables.

	β	*t*	*p*
Working memory	-1.352	-2.988	.004
Raven's intelligence	-2.495	-2.898	.005
Episodic memory	.028	.207	.837
Interaction WMC*Raven's	2.910	2.646	.011

As can be seen in Table 3, when the interaction between Gf (Raven's) and WMC was entered last in the hierarchical multiple regression, the β weights for the interaction was significant (Zero-order correlation r = .335 and semipartial

correlation r = .311), Gf (Raven's) and WMC also became significant while EM remained non-significant.

Testing the collected background variables, gender and age and their relation to overconfidence indicated that men were significantly less overconfident than women in *interval evaluation* $t(59) = 2.268$, $p = .027$ but not in *interval production* $t(58) = -.723$, $p = .473$.

There was a significant correlation between age and overconfidence in *interval production* $r = .264$, $p = .042$ while in *interval evaluation* there was virtually no correlation $r = -.002$, $p = .989$. The correlation between age and WMC was significant $r = -.198$, $p = .029$ but not the correlation between age and Gf $r = -.129$, $p = .158$.

Observe that there were no explicit hypotheses about the relationship between gender and under/overconfidence or age and under/overconfidence.

Discussion

As predicted by the *Naïve Sampling Model* (NSM), there was strong format dependence when the same subjective probability distributions were assessed. The *interval production* format elicited considerable overconfidence whereas the *interval evaluation* format only generated slight overconfidence. This result replicates the previous findings of format dependence between *interval production* and *interval evaluation* (Juslin et al., 1999: Klayman et al., 1999; Juslin et al., 2003). Previous research has mainly considered the format dependence between *interval production* and other assessment formats, for example half-range and full-range format where only one fractile of the subjective probability distribution is estimated (see Figure 1 for examples). The result also demonstrated that when completely new items are being used (to insure that inference has taken place) there is format dependence. A well calibrated person that indicates that she or he is on average 80% confident should provide approximately 80% correct answers. We see that when using the *interval production* format, the participants produced intervals that contained the true value 51% of the times when asked to produce 80% intervals. This shows that when this format is used people are far from well calibrated. People in *interval evaluation* were also overconfident but not to the same extent. Keep in mind that in both conditions learning was controlled and identical between the two conditions. Despite accurate feedback in the training phase, participants in the *interval production* condition were not able to calibrate their judgement.

Investigating the relationships between calibration (overconfidence) and other cognitive measures has not been done extensively in previous research. But doing so might help us understand, and predict the overconfidence phenomenon and in applied settings, cure or reduce it. The results were consistent with the hypothesis that working memory capacity (WMC) is negatively correlated to overconfidence in *interval production* but not in *interval evaluation*. Participants who scored high on the WMC tasks were less overconfident which indicates that they used a larger

sample size to make inferences about the population dispersion. The absence of a significant correlation between WMC and overconfidence in the *interval evaluation* format suggests that even people with relatively limited WMC could fairly accurately estimate the population proportion from the sample proportion. According to NSM, this is due to the fact that sample proportion is an unbiased estimator of the population proportion so smaller sample size does not result in large biases even when small samples are used.

The results also reveal that fluid intelligence (Gf), here measured by Raven's advanced progressive matrices, accounted for a significant amount of variance in overconfidence in *interval production* but not in *interval evaluation*.

Both WMC and Gf accounted for a significant amount of variance in overconfidence in the *interval production* condition. But contrary to the hypotheses, Gf had a higher correlation with overconfidence (r = -.323) than WMC (r = -.303) even if the difference was not significant. This result should be interpreted with caution however, because the correlations were similar in size and the correlation between working memory capacity and intelligence were high (r = .397). In fact it was higher then any of the independent variables and the dependent variable. The semipartial correlation showed a correlation of r = -.188 for WMC, r = -.218 for fluid intelligence, and r = -.014 for EM estimating the specific variance for each independent variable when the other two are held constant. None of the independent variables remained significant.

It seems as though fluid intelligence plays a greater roll in overconfidence than first hypothesized in this study. It is plausible that people with fairly large WMC also have a high fluid intelligence since they are at least similar, if not identical constructs. These persons might in addition to be able to use large samples to make inferences about populations, also posses other skills to facilitate the mental operation necessary to make these inferences. The fact that the interaction between WMC and Gf, still remained significant when the variation due to WMC, EM and Raven's (Gf) was controlled for, suggests that the relation between WMC and Gf in relation to overconfidence is a complicated one and demands further investigation.

Episodic memory (EM) did correlate negatively, but not significantly with overconfidence in both of the conditions. In *interval production*, the regression analysis revealed that when WMC and Gf are held constant there is virtually none of the variance in EM that explains any of the variance in overconfidence. The number of quarterly company incomes participants could correctly recall did also not correlate with overconfidence in any of the conditions. This suggests that long-term memory does not play a roll in overconfidence. It is not a good memory in general that leads to good calibration. These findings are consistent with earlier research (Hansson et al. 2005). It is also consistent with studies conducted outside the laboratory by Russo and Schoemaker (1992), finding that experts are not less overconfident even though they posses more experience and information.

The analysis of the collected background variables showed that there was no difference between the genders when the *interval production* format was used, but in *interval evaluation* men were less overconfident than women. These results are the

total opposite of common beliefs about the tendencies of overconfidence of the genders. The women in this study were more overconfident than the men. Further studies must be done to investigate and verify or invalidate these findings since this study was not specifically designed to answer this question. Nor did it consider any theoretical indications that led to any hypothesis that could explain these results.

The second background variable, age, was significant positively correlated with overconfidence in *interval production* but not in *interval evaluation*. This result is in line with previous research where older subject were more overconfident compared to younger subjects (Crawford and Stankov, 1996). This can be interpreted as support for the NSM because age is widely assumed to correlate negatively with working memory. NSM predicts that these types of cognitive abilities will correlate negatively with overconfidence in *interval production* but not in *interval evaluation*. NSM therefore, indirectly predicts that age should correlate with overconfidence in *interval production* but should not correlate with overconfidence in *interval evaluation*, which is also the case in this study. There was a significant negative correlation between WMC and age but not between age and Gf.

One possible criticism of the present study is that some participant only managed to store a very small sub-sample of company income figures in their long-term memory (indicated by the proportion learned). 17 participants could correctly recall 4 or less company incomes, which suggest that they did not store enough items in long-term memory to maximize their working memory capacities, when asked to make an inference about an unknown company's income. Even though individuals remembered more than four company incomes, they might not have had enough material to draw from to maximize their working memory capacity for all of the respective regions (A-E). These persons were perhaps not limited by their working memory at all. It is also hard to ensure (in this study) that even participants who can recall enough company incomes to be able to maximize their WMC, can remember to which region the specific company income figure belonged to. It is also concievable that participants recalled some company income figures without correctly remembering to which company (or which region) it belonged. This is espesially true for figures that are extremely low or extremely high. This however would cause less overconfidence since it would contribute to a higher perceived variance.

Another criticism is that the three tests used to estimate working memory capacity (WMC) did not measure WMC but short-term memory (STM) or a mixture of WMC and STM. STM is a more passive memory often referred to as a storage space, whereas WM is more of an operating space where the information is being actively used. Even though these two types of memories are highly related, a better test of WMC would certainly increase the validity of the study.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131, 30-60.

Block, R. A. and Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes*, 49, 188-207.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioural and Brain Sciences*, 24, 87-114.

Crawford, J. D., & Stankov, L. (1996). Age differences in the realism of confidence judgements: A calibration study using tests of fluid and crystallized intelligence. Learning and Individual Differences, 8, 83-103.

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31, 968-882.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. Psychological Review, 107, 659-676.

Fiedler, K., & Juslin, P. (in press). Taking the interface between mind and environment seriously. In K. Fiedler, & Juslin, P. (Eds.), *Information sampling as a key to understand adaptive cognition*. New York: Cambridge University Press.

Hansson, P., Juslin, P., & Winman, A. (2005). *Sampling in confidence judgment: Constraints on sample size and biased input distributions*. Unpublished manuscript, Umeå University.

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(2) 384-396.

Juslin, P., Wennerholm, P., & Olsson, H. (1999). Format dependence in subjective probability calibration. Journal of Experimental Psychology: Learning, Memory and Cognition, 28, 1038-1052.

Justlin, P., Winman, A., & Hansson, P. (2004). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. Unpublished manuscript.

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

Kane, M. J., Hambrick, D. Z., & Conway, A. R. A. (2005). Working memory capacity and fluid intelligence are strongly related constructs: Comments on Ackerman, Beier, & Boyle, (2005). *Psychological Bulletin,* 131, 66-71.

Kareev, Y., Arnon, S., & Horwitz-Zeliger, R. (2002). On the misperception of variability. *Journal of Experimental Psychology: General*, 131, 287-297.

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.

Korivat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 17-118.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of subjective probabilities. The state of art up to 1980. In D. Kahneman, P. Slovic, & Tversky (Eds.), *Judgements under uncertainty: Heuristics and biases*, (pp. 306-334). New York: Cambridge University Press.

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

Oberauer, K., Schulze, R., Wilhem, O., & Süβ, H.-M. (2005). Working memory and intelligence – Their correlation and their relation: Comments on Ackerman, Beier, & Boyle, (2005). *Psychological Bulletin,* 131, 61-65.

Peterson, C. R., & Beach, L. R. (1967). Man as initiative statistician. *Psychological Bulletin*, 68, 29-46.

Raven, J. C, Court, J. H., & Raven, J. (1988). *Manual for Raven's Progressive Matrices and Vocabulary Scales: Section 4*. London: Oxford Psychologists Press.

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

Winman, A., Hansson P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 1167-1175.

Yates, J. F. (1990). *Judgement and decision making*. Englewood Cliffs, NJ: Prentice Hall.