Separating Accuracy from Forecast Certainty: a Modified Miscalibration Measure*

Doron Sonsino, Yaron Lahav, Amir Levkowitz

Abstract: Interval forecasting tasks are commonly used to test for forecast-overconfidence. Pointing at deficiencies of the methodology, we advance a modified assignment, where subjects provide point predictions and assess the likelihood of return falling within small intervals around their estimates. The difference between the subjective likelihood assessments and the realized hit rates is advanced as an improved forecast-overprecision measure. Over three incentivized studies, 163 of 195 participants overestimate their hit rates, and a closer look at the data illustrates that inaccuracy and excessive-certainty act as distinct sources of overprecision. Applications where the adapted task may prove more powerful than standard interval forecasting are discussed.

Keywords: Interval forecasting, miscalibration, overconfidence, forecast-accuracy, trading

JEL classifications: C9, G4, D8

*Forthcoming in Behavioral Finance: A Novel Approach; a collection of papers edited by Itzhak Venetia. Doron Sonsino (corresponding author: sonsino.doron@gmail.com) is at the College of Law and Business, Ramat Gan, Israel and adjunct at the Economics Department of Ben-Gurion University. Yaron Lahav is from the Guilford Glazer Faculty of Business and Management (GGFBM) at Ben-Gurion University and Amir Levkowitz is a doctoral student at GGFBM. Studies 1 and 2 were conducted while Doron Sonsino was at the College of Management Academic Studies (COMAS) in Israel and funded by the research authority of COMAS. We thank participants at the Warwick University 2016 FUR conference, the Tucson ESA 2016 meetings, the SPUDM 2017 Technion conference, the IAREP 2017 COMAS meetings, the Vienna ESA 2017 meetings and seminars at the Gothenburg Research Institute, Ben-Gurion university and the Technion for helpful comments. We particularly thank David Budescu and Nigel Harvey for directing us to related psychology research on interval evaluation.

1. Introduction

The false belief in the accuracy of subjective assessments is considered an especially persistent form of judgmental overconfidence (Alpert and Raiffa, 1982; Klayman et al., 1999; Moore et al., 2015). Moore and Healy (2008) term the excessive certainty in beliefs *overprecision*, separating it from other facets of overconfidence such as *overestimation* of individual abilities or *overplacement* relatively to others. Diverse studies propose that overprecision contaminates the decision of traders (Daniel and Hirshleifer, 2015), executives (Malmendier and Taylor, 2015) and entrepreneurs (Astebro et al., 2014). Daniel et al., (1998), for a specific example, develop a model where overprecision explains the persistence of non-profitable trading, generating predictability in stock returns.

Finance experiments and surveys commonly use confidence interval tasks to test for forecast-overprecision (for diverse examples see Glaser and Weber, 2007; Oberlechner and Osler, 2012; Ben-David et al., 2013; Sonsino and Regev, 2013; Fellner-Röhling and Krügel, 2014; Broihanne et al., 2014; Merkle, 2017; Grosshans and Zeisberger, 2018). The confidence level is exogenously provided and the forecaster submits lower and upper bounds for the target return. The participants are faced with several such interval-production assignments and the collection of intervals is utilized to derive the forecast-overconfidence measures.[1] While the results of these studies uniformly support the overconfidence hypothesis, the overprecision metrics derived from forecast intervals show inconclusive statistically-weak results in empirical tests of the hypothesized overprecision-trading links (Glaser and Weber, 2007; Fellner-Röhling and Krügel, 2014; Broihanne et al., 2014; Merkle, 2017).

The current chapter points at deficiencies that arise with using interval forecasting tasks to derive overprecision measures, advancing a more direct method that avoids the obstacles. Essentially, the alternative approach diverts from *interval-production* to *interval-evaluation* (Winman et al., 2004). Subjects are asked to provide a point estimate (F) for the target return first, and then assess the likelihood of return falling within a fixed length interval ($F \pm \delta$) around their point forecast. Essentially, the subjects assess the likelihood of exhibiting smaller than δ errors when forecasting the returns on designated stocks. Overprecision is

[1] The terms *forecast-overconfidence*, *forecast-overprecision* and *forecast-miscalibration* are used interchangeably henceforth.

supported if subjects overestimate the hit rates of the small-error intervals around their point predictions.

We report the results of three incentivized experiments, applying the altered approach on N=195 competent subjects. Utilizing the modified task to illustrate that inaccuracy and excessive-certainty complementarily contribute to forecast-overconfidence, we argue that the modified task may prove more powerful than standard interval forecasting in testing the hypothesized overprecision-trading links.

Figure 1: The Standard Interval Forecasting Task (SIFT)

-Submit a median prediction for the return on *StockName* in April-June, 2018 (the second quarter of 2018) _____

- With probability 95%, I believe that the return on *StockName* in the second quarter of 2018 will be smaller than _____

- With probability 95%, I believe that the return on *StockName* in the second quarter of 2018 will be larger than _____

2. Motivating discussion

Figure 1 illustrates the standard interval forecasting task (SIFT) as employed in diverse overconfidence studies (similar formats are used, for instance, in Glaser and Weber, 2007; Oberlechner and Osler, 2012; Ben-David et al., 2013). The forecaster provides a point prediction and 95% confidence limits for some future return, so that the interval between the limits represents a 90% confidence interval for the target return. In typical surveys and experiments, the forecaster is faced with a collection of such forecasting assignments. Calibration is tested when the returns R are realized, and the term HIT is used for cases where *r* falls within the submitted interval. The difference between the exogenous confidence level and the actual hit rate, addressed as the *miscalibration score*, is utilized as an applicable forecast-overprecision metric. If the hit rate of 90% confidence intervals, for instance, is only 50% then the forecaster appears to overestimate the precision of the submitted interval forecasts by 40%, in line with forecast-overprecision.

The hit rates of forecast intervals, however, jointly depend on the accuracy of the intervals and their length. High miscalibration scores may emerge when the forecasts are inaccurate or when the intervals are too narrow (see Klayman et al., 1999; Juslin et al., 2007 for related judgmental psychology discussions).[2] If the target return, for a formal example, is normally distributed around μ_t with standard deviation (volatility) σ_t, then underestimation of the volatility by 25% would decrease the hit rate to 78%, while 50% discount of σ_t would cut the hit rate to 59%. Hit rates of 78% (59%), however, alternatively emerge when the point estimate is 0.84 (1.4) standard deviations from the true mean. Indeed, in McKenzie et al. (2008), IT experts provide more accurate, but shorter, confidence intervals for quantities related to their expertise. The contradicting effects cancel out and the experts' miscalibration scores are similar to those of non-expert students. Forecasting studies more generally show that domain-knowledge and expertise affect the accuracy of forecasts and the hit rates of prediction intervals (Lawrence et al., 2006; Hyndman and Athanasopoulos, 2018). The confounding knowledge and accuracy effects cast doubt on using miscalibration scores as meaningful measures of excessive forecast-certainty.

As an alternative approach, recent studies use the confidence intervals to estimate the *perceived volatility* of the forecasted return. The perceived volatility estimates are contrasted with empirical benchmarks and the difference (empirical volatility minus perceived volatility) is adopted as a direct measure of overprecision. This alternative method, however, suffers from the weakness of resting on explicit assumptions regarding the stochastic process of stock returns. Indeed, the papers employing this methodology use a wide range of models to approximate the perceived and empirical volatilities. Graham and Harvey (2001), Glaser et al. (2013) and Merkle (2017) use the Keefer and Bodily (1983) approximations to derive the standard deviation of the forecasted return. Oberlechner and Osler (2012) assume GARCH processes and use option implied volatilities to derive the empirical benchmarks. Sonsino and Regev (2013) compare the length of the forecast intervals with the realized spreads in recent histories, illustrating that the extent of overprecision may strongly vary with the length of history-window selected for the comparison. The sensitivity to statistical assumptions again

[2] We separate between *financial forecasting studies* where the prediction targets are future returns or prices and *judgmental psychology studies* where subjects estimate hidden quantities such as the population of given cities. Interval-evaluation tasks were tested in few judgmental psychology studies (see Section 3), but we are not aware of financial forecasting studies employing the methodology.

raises the concern that the proposed overprecision measures cannot be used to effectively rank investors in terms of relative forecast-certainty. Based on the lengths of the intervals, an investor that submits a [10%, 30%] interval, would be classified as less confident than one that submits a [0%, 10%] prediction, but the ranking may overturn if the heteroskedasticity of return processes is taken into account.

On top of these problems, few preceding overconfidence studies point at nonsensical results, questioning the internal validity of the standard confidence interval assignment. Teigen and Jørgensen (2005; experiment 3) find that 90% confidence intervals are only marginally longer, exhibiting similar hit rates as 50% intervals. In Langnickel and Zeisberger (2016) the lengths and hit rates of 90%, 60% and 30% confidence intervals are almost identical. Added to the accuracy-length confound and the statistical problems in deriving volatility estimates, these paradoxical results motivate a search for alternative improved measures of forecast-overprecision.

3. The Forecast Accuracy Assessment Task (FAAT)

The modified task, illustrated in Figure 2, consists of three steps:

Step 1: The subject submits a median forecast F for the target return. The instructions explain that the median is a point estimate, positive or negative, such that the provider assigns 50% likelihoods to larger or smaller returns.

Step 2: The instructions guide the subject to construct a fixed length interval around F, adding and subtracting a given margin δ from the median.

Step 3: The subject provides a likelihood assessment CONF of the interval $[F - \delta,\ F + \delta]$, estimating the likelihood of return falling within the 2δ interval centered at F. If, for example, the median forecast is 7% and the margin δ is 5%, then the subject estimates the likelihood of the [2%, 12%] interval. Likelihood assessments can take any value between 0% and 100%.

Figure 2: The Forecast Accuracy Assessment Task (FAAT)

-Submit a median prediction for the return on *StockName* in April-June, 2018 (the second quarter of 2018) _____

-What, in your opinion, is the probability that the return on *StockName* in the second quarter of 2018 would fall in a range of plus or minus 5% from the median?

The next diagram may help you develop your estimate:

[lower bound_____median forecast _____upper bound]

Add 5% to the median prediction and fill in the "upper bound" box.
Subtract 5% from the median prediction and fill in the "lower bound" box.

Submit the probability you assign to quarterly April-June 2018 *StockName* return within the interval between the lower and upper bounds (i.e., in a range of plus or minus 5% from your median forecast): _____ (between 0% and 100%)

In the studies described next, the subjects are faced with a sequence of such *Forecast Accuracy Assessment Tasks* (FAATs). Again, the term HIT is used for cases where *r* falls within the given interval, and the difference between the average likelihood that the subject assigns to calibration ($\overline{CONF}$) and the average hit rate ($\overline{HIT}$) represents our modified overconfidence measure: OC=$\overline{CONF} - \overline{HIT}$. If the CONFs, for example, are 70%, 50%, 90%, 70% while the realized return falls within the interval in only 1 of the 4 cases, OC is 45%. The subject overestimates the hit rate by 45%, in line with the overprecision hypothesis.

While the overconfidence scores derived from the FAATs may appear essentially similar to the miscalibration scores derived from SIFTs, the competing measures are fundamentally different. In FAAT, a hit is recorded when the realized return (r) falls within the fixed length prediction interval; i.e., when $F - \delta \leq r \leq F + \delta$ or $r - \delta \leq F \leq r + \delta$. The hit rate therefore only depends on the accuracy of the median forecasts. As the overconfidence score is derived by subtracting the realized hit rate from the average likelihood that subjects assign to the

respective intervals, FAAT fixes the accuracy-length concern raised for the standard miscalibration measure. The impact of accuracy and confidence on miscalibration can be separately assessed. Subjects may classify as overconfident for showing low accuracy, exaggerated forecast certainty, or both (see sections 4-5 for examples). Such clean separation is impossible when miscalibration is measured using standard intervals.

In addition, FAAT improves on the standard interval forecasting in few methodological/ technical aspects:

- In FAAT, forecast-confidence is elicited in a familiar 0-100 percentile scale. Intuitively, probabilistic assessments are more natural and easier than quantile assessments. Abbas et al. (2008), for instance, compare two methods for approximating the beta distributions of random variables. The method that is based on likelihood assessments shows higher consistency rates than a parallel method based on quantile assessments. It is also ranked as easier and more popular amongst the subjects. Similar results emerge in Wallsten et al. (2016).[3]

- Assuming the experiment is incentivized by randomly picking one of the assignments as a *payment task* (Cubitt et al., 1998; Hey and Lee, 2005), the likelihood assessment of the FAAT is easier to incentivize compared to the quantile assessments of the standard task. A loss function for the α quantile of a distribution X, for example, is $|x - q| \cdot (\alpha \cdot 1_{\{x>q\}} + (1-\alpha) \cdot 1_{\{x \leq q\}})$, where q is the elicited α quantile and x is the realized X, while a quadratic scoring rule (QSR) for the likelihood p of event E is $p^2 \cdot 1_E + (1-p)^2 \cdot 1_{E^C}$ (Gneiting and Raftery, 2007). The quadratic score takes only two values, while the loss function for the quantile varies with x.

- FAAT avoids the statistical problems that arise with deriving perceived volatility estimates and comparing with empirical benchmarks. Overconfidence is measured directly with no need for exploring the (perceived) stochastic process of returns.

[3] In Abbas et al. (2008) and Wallsten et al. (2016) the probability-based methods also produce more accurate estimates of the true probability distributions, but this aspect of the results is contended in other studies (e.g., Bansal and Palley, 2017).

Methods similar to FAAT were utilized in few psychology overconfidence studies (see Murphy and Winkler, 1974 for an early example). A common result of these scarce studies is that miscalibration decreases in *interval-evaluation* compared to *interval-production*. In Winman et al. (2004), for example, the miscalibration rate decreases from 34% to 15%, when distinct subjects judge the likelihood of 90% intervals produced by their peers. Teigen and Jørgensen (2005; experiment 2) ask subjects to add and subtract 25% of their point assessments and estimate the likelihood of the resulting intervals; the confidence estimates and actual hit rates almost agree. Speirs-Bridge et al. (2010) alternatively advance a four-step procedure where experts submit a point prediction, lower and upper bounds, and then assess the likelihood of the resulting interval. Again, miscalibration significantly decreases compared to standard interval-production tasks. Given this consistent evidence, we utilize FAAT to test the robustness of overprecision, in addition to advancing it as an improved measure of forecast-overconfidence.

Finally, returning to the Moore and Healy (2008) taxonomy of overconfidence, note that as FAAT measures overprecision in terms of the distance between perceived and actual accuracy rates, OC may be classified a hybrid overprecision-overestimation measure. The tendency to overestimate one's skills or knowledge was witnessed in diverse economic studies (e.g., Bhandari and Deaves, 2006; Neyse et al., 2016; Bergu, 2019). Bhandari and Deaves (2006), for example, study a survey where Canadian pension plan members are presented with two multiple choice questions regarding the past performance of stocks and bonds and report how certain they are in each answer. On average, confidence exceeded the correct choice rate by 22%, showing that the survey participants overestimate their knowledge of past returns. The current FAAT adopts a comparable approach to test for forecast-overprecision.

4. Study 1

Method: The questionnaire of study 1 consisted of eight FAATs. The prediction stocks were members of TA25, the 25 leading stocks of the Tel-Aviv exchange, and the prediction period was the last quarter of 2016. We used distinct stocks, from different sectors, to decrease the correlation between the target returns.[4] In the first four assignments, the subjects did not receive

[4] Indeed, the realized quarterly returns ranged between -22% and +11%.

information except for the stock's name. The next four tasks were preceded by charts of the daily price trends and trading volumes in the first six months of 2016 (see Web Supplement A).[5] The questionnaire was distributed in-class to MBA students within four days, in early August 2016. The instructions were presented verbally and distributed in print, and the participation time was not effectively constrained. Surfing the Web and page turning were forbidden. To incentivize the FAATs we used random task selection combined with binarized scoring rules. The random task selection is employed to motivate subjects for independent work in each task (Cubitt et al., 1998; Hey and Lee, 2005; see Murad et al., 2016 for recent application). Binarized scoring rules are adopted to prohibit the bias that personal risk attitudes may impose on experimental responses (Hossain and Okui, 2013; Harrison et al., 2014). The probability of winning a fixed prize of 100 NIS (about 25 US$) decreased with the absolute prediction error (|F-r|) when the payment task was median forecasting. A standard QSR was used to derive the winning probability in the CONF assignments. The payment task was randomly drawn at the end of each session and used in January 2017 to determine eligibility for the 100 NIS payoff. The students were invited to supervise the process and keep record of the draws.[6] The next section reports the results for the N=72 students (60% males) that completed the eight FAATs with no errors.

Results: On average, the subjects assigned 75% likelihood to the fixed length intervals around their median forecasts, while the actual hit rate was only 55%. Almost 90% of the participants (64 of 72) showed $\overline{CONF} > \overline{HIT}$, so the hypothesis that subjects are as likely to show over- or under-confidence is easily rejected ($p<0.01$ in a sign-test or a Wilcoxon signed-ranks test). Closer look at the data suggests that low accuracy and exaggerated forecast-confidence equally contribute to miscalibration. The Spearman correlation between OC and $\overline{CONF}$ is similar in magnitude to the correlation between OC and $\overline{HIT}$ and the hypothesis that the two correlations are equal in absolute value cannot be rejected ($\rho(OC,\overline{CONF})=0.54$; $\rho(OC,\overline{HIT})=-0.64$; $p=0.48$ by Hotelling-Williams test for dependent correlations). The disjoint confidence-accuracy

[5] The Web supplements are available at http://www.bgu.ac.il/~sonsinod/

[6] The probability of winning the 100 NIS was 100% for prediction errors smaller than 1%, 98% for errors smaller than 2%, etc. The QSR was $100*[1-(1-P)^2]$, where P is the likelihood assigned to the realized event (hit or miss). When drawing the payment assignment, we also drew a 1-100 threshold. Subjects received the 100 NIS when their (payment task) winning probability exceeded the threshold.

effects on miscalibration clearly show in direct inspection of the data. At the OC=50 level, for instance, subject 53 exhibited ultimate confidence $\overline{CONF}$=100 with about average $\overline{HIT}$=50, while subject 11 showed about average $\overline{CONF}$=75 with half smaller than average $\overline{HIT}$=25 (for convenience, we omit the % sign when discussing $\overline{CONF}$, $\overline{HIT}$ and OC). The average confidence of the N=8 underconfident (OC<0) subjects, to take another perspective, ranged between 20 and 80, while their hit rates varied between 50 and 87.5. A comparison between the four tasks with history charts (HC) and the four preceding tasks where historical information was not provided (NHC) reveals that miscalibration is about half smaller in the tasks with history charts (mean OC 13 compared to 27; $p<0.01$). Still, 2/3 of the subjects exhibit overconfidence in the HC condition, illustrating that overprecision persists when forecasters are faced with graphs showing strong price volatility.

5. Study 2

Study 2 was distributed as a take-home incentivized survey to a convenience sample (Ferber, 1977) consisting of participants in a professional preparation course to the Israeli SEC exams (N=43), members of an Internet stock-trading forum (N=29), and MBA students (N=25). The questionnaire consisted of three slightly altered FAATs where subjects submit their best point prediction F for the target return and separately assess the likelihood of return exceeding $F + \delta 1$ and the likelihood of return smaller than $F - \delta 2$. The increments $\delta 1$ and $\delta 2$ were equal (10% and 5%) in two tasks, but asymmetric $\delta 1$=11% and $\delta 2 = 6\%$ in the third task. The likelihood CONF that subjects assign to the $[F - \delta 2, F + \delta 1]$ interval was calculated by subtracting the tail events' likelihoods from 100%; e.g., if the subject assigns 25% likelihood to return larger than $F + \delta 1$ and 30% likelihood to return smaller than $F - \delta 2$, then CONF=45. One prediction stock repeated in all the questionnaires. The other two stocks were randomly drawn from the 150 largest stocks of the Tel-Aviv exchange, to preclude the bias that unrepresentative task selection may bring (Gigerenzer et al., 1991). The data was collected in March-April 2015 and the prediction period was the six months starting in May 1-st, 2015. As the equality of $\overline{CONF}$, $\overline{HIT}$ and OC across the 3 subsamples could not be rejected, we report the results for the complete sample of N=97 subjects (see Web Supplement B for the sample specific results).[7]

[7] Study 2 was run before study 1 and motivated the simpler version of FAAT as presented in Figure 2.

Surprisingly, more than 1/3 of the participants (N=36 of 97) submitted at least one tail-likelihood strictly exceeding 50% (again, the violation rates did not differ significantly across the three subsamples). Intuitively, a large tail probability may represent a statement of lack-of-confidence in the submitted point forecast. Formally, the submission of tail likelihoods exceeding 50% connects to violations of *set inclusion* (monotonicity) documented in diverse psychology studies. Slovic et al. (1976), for example, ask subjects to estimate the likelihood of 3 events regarding some character Tom: (a) Tom will select journalism as his College major (b) but quickly become unhappy with his choice (c) and switch to engineering. The average likelihood that subjects assign to (a) alone was 0.21; the likelihood of (a) and (b) almost doubled to 0.39; and the average likelihood of the conjunction (a)-(c) slightly increased further to 0.41 (see Tversky and Koheler, 1994 for more examples). In the current application, subjects that violate monotonicity appear as if assigning 50% probability to the event $R \geq F$, while assigning larger probability to smaller events such as $R \geq F + \delta$.

The results for the complete sample and for the subjects that did not violate monotonicity are summarized in table I. The mean $\overline{CONF}$ was 49 for the complete sample (N=97) and 62 for the subjects that did not violate monotonicity (N=61). The respective $\overline{HIT}$ rates were 19 and 18, with 78% (90%) of the participants showing $\overline{CONF} > \overline{HIT}$. Overconfidence is stronger in the common task (mean OC of 41, with 75 of the 97 subjects showing OC>0), but it is still highly significant in the randomly drawn assignments (mean OC of 24, with 69 of the 97 subjects showing OC>0 when the common task is ignored). The hypothesis OC=0 is rejected at $p<0.01$ for each of the subsamples: SEC exams students, Web-forum traders, and MBAs. Again, the correlations between OC and $\overline{CONF}$ (0.74 for N=97 and 0.60 for N=61) are similar in magnitude to the negative correlations between OC and $\overline{HIT}$ (-0.63 and -0.75). The overconfidence scores of subjects 83 and 113, for brief example, are almost similar (15.33 and 16.67), but subject 113 is low in confidence (15.33) and calibration (0), while subject 83 shows $\overline{CONF}$=83.33 and $\overline{HIT}$=66.67.

Table I: Summary of results

	$\overline{CONF}$	$\overline{HIT}$	OC	%(OC>0)
Study 1 (N=72)	75	55	20	89%
Study 2 – full sample (N=97)	49	19	30	78%
Study 2 – restricted sample (N=61)	62	18	44	90%
Study 3 – FAAT (N=26)	64	26	38	88%
Study 3 – SIFT (N=24)	90	40	50	100%

The table presents the mean $\overline{CONF}$, $\overline{HIT}$ and OC for each sample. %(OC>0) is the proportion showing overconfidence. Confidence is exogenously fixed at 90% in SIFT.

6. Study 3

To compare the FAAT to the standard interval forecasting task (SIFT), we ran an exploratory small-sample study where N=50 finance undergraduate students were randomly assigned to one of the two formats. The questionnaires consisted of six forecasting tasks and the incentivization method was similar to the one of the preceding experiments (see Web Supplement C for details).[8] In the questionnaires with standard tasks, subjects submitted their median forecast and 95% confidence limits for the target return, as illustrated in Figure 1. The questionnaires with FAATs used the same six prediction targets and fixed margins of 5% around the median, as shown in Figure 2. The prediction targets included two domestic stocks, two index-linked certificates, and two leading stocks from the Paris and NYSE exchanges. The data was collected in March 2018 and the forecasting interval was the second quarter (April-June) of 2018. N=26 students completed the FAAT questionnaires and N=24 students completed the SIFT version. The results are summarized at the bottom panel of Table I.

On average, the 90% confidence intervals submitted by the SIFT subjects were 20.2% long, about twice longer than the FAAT 10% intervals. The hit rate of the 90% intervals was close to 40, generating an overconfidence score of about 50. The mean $\overline{CONF}$ of the subjects receiving the FAAT questionnaires was 64. The hit rate of the shorter FAAT intervals was only

[8] Since the incentivization of quantile assessments is relatively complex, the SIFT instructions (and, for symmetry, study 3's FAAT instructions) skipped the details regarding the incentivization of the quantile (likelihood) assessments. The instructions explained that we skip details for brevity but the method is designed so that truthful reporting maximizes the chances of winning the fixed NIS prize. Students were invited to send an email for details.

26, so that the mean overconfidence score was 38 for the FAAT subjects. While the miscalibration score in FAAT is 12 points smaller, a Pitman permutation test could not reject equality (two-tailed p=0.16). Overprecision, moreover, is clearly evident in both groups, with all 24 SIFT subjects showing $\overline{HIT}$ smaller than 90%, and 23 of the 26 FAAT subjects exhibiting $\overline{HIT} < \overline{CONF}$. While the sample size of study 3 is too small for drawing general conclusions, we note that extending the comparative analysis requires extensive effort since the results may strongly depend on the format of the tasks (Juslin et al., 1999) and the CONF/δ parameters assumed in defining the SIFT/FAAT, respectively (Teigen and Jørgensen, 2005; Speirs-Bridge et al., 2010).[9]

7. Discussion

The vast psychology research on overconfidence proves the robustness of miscalibration, illustrating that it is hard to debias subjects from over-trusting their private assessments (cf., Alpert and Raiffa, 1982 for an early report of failed debiasing attempts; Moore et al., 2015 for a recent survey). The results of the three current studies indeed reveal that the exaggerated confidence in subjective forecasts persists when prospective investors judge the likelihood of exhibiting prediction errors smaller than preassigned levels. Over three studies, 163 of 195 finance competent subjects overestimate their hit rates, in line with the forecast-overprecision hypothesis.

By separating forecast-accuracy from forecast-certainty, the three-step FAAT improves on standard interval forecasting, showing that inaccuracy and excessive-certainty act as distinct sources of forecast-miscalibration. As point forecasting and likelihood assessments are easier and more natural to subjects than quantile assessments (Abbas et al., 2008; Wallsten et al., 2016), FAAT may produce less noisy, more meaningful measures of forecast-overprecision. The FAAT overconfidence score may therefore prove more powerful in testing hypothesized links between forecast-overconfidence and aspects of financial decision (Skala, 2008). Empirical studies illustrate that intuitive proxies for overconfidence such as partition by gender (Barber and Odean, 2000), managers' inclination to hold on to their stock options (Malmendier and Tate, 2008), or traders' risk exposure in terms of position size (Forman and Horton, 2019)

[9] In Teigen and Jørgensen (2005) studies 2-3 for specific examples.

link with exaggerated non-profitable trading. The micro level, experimental or survey-based evidence regarding the overprecision-trading correlations, however, is inconclusive and statistically weak (cf., Biais et al., 2005; Glaser and Weber, 2007; Deaves et al., 2009; Nosic and Weber, 2010; Broihanne et al., 2014; Fellner-Röhling and Krügel, 2014; Merkle, 2017). By separating forecast-accuracy from forecast-certainty the FAATs of the present study open possibility for more efficient tests of the links between trading propensity/profitability and forecast-certainty, accuracy, and overconfidence.

References

Abbas, A. E., Budescu, D. V., Yu, H. T., and Haggerty, R. (2008). A comparison of two probability encoding methods: Fixed probability vs. fixed variable values. *Decision Analysis, 5*(4), 190-202.

Alpert, M., and Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, and A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Astebro, T., Herz, H., Nanda, R., and Weber, R. A. (2014). Seeking the roots of entrepreneurship: Insights from behavioral economics. *Journal of Economic Perspectives, 28*(3), 49-70.

Bansal, S., and Palley, A. (2017). Is It Better To Elicit Quantile Or Probability Judgments? A Comparison of Direct and Calibrated Procedures for Estimating a Continuous Distribution. Working paper.

Barber, B. M., and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance, 55*(2), 773-806.

Ben-David, I., Graham, J. R., and Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics, 128*(4), 1547-1584.

Bergu, K. (2019). Overconfidence and Trading Behavior: Does There Exist a Causal Link?. Working paper

Bhandari, G., and Deaves, R. (2006). The demographics of overconfidence. *The Journal of Behavioral Finance*, *7*(1), 5-11.

Biais, B., Hilton, D., Mazurier, K., and Pouget, S., (2005). Judgmental overconfidence, self-monitoring, and trading performance in an experimental financial market, *The Review of Economic Studies 72 (2),* 287-312.

Broihanne, M. H., Merli, M., and Roger, P. (2014). Overconfidence, risk perception and the risk-taking behavior of finance professionals. *Finance Research Letters*, *11*(2), 64-73.

Cubitt, R. P., Starmer, C., and Sugden, R. (1998). On the validity of the random lottery incentive system. *Experimental Economics*, *1*(2), 115-131.

Daniel, K., and Hirshleifer, D. (2015). Overconfident investors, predictable returns, and excessive trading. *The Journal of Economic Perspectives*, *29*(4), 61-87.

Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *the Journal of Finance*, *53*(6), 1839-1885.

Deaves, R., Lüders, E., and Luo, G., (2009). An experimental test of the impact of overconfidence and gender on trading activity. *Review of Finance 13,* 555–575.

Fellner-Röhling, G., and Krügel, S. (2014). Judgmental overconfidence and trading activity. *Journal of Economic Behavior and Organization*, *107*, 827-842.

Ferber, R. (1977). Research by convenience. *Journal of Consumer Research*, *4*(1), 57-58.

Forman, J., and Horton, J. (2019). Overconfidence, position size, and the link to performance. *Journal of Empirical Finance, 53,* 291-309.

Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: a Brunswikian theory of confidence. *Psychological review*, *98*(4), 506-528.

Glaser, M., Langer, T., and Weber, M. (2013). True Overconfidence in Interval Estimation Based on a New Measure Of Miscalibration. *Journal of Behavioral Decision Making, 26:* 405-417

Glaser, M. and Weber, M. (2007). Overconfidence and trading volume. *The Geneva Risk and Insurance Review, 32*(1), 1-36.

Gneiting, T., and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359-378.

Graham, J. R., and Harvey, C. R. (2001). *Expectations of equity risk premia, volatility and asymmetry from a corporate finance perspective* (No. w8678). National Bureau of Economic Research.

Grosshans, D., and Zeisberger, S. (2018). All's well that ends well? On the importance of how returns are achieved. *Journal of Banking & Finance*, *87*, 397-410.

Harrison, G. W., Martínez-Correa, J., and Swarthout, J. T. (2014). Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior and Organization*, *101*, 128-140.

Hey, J. D., and Lee, J. (2005). Do subjects separate (or are they sophisticated)?. *Experimental Economics*, *8*(3), 233-265.

Hossain, T., and Okui, R. (2013).The binarized scoring rule.*The Review of Economic Studies*, *80*(3), 984-1001.

Hyndman, R. J., and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. OTexts.

Juslin, P., Wennerholm, P., and Olsson, H. (1999). Format dependence in subjective probability calibration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1038-1052.

Juslin, P., Winman, A., and Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review, 114*(3), 678-703.

Keefer, D. L., and Bodily, S. E. (1983). Three-point approximations for continuous random variables. *Management Science, 29*(5), 595-609.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., and Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216-247.

Langnickel, F., and Zeisberger, S. (2016). Do we measure overconfidence? A closer look at the interval-production task. *Journal of Economic Behavior and Organization, 128*, 121-133.

Lawrence, M., Goodwin, P., O'Connor, M., and Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25years. *International Journal of Forecasting*, *22*(3), 493-518.

Malmendier, U., and Tate, G. (2008). Who makes acquisitions? CEO overconfidence and the market's reaction. *Journal of Financial Economics*, *89*(1), 20-43.

Malmendier, U. and Taylor, T. (2015). On the verges of overconfidence. *Journal of Economic Perspectives*. *29(4)*, 3-8.

McKenzie, C. R., Liersch, M. J., and Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you?. *Organizational Behavior and Human Decision Processes*, *107*(2), 179-191.

Merkle, C. (2017). Financial overconfidence over time: Foresight, hindsight, and insight of investors. *Journal of Banking and Finance*, *84*, 68-87.

Moore, D. A., and Healy, P. J. (2008). The trouble with overconfidence. *Psychological review*, *115*(2), 502-517.

Moore, D. A., Tenney, E. R., and Haran, U. (2015). Overprecision in judgment. *The Wiley Blackwell Handbook of Judgment and Decision Making*, 182-209.

Murad, Z., Sefton, M., and Starmer, C. (2016). How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, *52*(1), 21-46.

Murphy, A. H., and Winkler, R. L. (1974). Probability forecasts: A survey of National Weather Service forecasters. *Bulletin of the American Meteorological Society*, *55*(12), 1449-1452.

Neyse, L., Bosworth, S., Ring, P., and Schmidt, U. (2016). Overconfidence, incentives and digit ratio. *Scientific reports*, *6*, 23294.

Nosic, A., and Weber, M. (2010). How riskily do I invest? The role of risk attitudes, risk perceptions, and overconfidence. *Decision Analysis*, 7(3), 282-301.

Oberlechner, T., and Osler, C. (2012). Survival of overconfidence in currency markets. *Journal of Financial and Quantitative Analysis*, *47*(01), 91-113.

Skala, D. (2008). Overconfidence in psychology and finance - an interdisciplinary literature review. *Bank i Kredyt*, *39(4)*, 33–50.

Slovic, P., Fischhoff, B., and Lichtenstein, S. (1976). Cognitive processes and societal risk taking. In *Decision making and change in human affairs* (pp. 7-36). Springer Netherlands.

Sonsino, D., and Regev, E. (2013). Informational overconfidence in return prediction–More properties. *Journal of Economic Psychology*, *39*, 72-84.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., and Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis*, *30*(3), 512–23.

Teigen, K. H., and Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: On the credibility of credible intervals. *Applied Cognitive Psychology*, *19*(4), 455-475.

Tversky, A., and Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547-567.

Wallsten, T. S., Shlomi, Y., Nataf, C., and Tomlinson, T. (2016). Efficiently encoding and modeling subjective probability distributions for quantitative variables. *Decision*, *3*(3), 169-189.

Winman, A., Hansson, P., and Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval-evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1167–1175.