Getting more wisdom out of the crowd: The case of competence-weighted aggregates

Michael Goedde-Menke, Enrico Diecidue, Andreas Jacobs, and Thomas Langer
University of Münster, Germany *INSEAD, France*

Link to Working Paper:

Summary

We show that group discussions can serve as an instrument to improve individuals' calibration, which in turn strongly increases the accuracy of competence-weighted, statistical aggregates. We conduct an experiment in which participants estimate quantities and report their self-perceived competence for various judgment problems. In addition, they engage in group discussions with other judges on unrelated judgment tasks. We find that prior to participating in the group discussions, judges' self-perceived competence and their estimation accuracy are poorly aligned, which causes competence weighting to perform worse than prediction markets and simple averaging. However, the information exchange facilitated by the group discussions improved judges' calibration, raising the accuracy of competence-weighted aggregates on subsequent judgment problems to prediction market levels and beyond.

1. Motivation & Contribution

Motivation

Managerial decisions often involve singular judgment problems

- Specific estimates (future outcomes) required
- Accuracy of estimates affects returns to shareholders and other stakeholders

Estimation accuracy improves by exploiting the 'wisdom of the crowd'
(Clemen and Winkler 1986, Hastie and Kameda 2005, Larrick and Soll 2006, O'Hagan 2019)

- Frequently used: Simple averaging

$$\hat{x}_{AVG} = \frac{1}{n} \cdot \sum_{i=1}^{n} \hat{x}_i$$

- Less accurate than prediction markets though
(Atanasov et al. 2017, Palan et al. 2020, Wolfers and Zitzewitz 2004)

But: Equal weighting ignores differences in competence across judges

- Competence weighting: Give more weight to more competent judges
(e.g., Aspinall 2010)

$$\hat{x}_{AVG(comp)} = \frac{\Sigma_{i=1}^{n} competence_i \cdot \hat{x}_i}{\Sigma_{i=1}^{n} competence_i}$$

- Problem: Reliably identifying competence based on
  - past performance → Often not available
  - self-perceived competence → Individuals poorly calibrated

Competence weighting: Estimation accuracy depends on calibration

- Estimation accuracy: $SE_i = (\hat{x}_i - x_{true})^2$
- Calibration: $\rho_{spearman}(SE_i, competence_i)$
- Example: How much of German electricity demand was produced by fossil fuels in 2021 (in %)?
  - True quantity (x_{true}): 60%

Group A

$Judge_i$	$\hat{x}_i$	$competence_i$	SE_i
1	20%	1	16.0%
2	32%	3	7.8%
3	35%	2	6.3%
4	48%	4	1.4%
5	50%	5	1.0%
6	65%	7	0.3%

- Well calibrated ($\rho = -0.86$)
- Accuracy:
  - $SE_{AVG} = 3.4\%$
  - $SE_{AVG(comp)} = 1.2\%$

→ Competence weighting more accurate than simple averaging

Group B

$Judge_i$	$\hat{x}_i$	$competence_i$	SE_i
7	20%	6	16.0%
8	32%	4	7.8%
9	35%	5	6.3%
10	48%	1	1.4%
11	50%	2	1.0%
12	65%	3	0.3%

- Poorly calibrated ($\rho = +0.86$)
- Accuracy:
  - $SE_{AVG} = 3.4\%$
  - $SE_{AVG(comp)} = 5.5\%$

→ Competence weighting less accurate than simple averaging

Contribution

⇒ **Group discussions can improve individuals' calibration**

⇒ **Competence-weighted aggregates can get more wisdom out of the crowd**

2. Experimental Design

Mechanism: How group discussions can improve individual calibration

Miscalibration discourages use of competence-weighted aggregates

- Overconfidence and underconfidence widespread
(Griffin and Tversky 1992, Jose et al. 2014, Kruger 1999, Larrick et al. 2007, Lichtenstein et al. 1982, Moore et al. 2017)

Key sources of miscalibration: Unawareness and confirmatory bias

- Individuals are simply unaware of their miscalibration
(Arkes 1991, Benson and Önkal 1992, Sharp et al. 1988)
- Individuals primarily seek information that confirms their hypothesis
(Koriat et al. 1980)

Group discussions likely improve calibration because they address both sources

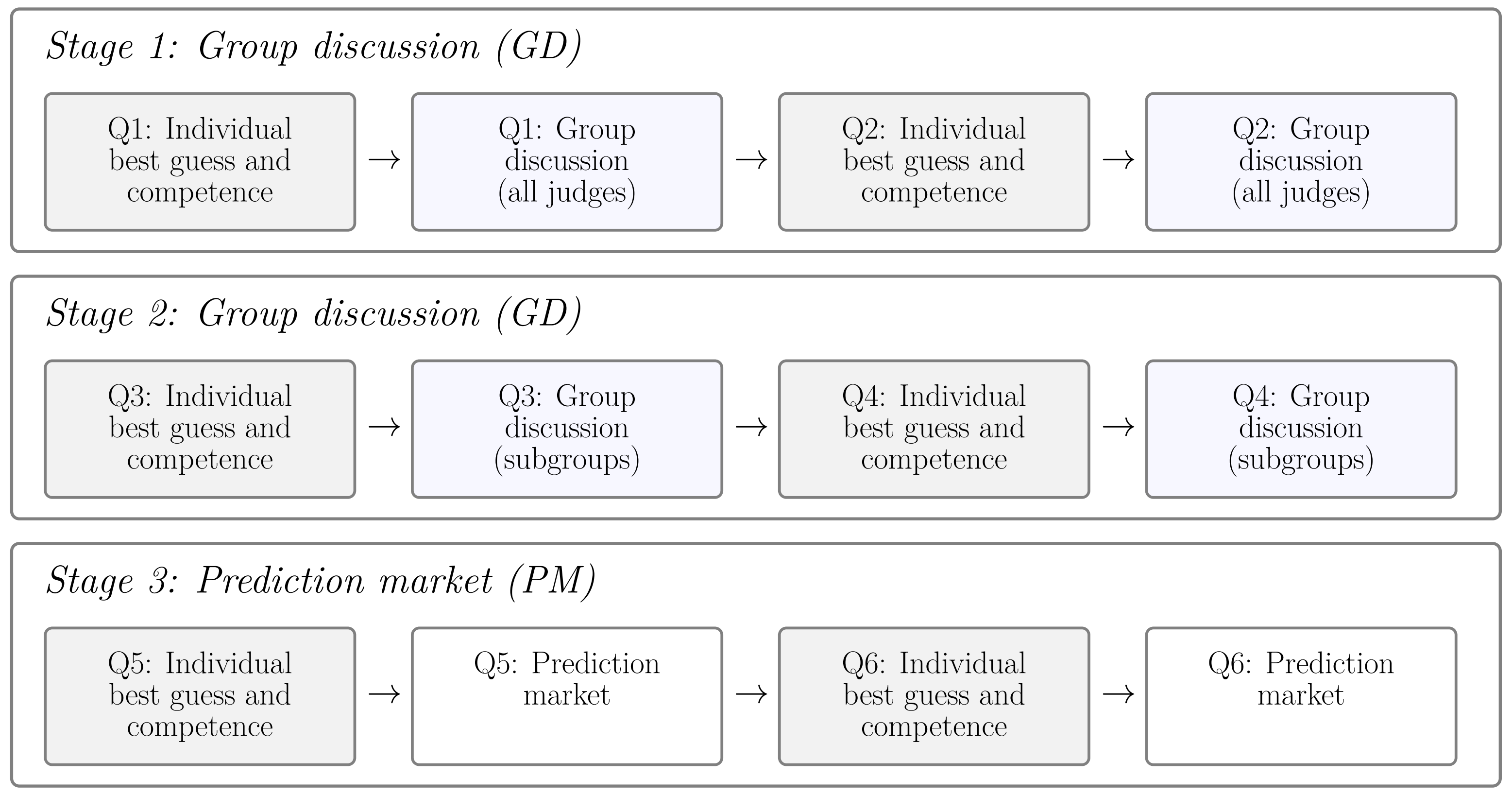
- Through calibration feedback:
Individuals obtain cues from others whether their self-perceived competence is justified
(Larson and Christensen 1993, Schultze et al. 2012) → tackles unawareness
- Through in-depth reflection of the estimation process:
Individuals must defend their own beliefs while absorbing the reasoning of others
(Keck and Tang 2021, Minson et al. 2018, Trouche et al. 2014) → tackles confirmatory bias

Lab experiment: Details and sequential stages

Subjects: 288 undergraduate students from the University of Münster, split into groups of 12

- Each subject: Best guess and self-perceived competence for six real-world quantities (0-100%)
- Each session: 2 hours and 15 minutes; about €21 average payout (fix + variable components)

Each group participated in three sequential stages (counterbalanced design):



Stage 1: Group discussion (GD)
Q1: Individual best guess and competence → Q1: Group discussion (all judges) → Q2: Individual best guess and competence → Q2: Group discussion (all judges)

Stage 2: Group discussion (GD)
Q3: Individual best guess and competence → Q3: Group discussion (subgroups) → Q4: Individual best guess and competence → Q4: Group discussion (subgroups)

Stage 3: Prediction market (PM)
Q5: Individual best guess and competence → Q5: Prediction market → Q6: Individual best guess and competence → Q6: Prediction market

Key: Ordering of sequential stages determines degree of information exchange

	Rounds of Information Exchange		
Stage ordering	Before first stage	Before second stage	Before third stage
PM → GD → GD	0 rounds	0 rounds	2 rounds
GD → PM → GD	0 rounds	2 rounds	2 rounds
GD → GD → PM	0 rounds	2 rounds	4 rounds

Robustness

We conduct several tests to show that our results are robust:

-Alternative analyses to quantify the impact of information exchange on calibration and accuracy
-Alternative measure of accuracy: Absolute error rather than squared error
-Alternative aggregation: Group medians rather than group means
-Alternative problem sets: Bootstrapping questions

3. Results

Our findings in a nutshell

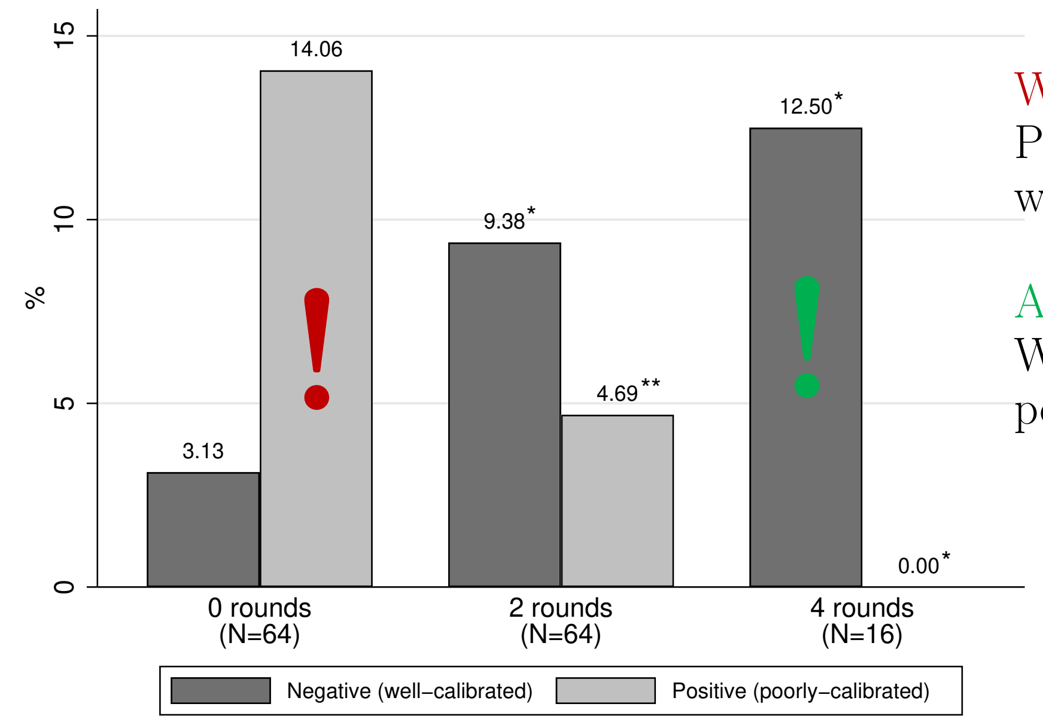
⇒ **Information exchange through group discussions improves individuals' calibration in subsequent and unrelated judgment problems**

⇒ **Improved calibration boosts the accuracy of competence-weighted aggregates to prediction market levels and beyond**

Main results

More rounds of information exchange improve judges' calibration

The figure below shows the proportions of poorly- and well-calibrated groups by rounds of information exchange

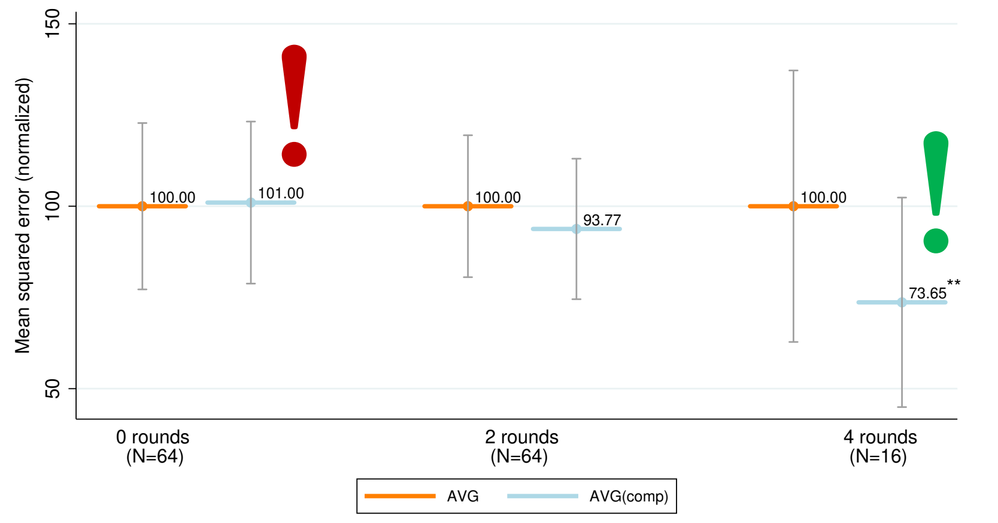


Without information exchange (0 rounds):
Poorly-calibrated groups outnumber well-calibrated groups

After 4 rounds of information exchange:
Well-calibrated groups outnumber poorly-calibrated groups

Improved calibration boosts accuracy of competence-weighted aggregates

The figure below compares the accuracy of competence weighting to simple averaging based on the normalized mean squared error (MSE) and by rounds of information exchange

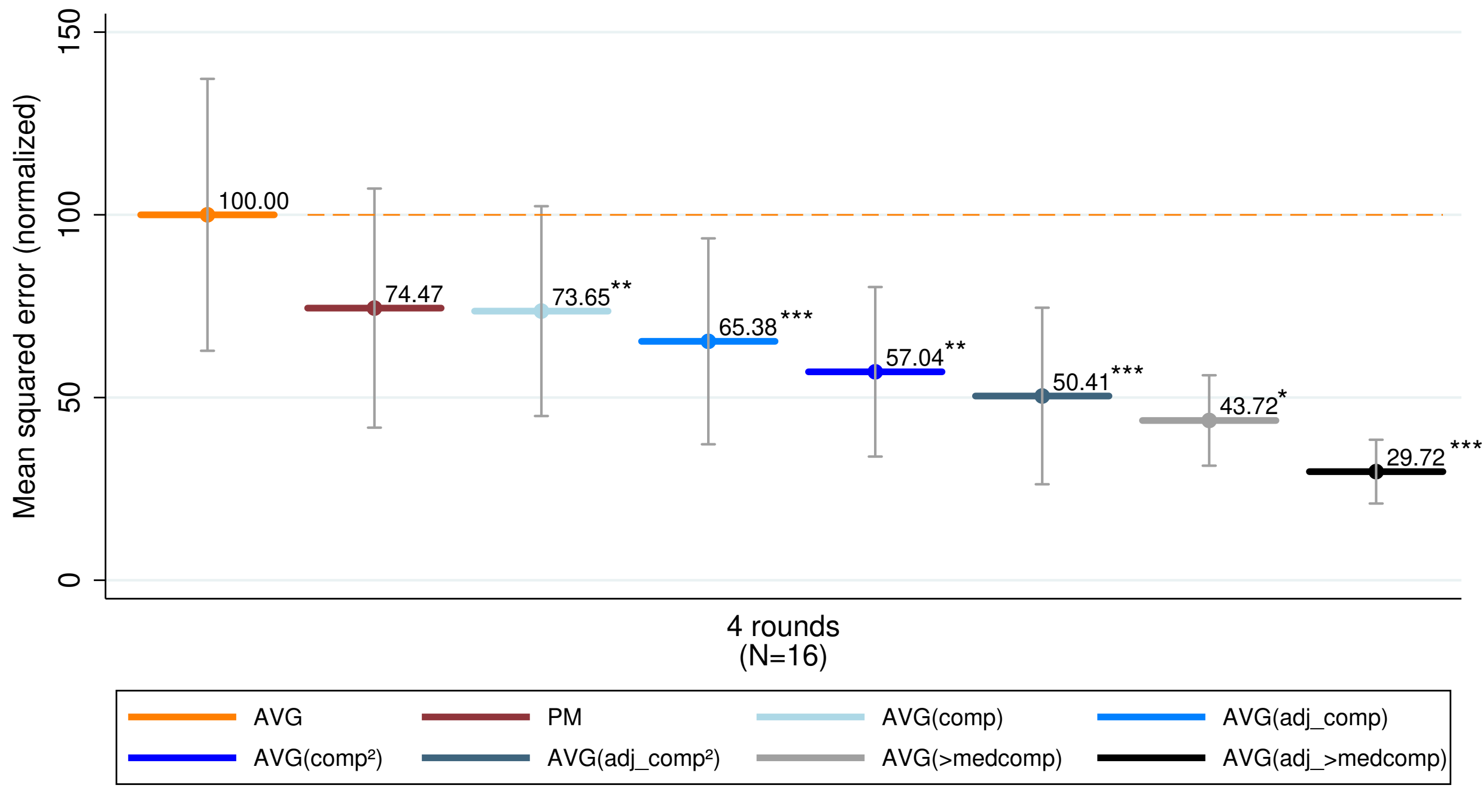


Without information exchange (0 rounds):
Competence weighting less accurate than simple averaging

After 4 rounds of information exchange:
Competence weighting more accurate than simple averaging

Advanced competence-weighted aggregates reduce estimation error by 60% compared to prediction markets

The figure below compares the accuracy of more advanced competence-weighted aggregates (adjusted competence information, select crowds) to prediction markets and simple averaging based on the normalized mean squared error (MSE) and after 4 rounds of information exchange


Mean squared error (normalized)
73.65**
65.38***
57.04**
50.41***
43.72*
29.72***
4 rounds (N=16)
AVG
PM
AVG(comp)
AVG(adj_comp)
AVG(comp²)
AVG(adj_comp²)
AVG(>medcomp)
AVG(adj_>medcomp)