nature human behaviour

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41562-017-0247-6

In the format provided by the authors and unedited.

The wisdom of the inner crowd in three large natural experiments

Dennie van Dolder[1,2]* and Martijn J. van den Assem[2]

[1]Centre for Decision Research and Experimental Economics, University of Nottingham, Nottingham, UK. [2]School of Business and Economics, VU Amsterdam, Amsterdam, The Netherlands. *e-mail: d.van.dolder@vu.nl

Supplementary Notes

Supplementary Note 1: Data cleaning and linking

Data cleaning

The raw data contained 369,297 pseudonymized entries from 2013, 388,355 from 2014, and 407,627 from 2015. For each entry we know the voucher code, the date and time (in milliseconds) of submission, and the player's gender and birthdate (see Supplementary Figures 2-5 for distributions). For entries submitted at the same time we know the order in which they were entered. We do not know whether entries are submitted via a terminal inside a casino or via the internet from elsewhere.

For 2013 a relatively small number of 27 entries appeared in the data twice. We have removed these duplicate entries, bringing the total number of observations from 2013 to 369,270. There were no duplicate entries in the data from 2014 and 2015.

The competitions officially spanned the periods of 12 November 2013 through 1 January 2014, 11 November 2014 through 2 January 2015, and 10 November 2015 through 3 January 2016. A small number of entries was submitted prior to the official start date: 10 in 2013, 3 in 2014, and 5 in 2015. As these entries were likely made as test inputs by the casino we excluded them from the analyses.[1] This filtering leaves us with 369,260 observations for 2013, 388,352 observations for 2014, and 407,622 for 2015.

In each year there is a considerable number of entries after the official end date. For the 2013 competition we observed 1,300 entries on January 2, all before 10 a.m. For the 2014 competition there are 2,979 entries after the official deadline. These are slightly more spaced out in time: 2,759 on January 3, 173 on January 4, and 47 on January 5. For the 2015 competition there were 1,385 entries on January 4 and 47 on January 5. The natural interpretation is that it was technically possible to submit entries past the official deadline. Further inspection of the late entries did not reveal anything that was out of the ordinary, and hence we included these observations in the analyses.

[1] In 2014 the early entries were 1234567, 9876543, and 12345; in 2015 the early entries were 1, 2, 3, 4, and 5. In both years the numbers were all entered a few days before the start of the competition (November 5 in 2014, November 6 in 2015). In 2013 the early entries were not noticeably contrived and they were entered only one day prior to the official start date (November 11), but for consistency we excluded these as well.

Linking entries

To link entries from the same individual, the casino used the initials, surname, and date of birth provided by the participant (all mandatory fields). This approach distinguished 169,913 different individuals in 2013, 159,793 in 2014, and 167,146 in 2015, all represented in our data by a unique numeric pseudonym.

The data also contained a unique numeric string for each unique e-mail address and for each unique phone number.[2] Upon inspection it turned out that different pseudonyms sometimes shared the same numeric e-mail or phone string, indicating that these pseudonyms might represent the same person. From correspondence with the casino we learned that some of the participants with multiple initials had sometimes used their full set of initials and sometimes only their first. As a consequence of the casino's matching procedure their entries were appearing in our data as entries from two different participants.

To solve this issue we first merged pseudonyms sharing the same numeric e-mail string, gender, and date of birth. Gender and date of birth were mandatory fields, and we have a numeric e-mail string for 76.6% of the entries from 2013, for 80.3% from 2014, and for 84.5% from 2015. After this first adjustment, there were 164,786 different individuals in the sample from 2013, 155,511 in 2014, and 162,748 in 2015.

Second, for all pseudonyms with no numeric e-mail string, we merged those sharing the same numeric phone string, gender, and date of birth.[3] We have a numeric phone string for 97.5% of the entries from 2013 that had no numeric e-mail string. For 2014 and 2015 these percentages are 96.1 and 97.2, respectively. This second and final adjustment further lowered the number of unique players to 163,719 in 2013, 154,790 in 2014, and 162,275 in 2015.

[2] In the data provided by the casino, each unique pseudonym had a consistent numeric e-mail string across all entries. This was not the case for the numeric phone string. Possible explanations are that players have multiple phone numbers, that they entered their number in different formats, or that they intentionally or unintentionally entered a wrong number.

[3] All pseudonyms with no numeric e-mail string did have a consistent numeric phone string across all entries.

Supplementary Note 2: Additional analyses

The paper uses the mean squared error relative to the true value to measure the quality of estimates, and reports the results of analyses with the transformed estimates. Below we show that (i) using the mean absolute error instead of the mean squared error and (ii) using the untransformed estimates instead of the transformed estimates both lead to similar conclusions.

Mean absolute error, transformed data

We observe statistically significant benefits from within-person aggregation when we use the mean absolute error (MAE). In 2013, the MAE of the average of the first two estimates from one individual was lower than the MAE of either estimate alone ($MAE_1 = 1.33$, $MAE_2 = 1.24$, $MAE_{1\&2} = 1.20$, with $t(60{,}869) > 16.90$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 ($MAE_1 = 1.36$, $MAE_2 = 1.28$, $MAE_{1\&2} = 1.23$, $t(59{,}156) > 22.27$, $p < 0.0001$) and in 2015 ($MAE_1 = 1.39$, $MAE_2 = 1.35$, $MAE_{1\&2} = 1.29$, $t(61{,}893) > 28.14$, $p < 0.0001$). Across the three events Cohen's d varies between 0.09 and 0.13 for comparisons between the average and the first estimate, and between 0.04 and 0.05 for comparisons between the average and the second estimate.

Supplementary Figure 13 displays the MAE of aggregations across T different players and the MAE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, and shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation at best only approximates the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 13 also displays the MAE of individual consecutive estimates, showing that estimates improved over time and that this improvement did not match the improvement that could have been obtained by averaging.

Defining the accuracy gain as in the paper, but for the absolute instead of the squared error, Supplementary Figure 14 shows that the accuracy gain from aggregation is larger if the estimations are further apart in time.

Mean squared error, untransformed data

We observe statistically significant benefits from within-person aggregation when we work with the untransformed data instead of the transformed data. In 2013, the MSE of the

geometric mean of the first two estimates from one individual was lower than the MSE of either estimate alone (MSE_1 = 2.28e+11, MSE_2 = 2.29e+11, $MSE_{1\&2}$ = 9.44e+10, with $t(60{,}869) > 11.57$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 (MSE_1 = 3.34e+11, MSE_2 = 3.61e+11, $MSE_{1\&2}$ = 1.58e+11, $t(59{,}156) > 12.99$, $p < 0.0001$) and in 2015 (MSE_1 = 6.75e+11, MSE_2 = 7.62e+11, $MSE_{1\&2}$ = 3.94e+11, $t(61{,}893) > 17.49$, $p < 0.0001$). Across the three events Cohen's d varies between 0.05 and 0.06 for comparisons between the aggregate and the first estimate, and between 0.05 and 0.07 for comparisons between the aggregate and the second estimate.

Supplementary Figure 15 displays the MSE of aggregations across T different players and the MSE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, using the untransformed data. The figure shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation is lower than the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 15 also displays the MSE of individual consecutive untransformed estimates, showing no improvement over time, whereas improvement could have been obtained by averaging. The difference with the pattern of decreasing errors for the individual transformed estimates seems to be caused by outliers, as further (unreported) analyses show that the *median* squared error *did* decrease with the number of estimates previously made. Our transformation decreases the influence of these outliers.

Supplementary Figure 16 shows that the accuracy gain from aggregation is larger if the untransformed estimations are further apart in time.

Mean absolute error, untransformed data

We observe statistically significant benefits from within-person aggregation when we use the mean absolute error (MAE) and work with the untransformed data. In 2013, the MAE of the average of the first two estimates from one individual was lower than the MAE of either estimate alone (MAE_1 = 77,452, MAE_2 = 73,125, $MAE_{1\&2}$ = 51,014, with $t(60{,}869) > 17.67$ and two-sided $p < 0.0001$ in the two comparisons). This was also true in 2014 (MAE_1 = 108,303, MAE_2 = 107,459, $MAE_{1\&2}$ = 77,621, $t(59{,}156) > 19.10$, $p < 0.0001$) and in 2015 (MAE_1 = 216,291, MAE_2 = 222,804, $MAE_{1\&2}$ = 169,401, $t(61{,}893) > 25.04$, $p < 0.0001$). Across the three events Cohen's d varies between 0.06 and 0.07, both for comparisons

between the aggregate and the first estimate and for comparisons between the aggregate and the second estimate.

Supplementary Figure 17 displays the MAE of aggregations across T different players and the MAE of aggregations across the first t consecutive estimates for players who provided at least $K = 5$ or $K = 10$ estimates in a given year, using the untransformed data. The figure shows that aggregating across individuals works substantially better than aggregating judgements from the same individual. The expected potential benefit from within-person aggregation is lower than the benefit of combining the first estimates of two randomly selected individuals.

Supplementary Figure 17 also displays the MAE of individual consecutive untransformed estimates, showing no improvement over time, whereas improvement could have been obtained by averaging. The difference with the pattern of decreasing errors for the individual transformed estimates seems to be caused by outliers, as further (unreported) analyses show that the *median* absolute error *did* decrease with the number of estimates previously made. Our transformation decreases the influence of these outliers.

Defining the accuracy gain as in the paper, but for the absolute instead of the squared error, Supplementary Figure 18 shows that the accuracy gain from aggregation is larger if the untransformed estimations are further apart in time.

Supplementary Tables

Supplementary Table 1

Arithmetic and geometric mean across first estimates

Aggregation measures are calculated across all players' first estimates. *N* is the number of players. Percentage deviations relative to the true values are in parentheses.

	N	True value	Arithmetic mean	Geometric mean
2013	163,719	12,564	79,973 (+537%)	10,618 (-15%)
2014	154,790	23,363	128,414 (+450%)	17,652 (-24%)
2015	162,275	22,186	271,573 (+1,124%)	48,790 (+120%)

Supplementary Table 2

Estimation error decomposition for sub-samples

The table displays the maximum likelihood parameter estimates for the components of the estimation error (see Methods), using the first K estimates of players who provided at least $K = 5$ or $K = 10$ estimates in a given year. Shown are the population bias (μ), the variance of idiosyncratic bias (τ^2), and the variance of random noise (σ^2). Implied T_∞^* follows from the parameter estimates using $T_\infty^* = 1 + \sigma^2 / \tau^2$. Original T_∞^* is derived from the best-fitting hyperbolic function $MSE = a / t + b$ (see Figure 1), and shown for the purpose of comparison. N is the number of players.

	K	N	μ	τ^2	σ^2	Implied T_∞^*	Original T_∞^*
2013	5	15,056	-0.26	1.74	0.91	1.52	1.52
	10	4,989	-0.28	1.47	0.82	1.56	1.56
2014	5	17,047	-0.47	1.54	0.81	1.52	1.52
	10	6,538	-0.51	1.28	0.73	1.57	1.57
2015	5	17,630	0.44	1.73	0.80	1.46	1.46
	10	6,801	0.39	1.49	0.76	1.51	1.51

Supplementary Table 3

Estimation error decomposition for all data

The table displays the maximum likelihood parameter estimates for the components of the estimation error (see Methods), using all 369,260 estimates from the 163,719 players in 2013, all 388,352 estimates from the 154,790 players in 2014, and all 407,622 estimates from the 162,275 players in 2015. Definitions are as in Supplementary Table 2.

	μ	τ^2	σ^2	Implied T_∞^*
2013	-0.17	1.95	0.88	1.45
2014	-0.28	2.08	0.77	1.37
2015	0.79	2.24	0.80	1.36

Supplementary Table 4

Estimation error decomposition with delay-dependent co-variance

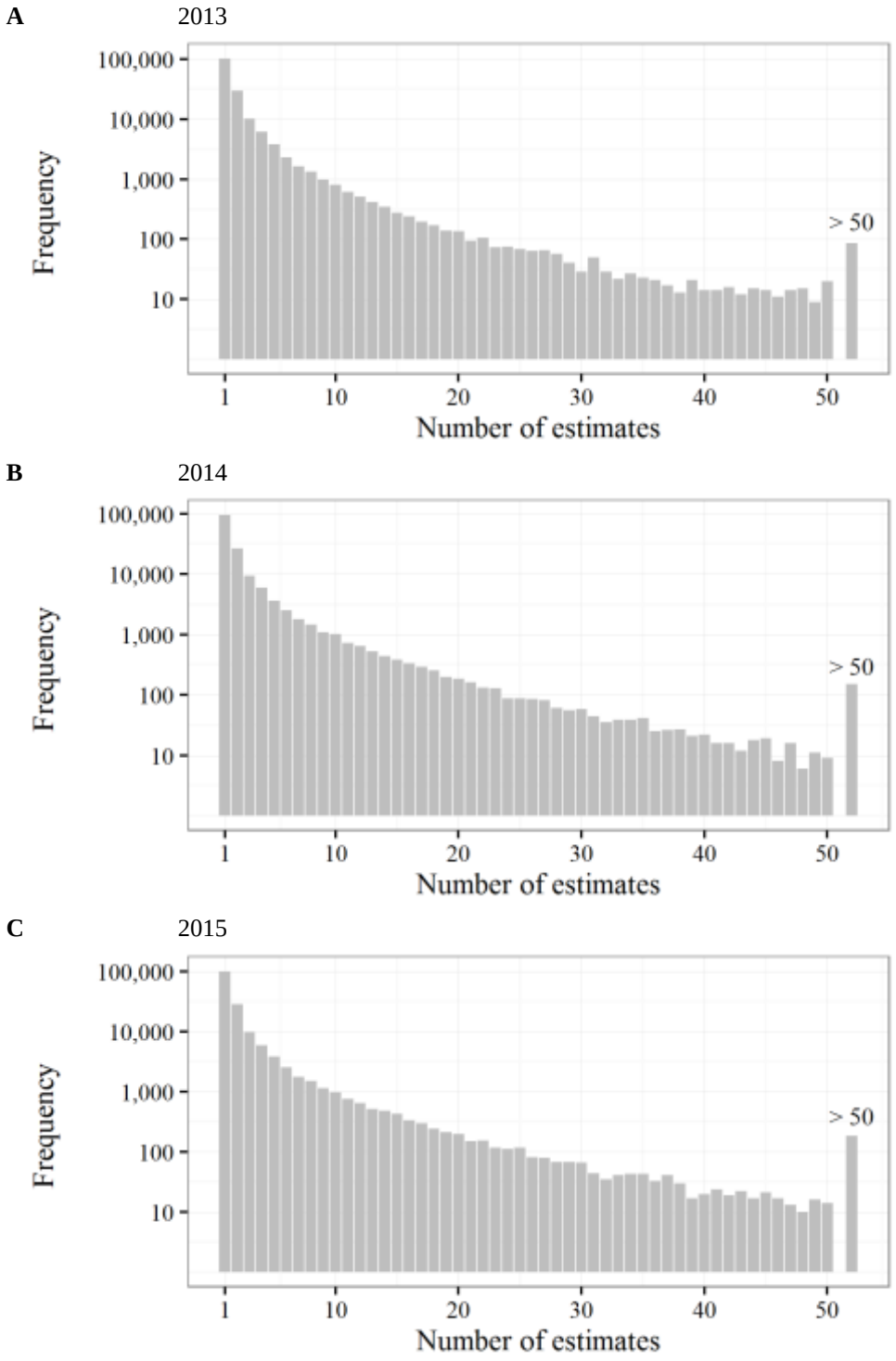
The table displays the maximum likelihood parameter estimates for the components of the estimation error with delay-dependent covariance (see Methods), using all 369,260 estimates from the 163,719 players in 2013, all 388,352 estimates from the 154,790 players in 2014, and all 407,622 estimates from the 162,275 players in 2015. λ determines the speed of decay of the delay-dependent covariance, and (1 - δ) allows for a discontinuous jump in the covariance structure such that estimates provided simultaneously are not required to be perfectly correlated. $t_{1/2}$ is the half-life of the delay-dependent covariance (in days). Limiting T_{∞}^{*} gives the estimated value of T_{∞}^{*} for infinitely long time delays, such that all delay-dependent covariance has dissipated.

	μ	τ^2	σ^2	1 - δ	λ	$t_{1/2}$	Limiting T_{∞}^{*}
2013	-0.17	1.41	1.40	0.62	0.083	8.39	1.99
2014	-0.28	1.60	1.24	0.63	0.086	8.01	1.78
2015	0.78	1.73	1.29	0.63	0.084	8.22	1.75

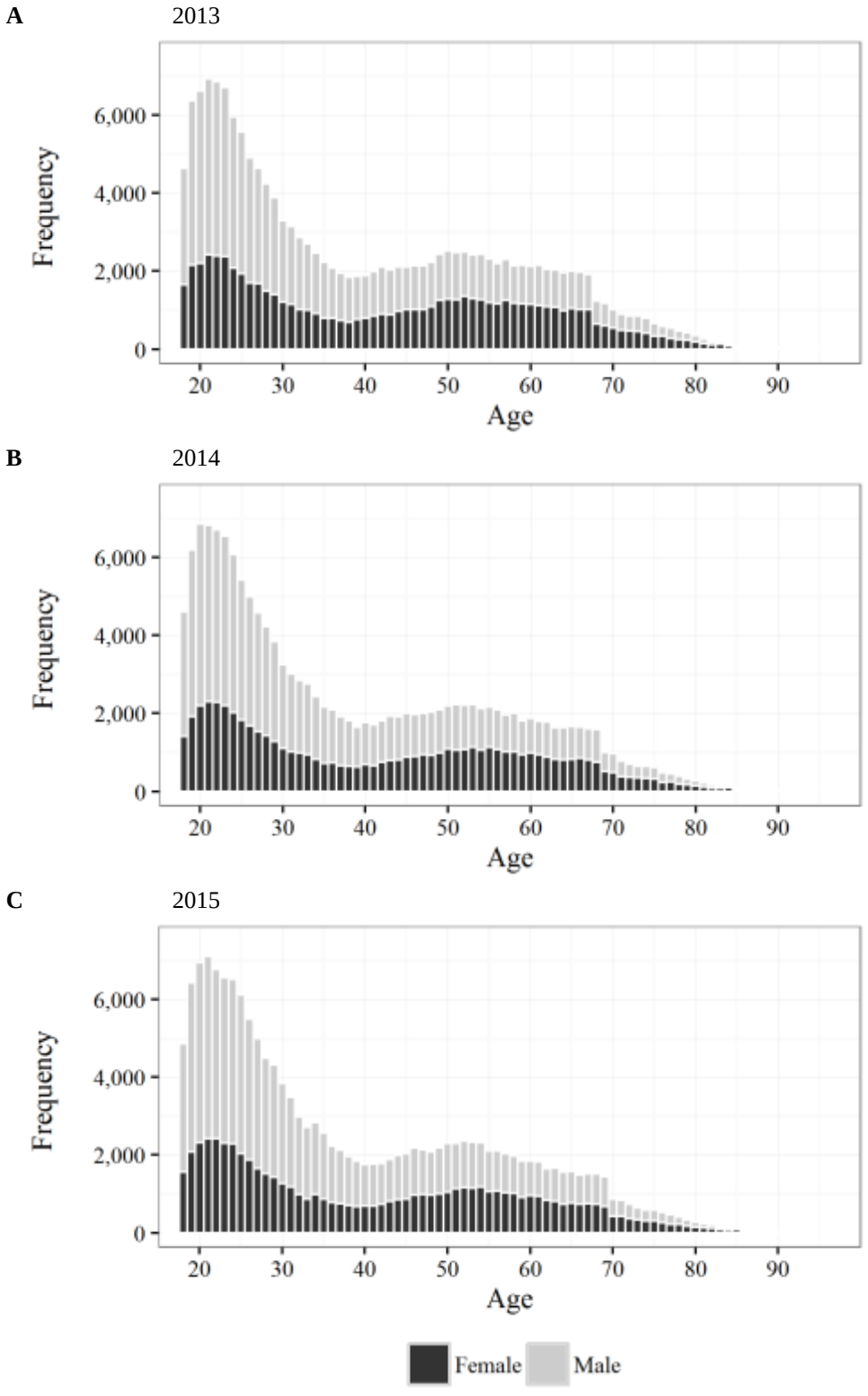
Supplementary Figures

A **B** **C**

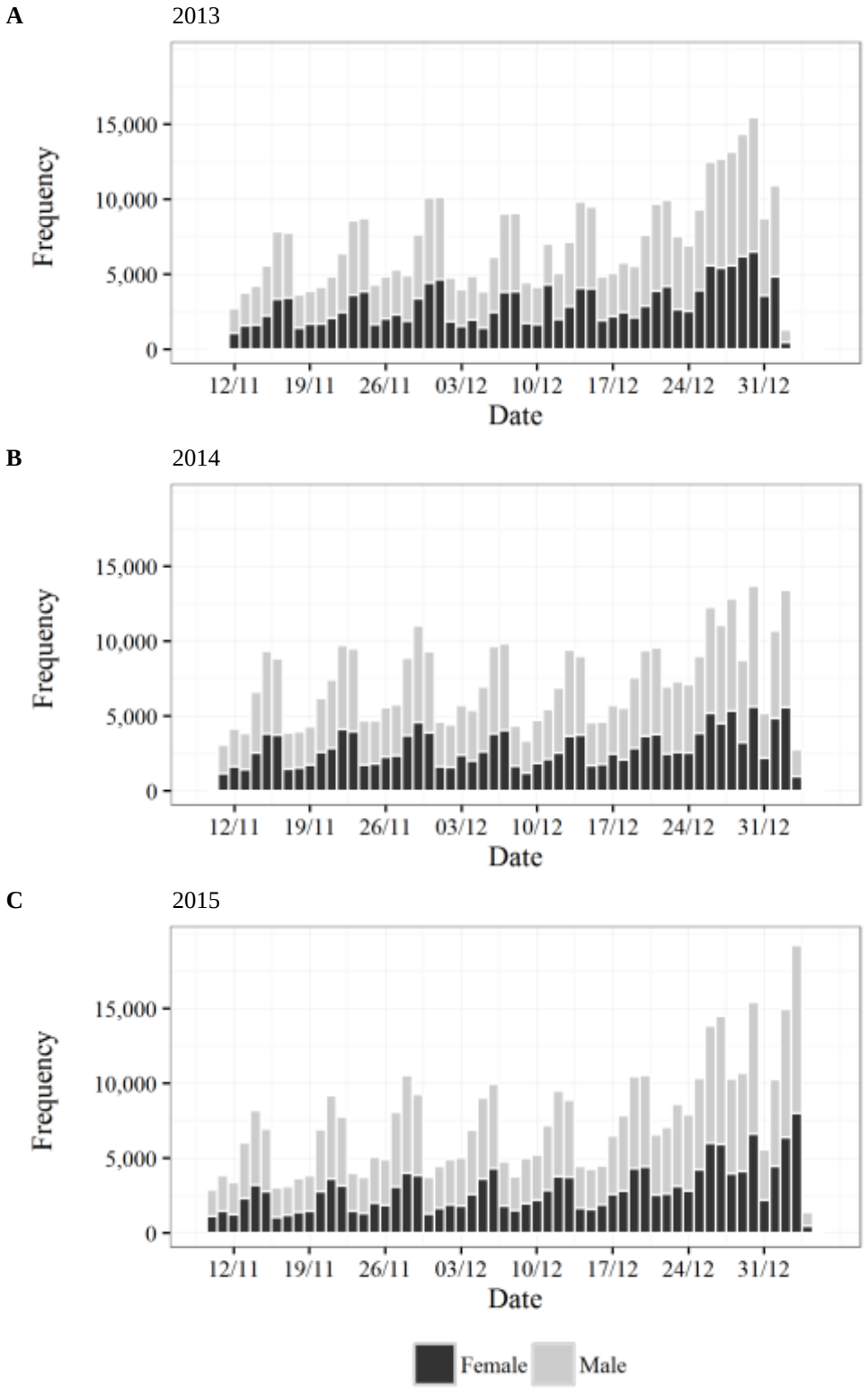
Supplementary Figure 1: Stimuli in the natural experiments. The three pictures display the champagne glass resembling plastic container that was filled with objects representing pearls in 2013 (picture A), pearls and diamonds in 2014 (B), and casino chips in 2015 (C).



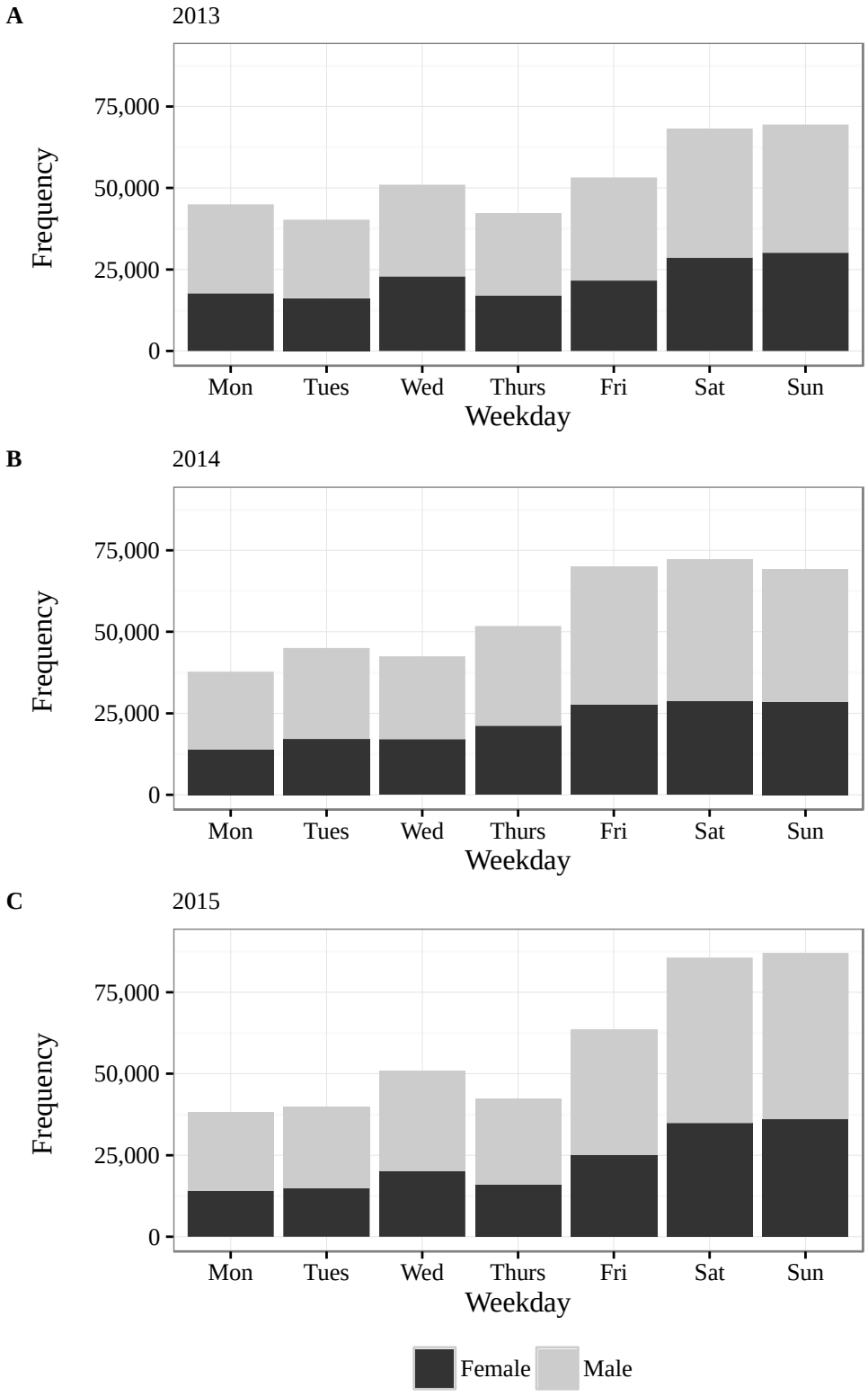
Supplementary Figure 2: Number of estimates per participant. The figure displays the distribution of the number of estimates provided by the 163,719 participants in the 2013 event (Panel A), the 154,790 participants in the 2014 event (B), and the 162,275 participants in the 2015 event (C).



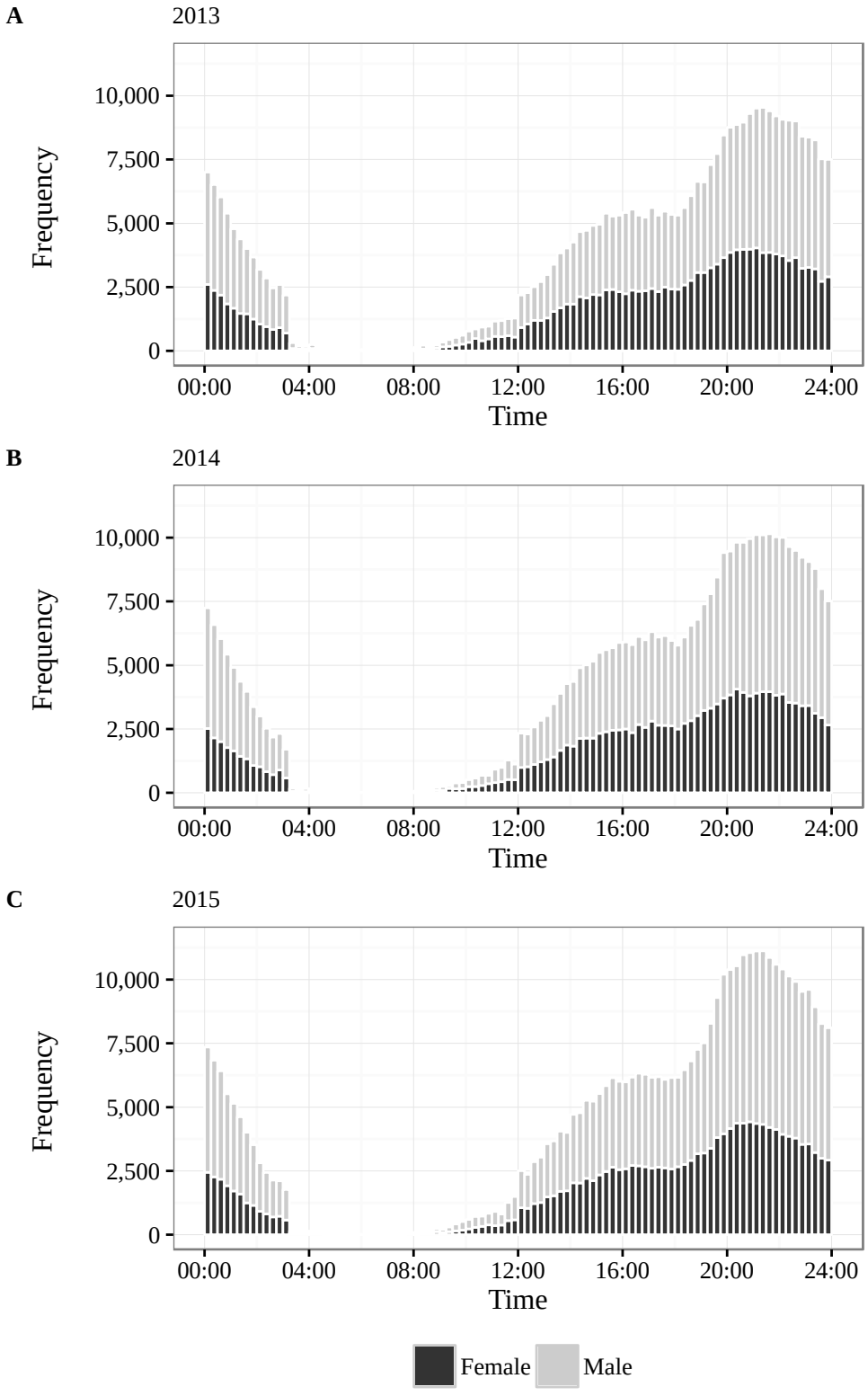
Supplementary Figure 3: Age. The figure displays the distribution of the age of the 163,719 participants in the 2013 event (Panel A), the 154,790 participants in the 2014 event (B), and the 162,275 participants in the 2015 event (C). Bars are split into segments by gender. Age is measured on January 16 after the event. The slightly abrupt decrease just below 70 years reflects the birth rate increase after World War II.



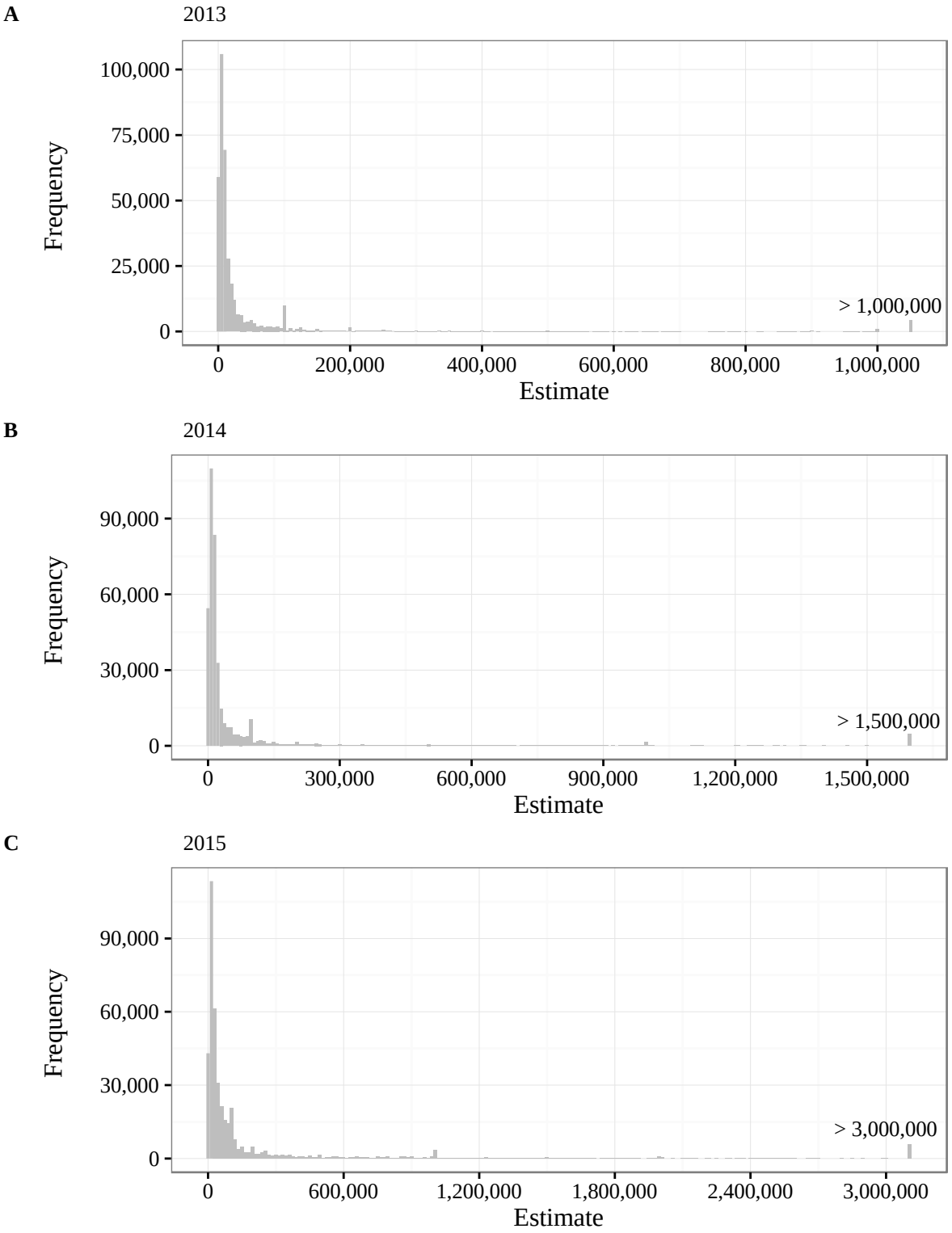
Supplementary Figure 4: Date. The figure displays the distribution of the submission date (day/month) of the 369,260 estimates in the 2013 event (Panel A), the 388,352 estimates in the 2014 event (B), and the 407,622 estimates in the 2015 event (C). Bars are split into segments by gender. Much of the variation reflects day-of-the-week effects (see Supplementary Figure 5).



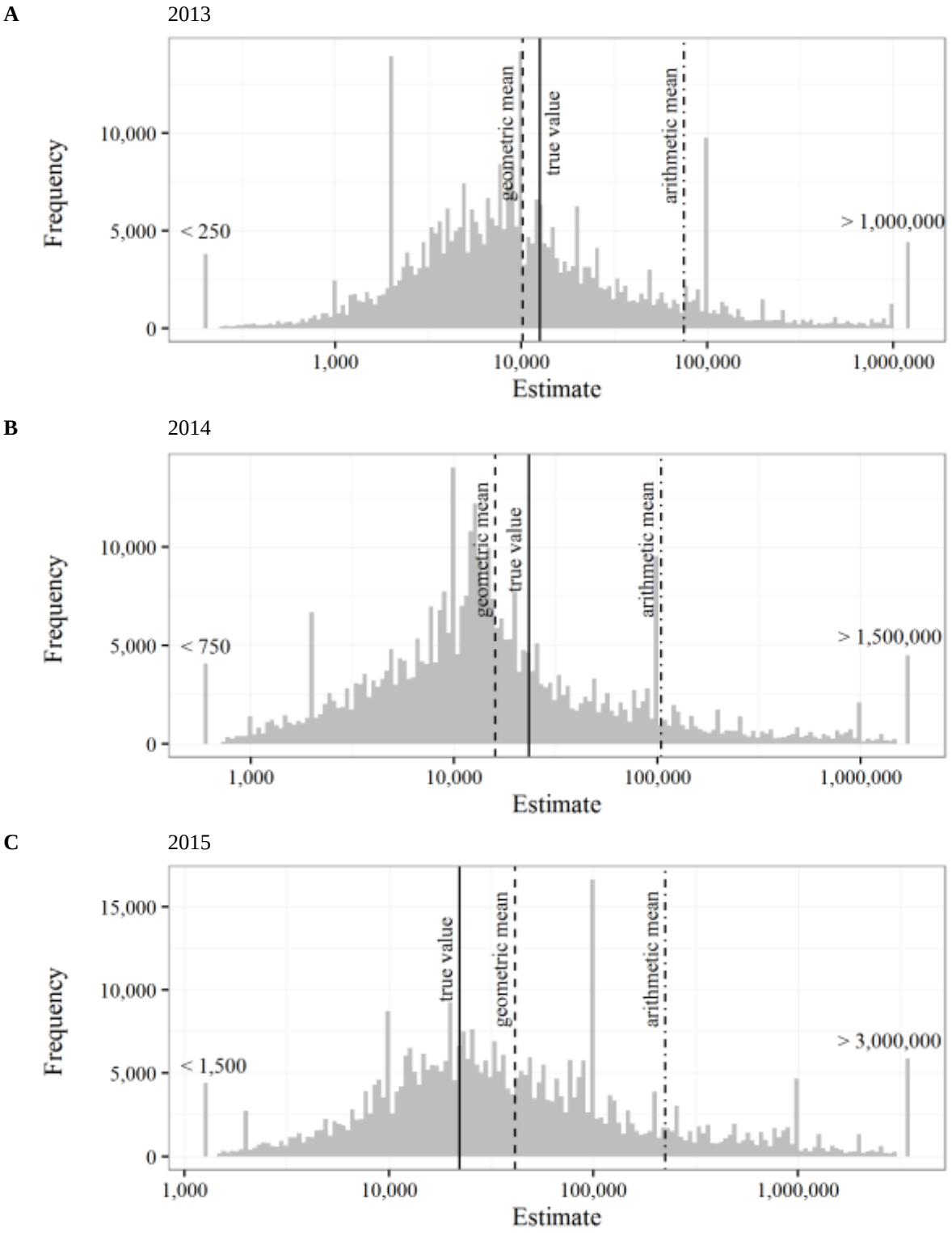
Supplementary Figure 5: Weekday. The figure displays the distribution of the day of the week for the 369,260 submitted estimates in the 2013 event (Panel A), the 388,352 submitted estimates in the 2014 event (B), and the 407,622 submitted estimates in the 2015 event (C). Bars are split into segments by gender.



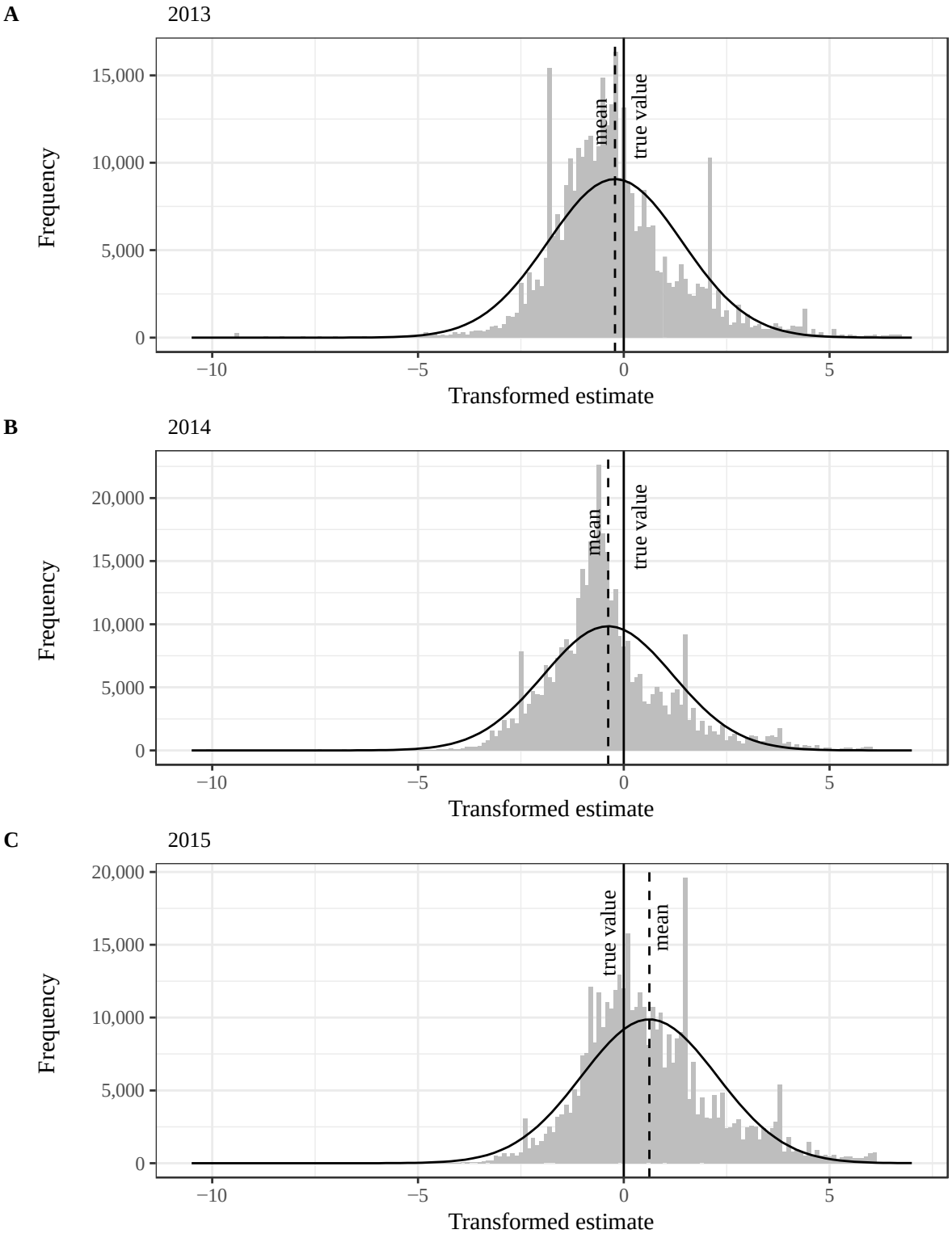
Supplementary Figure 6: Time. The figure displays the distribution of the time of the day for the 369,260 submitted estimates in the 2013 event (Panel A), the 388,352 submitted estimates in the 2014 event (B), and the 407,622 submitted estimates in the 2015 event (C). Bars are split into segments by gender. Outside opening hours, entries could not be submitted via the terminals inside the casino (only via the internet from elsewhere, explaining the abrupt changes in frequency around 03:00 and 12:00).



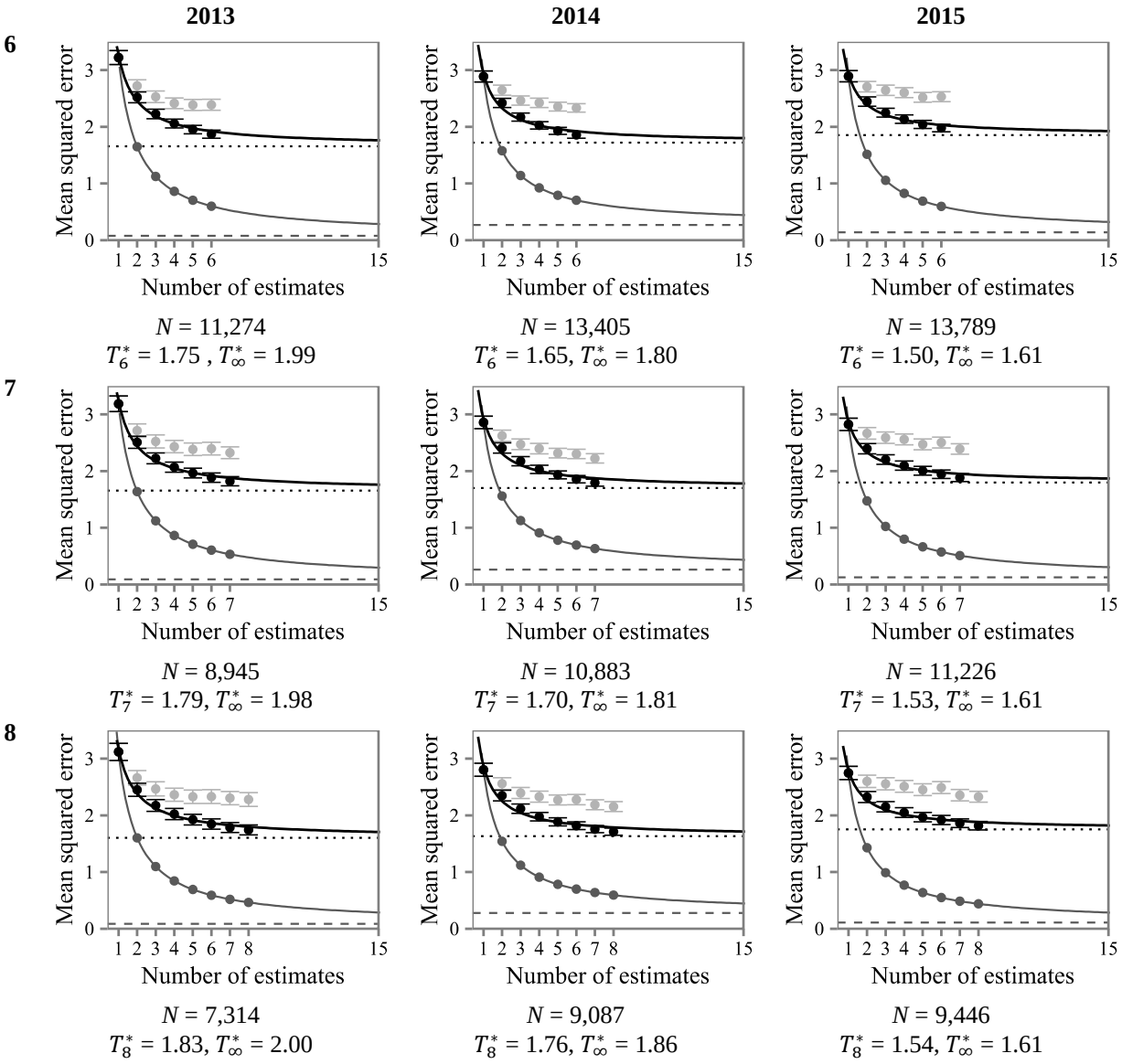
Supplementary Figure 7: Untransformed estimates. The figure displays the distribution of all 369,260 untransformed estimates in the 2013 event (Panel A), all 388,352 untransformed estimates in the 2014 event (B), and all 407,622 untransformed estimates in the 2015 event (C).



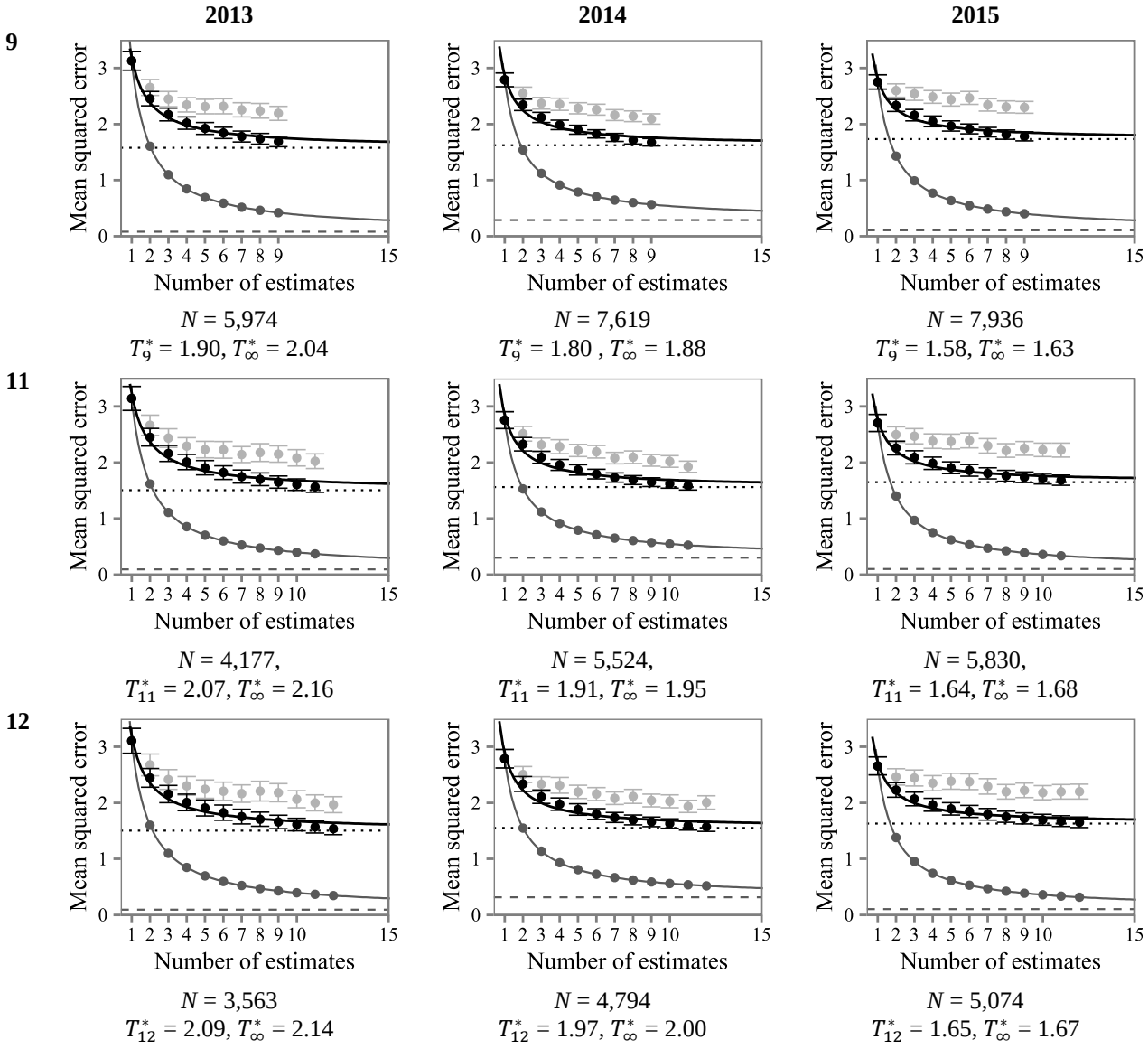
Supplementary Figure 8: Estimates on a logarithmic scale. The figures use a logarithmic scale to display the distribution of all 369,260 estimates in the 2013 event (Panel A), all 388,352 estimates in the 2014 event (B), and all 407,622 estimates in the 2015 event (C).



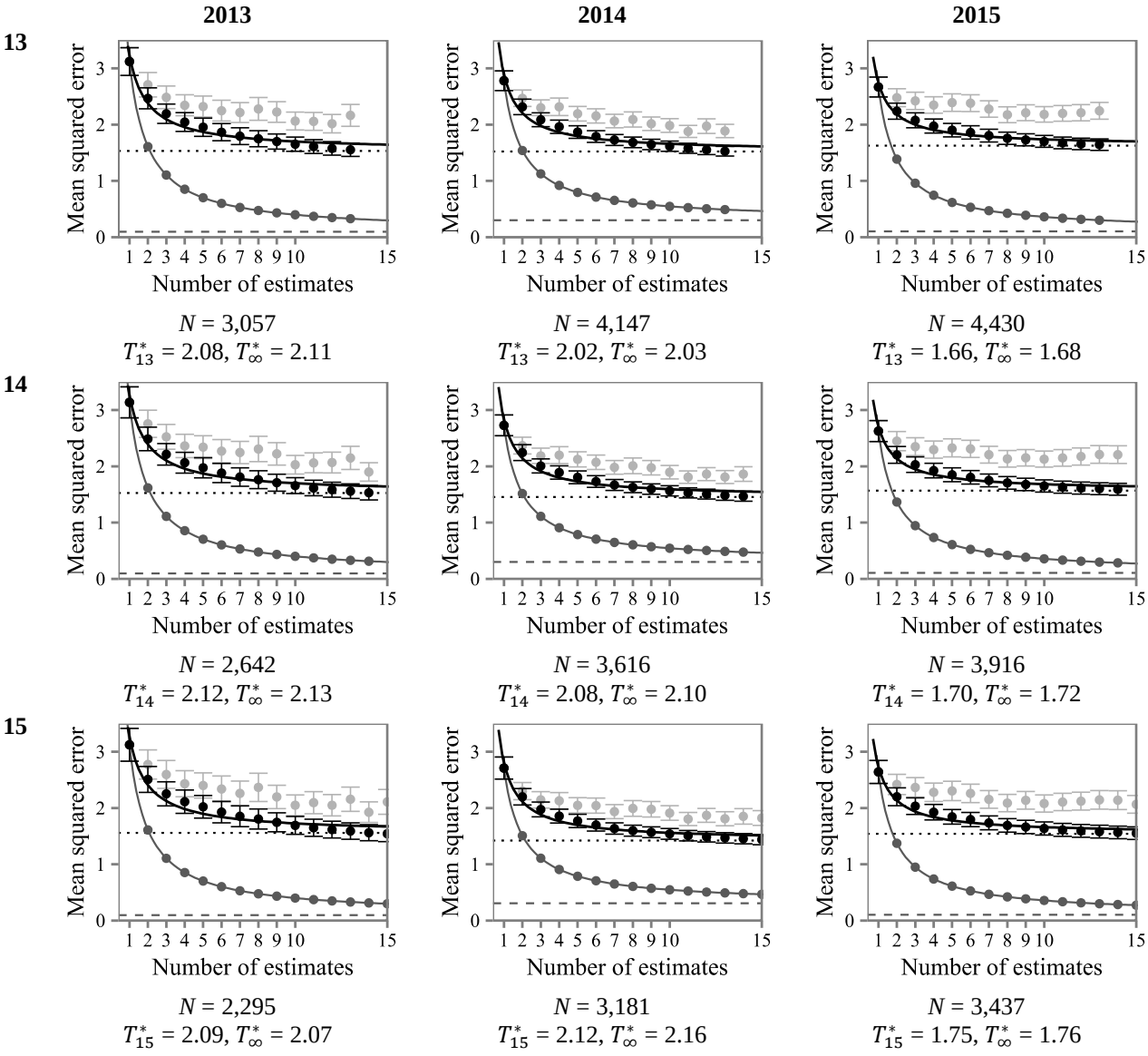
Supplementary Figure 9: Transformed estimates. The figure displays the distribution of all 369,260 transformed estimates in the 2013 event (Panel A), all 388,352 transformed estimates in the 2014 event (B), and all 407,622 transformed estimates in the 2015 event (C). Transformed estimates represent the logarithm of the ratio of the untransformed estimate and the true value. The plotted curves represent the best-fitting normal distribution.



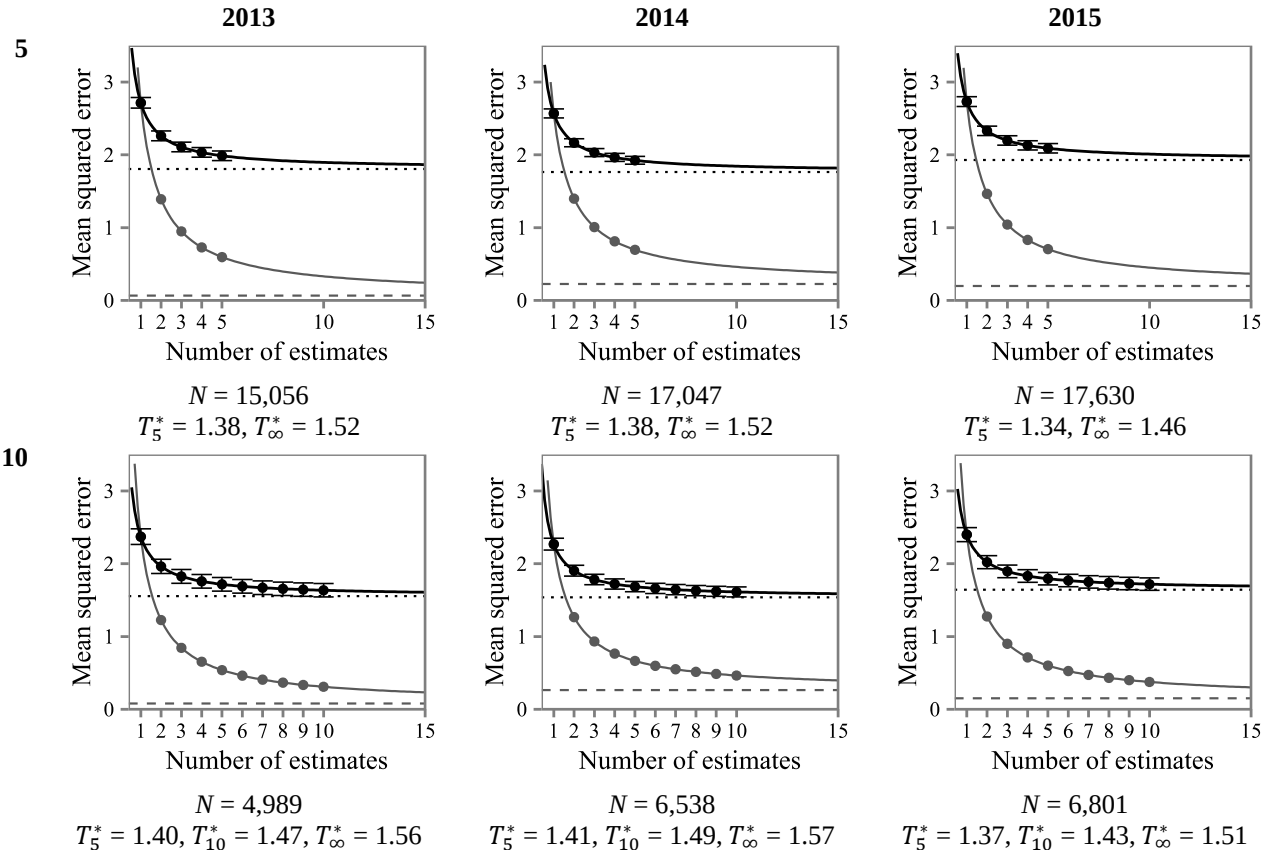
Supplementary Figure 10: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of included estimates. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for K = 6-9 and K = 11-15. The curve for the inner crowd represents the best-fitting hyperbolic function $MSE = a / t + b$ (using non-linear least squares); the dotted line represents b. Values for the outer crowd are mathematically determined using the diversity prediction theorem (see Methods); the dashed line represents the limit as the number of included estimates goes to infinity. The graphs also show the MSE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players. T_t^* is provided for $t = K$ and $t = \infty$ and equals the number of estimates one needs to average across individuals to achieve the same squared error as the squared error that results from averaging t estimates from a single individual. N is the number of players.



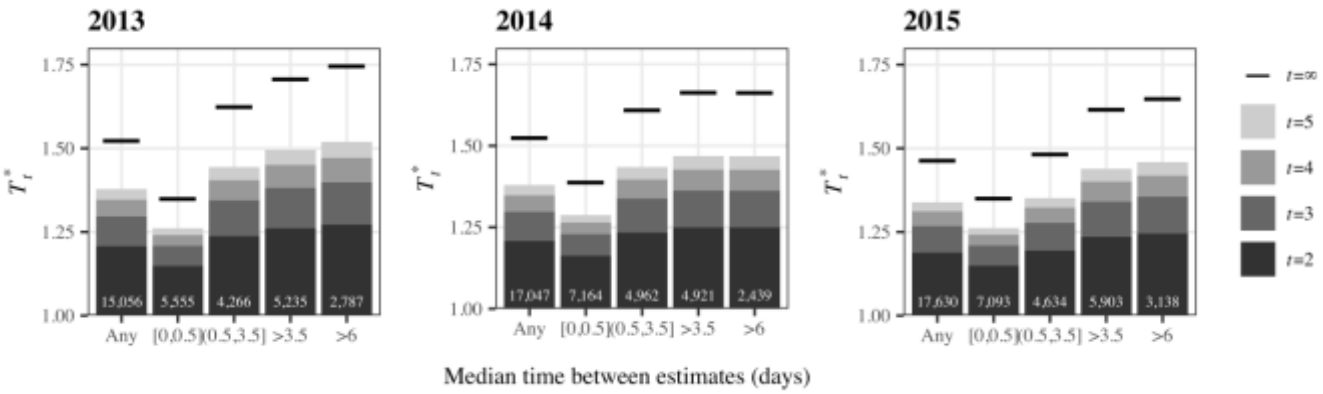
Supplementary Figure 10 (cont'd)



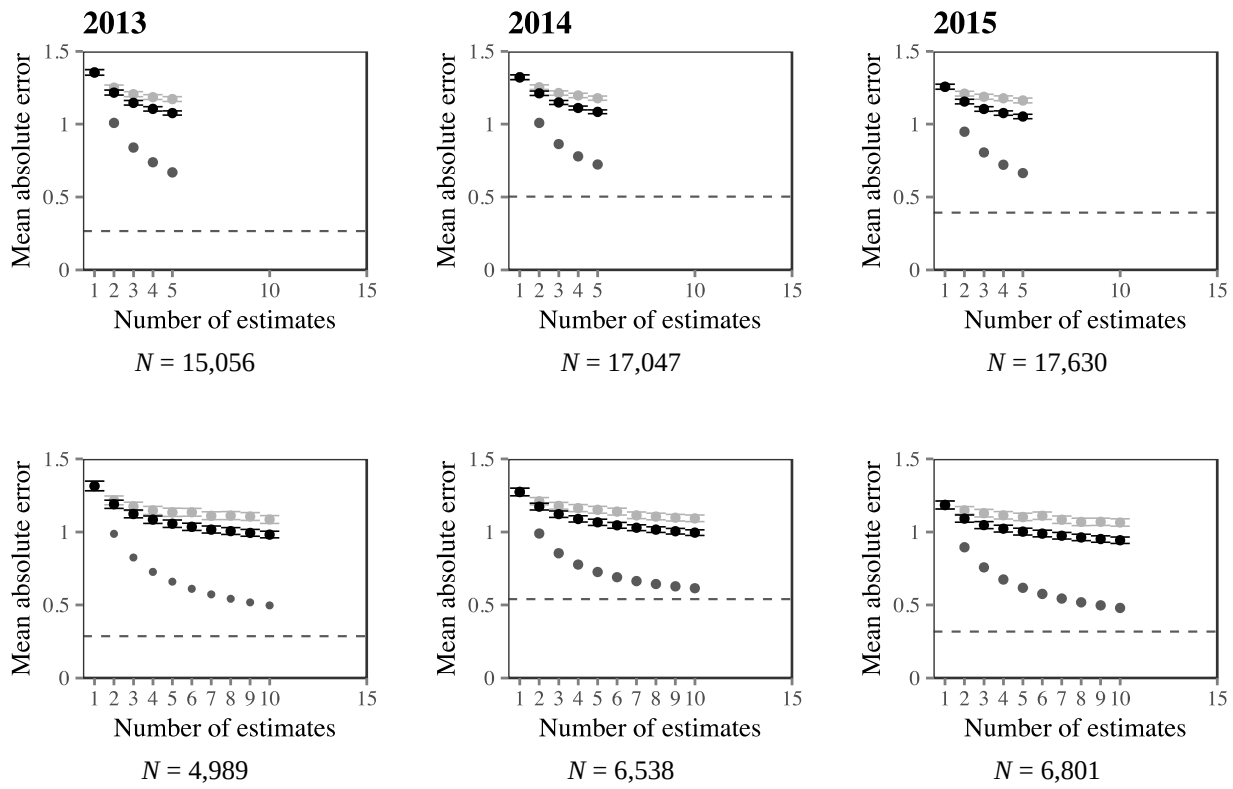
Supplementary Figure 10 (cont'd)



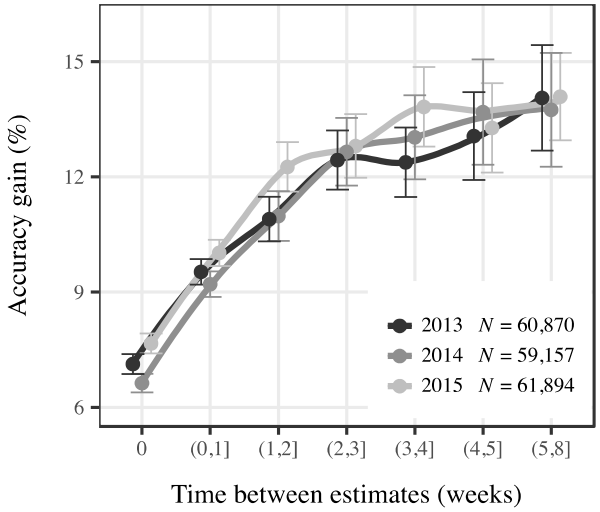
Supplementary Figure 11: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of included estimates (unordered). The upper (lower) graphs use the first *K* estimates of players who submitted at least $K = 5$ ($K = 10$) estimates in a given year. For the inner crowd results, estimates from the same person are aggregated in a random order. The outer crowd results are averages across the separate results of between-person aggregation for each of the *K* estimates. The curve for the inner crowd represents the best-fitting hyperbolic function $MSE = a / t + b$ (using non-linear least squares); the dotted line represents *b*. Values for the outer crowd are mathematically determined using the diversity prediction theorem (see Methods); the dashed line represents the limit as the number of included estimates goes to infinity. Error bars represent 95 percent confidence intervals. *N* is the number of players.



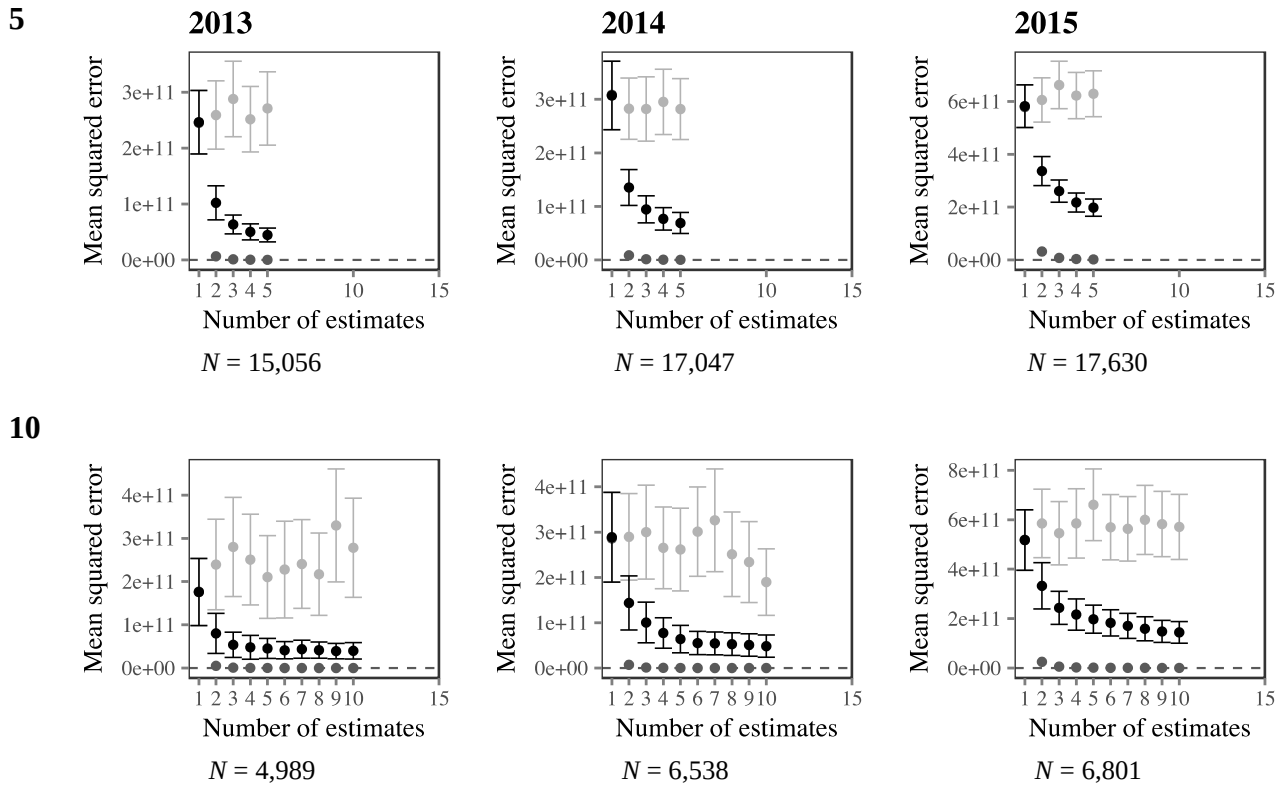
Supplementary Figure 12: Number of estimates T_t^* one needs to average across individuals to achieve the same squared error as the squared error that results from averaging t estimates from a single individual (unordered). The graphs use the first five estimates of players who submitted at least five estimates in a given year. Within-person aggregation uses estimates that are randomly selected from the five estimates. Between-person aggregation results are averages across the separate results of between-person aggregation for each of the five estimates. Results are shown for the full samples and for subsamples that differ in terms of the median time between the estimates. The numbers at the bottom of the bars represent the numbers of included players.



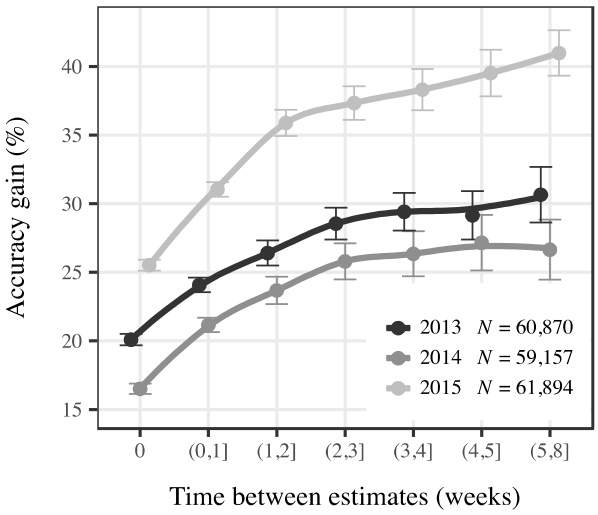
Supplementary Figure 13: MAE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of transformed estimates included. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MAE to which the outer crowd converges. The graphs also show the MAE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players.



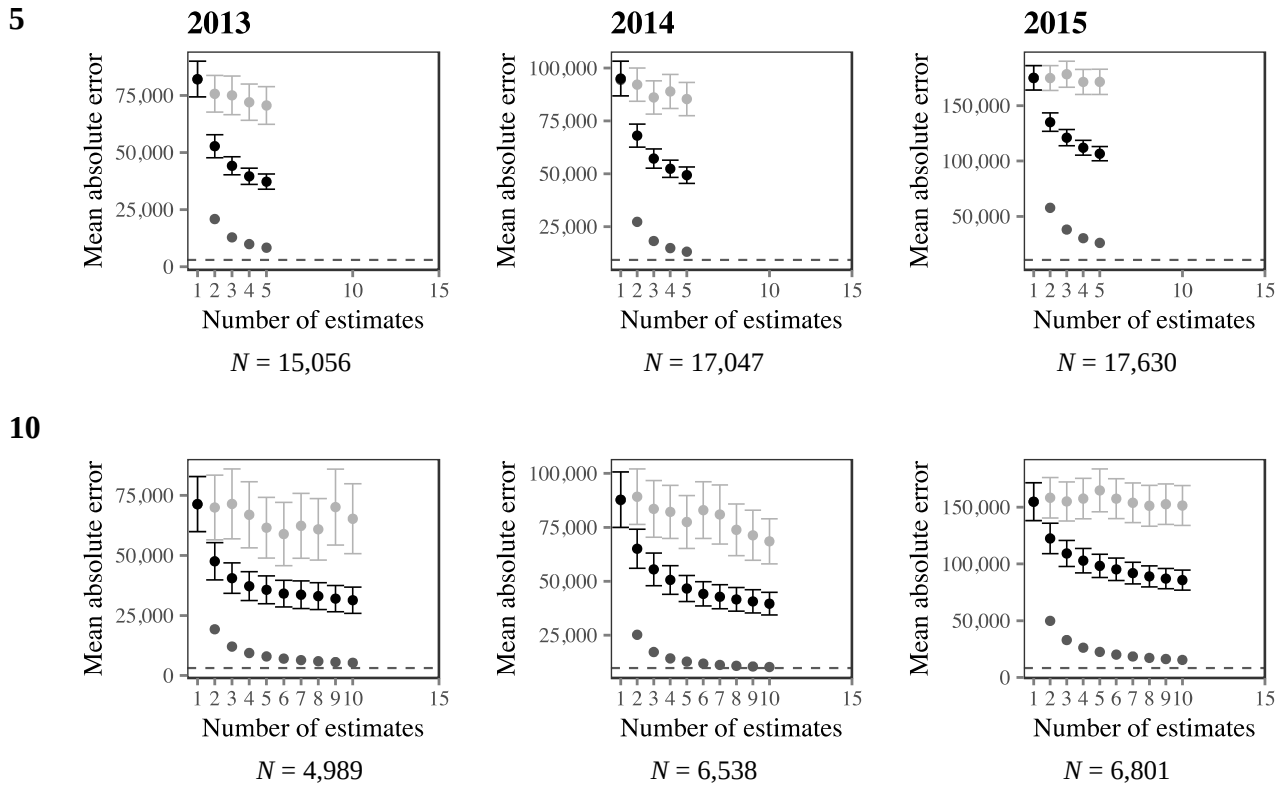
Supplementary Figure 14: Accuracy gain in terms of absolute error as a function of the time between the transformed estimates. Accuracy gain is defined as the decrease in absolute error obtained by aggregation (absolute error of the average of the estimates relative to the average absolute error of the individual estimates). Error bars represent 95 percent confidence intervals.



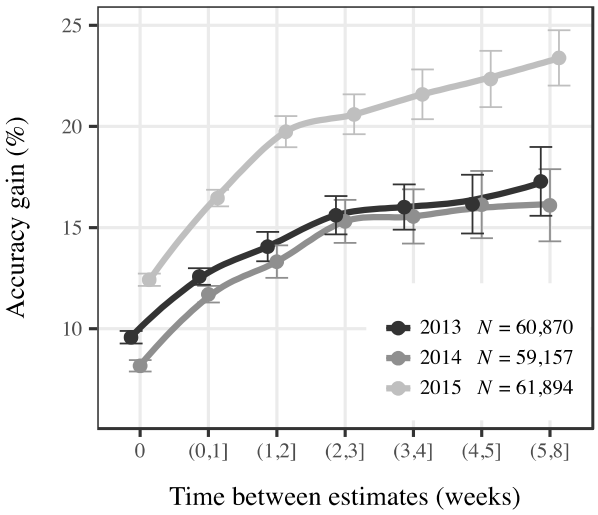
Supplementary Figure 15: MSE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of untransformed estimates included. The graphs use the estimates of players who submitted at least K estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MSE to which the outer crowd converges. The graphs also show the MSE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. N is the number of players.



Supplementary Figure 16: Accuracy gain in terms of squared error as a function of the time between the untransformed estimates. Accuracy gain is defined as the decrease in squared error obtained by aggregation (squared error of the average of the estimates relative to the average squared error of the individual estimates). Error bars represent 95 percent confidence intervals.



Supplementary Figure 17: MAE of the inner crowd (black) and the outer crowd (dark gray) as a function of the number of untransformed estimates included. The graphs use the estimates of players who submitted at least *K* estimates in a given year. Panels are shown for $K = 5$ and $K = 10$. Values for the outer crowd are determined by randomly combining estimates from different individuals 5,000,000 times. The dashed line depicts the MAE to which the outer crowd converges. The graphs also show the MAE of individual consecutive estimates (light gray). Error bars represent 95 percent confidence intervals. *N* is the number of players.



Supplementary Figure 18: Accuracy gain in terms of absolute error as a function of the time between the untransformed estimates. Accuracy gain is defined as the decrease in absolute error obtained by aggregation (absolute error of the average of the estimates relative to the average absolute error of the individual estimates). Error bars represent 95 percent confidence intervals.