"Worker Overconfidence: Field Evidence and Implications for Employee Turnover and Firm Profits": Online Appendix

Mitchell Hoffman and Stephen V. Burks

The Online Appendix is organized as follows. Appendix A provides additional discussion and analysis on belief survey non-response, as well as on other issues. Appendix B provides more information on the Firm B field experiment. Appendix C uses a one-period version of the structural model to briefly show that differential overconfidence lowers the probability of quitting. Appendix D provides omitted derivations from the structural model. Appendices E-F collect additional figures and tables. Appendix G provides further discussion on measuring productivity. Appendices here may differ slightly from the typeset version due to changes in typesetting.

A Additional Discussion and Results

A.1 Non-response to the Firm A Productivity Beliefs Survey

As discussed in Section 3 of the main text, the overall response rate to the Firm A beliefs survey was 28%, computed by averaging over all drivers and weeks (restricted to weeks where miles are greater than zero) in the estimation sample. Non-response could be a source of bias for our paper if non-response is correlated with overconfidence (or, more generally, if the true structural model is different for responders vs. non-responders). We provide qualitative and quantitative evidence that non-response bias is unlikely to be driving the paper's conclusions.

We suspect one reason our survey's response rate was only 28% is because the survey was entirely voluntary, with no communication at all regarding the survey from supervisors. We deliberately conducted the survey this way so as to eliminate any desirability pressure from supervisors. Although we may have gotten a higher response rate if supervisors had encouraged workers to participate, doing so could have risked contaminating worker response.

Each week, in addition to asking the productivity beliefs survey question, we asked a standard work happiness question.[1] The advantage of asking two questions each week instead of one is that it was less clear to drivers that the survey was "about" one topic or the other.

About 62% of data subset workers respond to at least one survey. Drivers who are white, male, older, and have higher schooling are more likely to respond each week, as seen in column 1 of Table A1. Among drivers who respond to at least one survey, there is significant variation in the average response rate, as seen in Figure E1.

Intuitively, what is driving the variation in driver response rates? Based on conversations with Firm A managers, we believe that differences in driver response rates may simply reflect that some people tend to be more likely to respond to surveys *in general* than others. Besides the weekly productivity beliefs survey, drivers in the data subset were also asked to participate in a number of other surveys: (i) a long computer survey during training on cognitive and non-cognitive skills and experimental preferences; (ii) "continuing driver surveys" mailed every 6 months (until the driver exits the firm) about driving conditions, traffic, work satisfaction, family life, and worker-supervisor relations; and (iii) our exit survey. Outside of the surveys administered by the researchers, which were clearly marked as such, drivers also received many queries from the firm over the same Qualcomm message system. (A Qualcomm is a very basic computer in trucks used for sending and receiving messages.)

A more worrisome scenario would be if driver response reflected overconfidence. A bias could presumably go in either direction, with more overconfident people either being more likely to respond to the survey (e.g., because it is more fun to respond if you expect to do better) or being less likely to respond to the survey (e.g., because overconfident people mis-predict more frequently and it is more embarrassing to respond). We believe that selection on overconfidence into survey-taking is unlikely because the workers took a number of different surveys and specific surveys seem unlikely to have been particularly salient to them. There is a strong correlation in whether a driver responds across most of the different surveys. For example, drivers who respond to the continuing driver surveys every 6 months (on traffic, driving conditions, etc.) have a significantly higher response rate to the weekly productivity beliefs survey.

Having provided intuition why we do not believe non-response is significantly biasing our results, we turn now to quantitative tests. To address non-random response, we use Inverse Probability

[1] The question wording was: "Overall, how happy are you with your job right now?" where 1=Very Unhappy, 2=Somewhat Unhappy, 3=Neutral, 4=Somewhat Happy and 5=Very Happy. In all cases, drivers either responded to both the productivity beliefs and happiness questions in a week, or didn't respond at all.

Table A1: How Do Driver Characteristics Predict Survey Response and Overconfidence?

Dep var:	(1) Survey Response	(2) Overconf	(3) Survey Response	(4) Survey Response	(5) Survey Response	(6) Overconf
Model:	OLS	OLS	OLS	OLS	Heckman	
L. Average productivity beliefs to date (in hundreds of miles per week)			0.001			
			(0.002)			
L. Average overconfidence to date (in hundreds of miles per week)				-0.004		
				(0.003)		
Response rate to continuing driver surveys					0.573	
					(0.127)	
ρ (correlation between error terms)						-0.058
						(0.257)
Black	-0.063	-24.261	-0.080	-0.078	-0.144	-304.933
	(0.034)	(91.907)	(0.044)	(0.044)	(0.222)	(147.600)
Hispanic	-0.255	-619.039	-0.280	-0.298	-1.174	-540.115
	(0.030)	(196.898)	(0.034)	(0.034)	(0.111)	(1,129.850)
Female	-0.056	-22.149	0.044	0.034	-0.268	62.407
	(0.036)	(129.783)	(0.045)	(0.048)	(0.225)	(186.843)
Married	0.041	-17.160	0.048	0.047	0.116	-16.835
	(0.022)	(62.615)	(0.025)	(0.026)	(0.113)	(79.443)
Age at a given time	0.004	0.224	0.004	0.004	0.016	0.495
	(0.001)	(2.316)	(0.001)	(0.001)	(0.005)	(4.573)
Years of schooling	0.010	-11.602	0.012	0.010	-0.001	-6.282
	(0.007)	(18.070)	(0.008)	(0.009)	(0.032)	(23.153)
Observations	28,039	8,121	21,397	19,877	7,836	

Notes: An observation is a driver-week. Standard errors in parentheses clustered by driver. Columns 1, 3, and 4 are linear probability models of whether a driver responds to the belief survey in a given week. Column 2 regresses overconfidence (defined here as the miles prediction made in t about $t+1$ minus actual miles in week $t+1$) on characteristics. Column 4 is restricted to observations where lagged average weekly overconfidence to date is greater than or equal to -1,500 miles and less than or equal to +1,500 miles. The variable is restricted to [-1,500mi, +1,500mi] to reduce the influence of outliers. Lagged average weekly overconfidence to date is calculated excluding zero mile weeks. Columns 5-6 present a Heckman selection model estimated by one-step full information maximum likelihood. The continuing driver surveys were given to drivers after 26 and 52 weeks of tenure. In order so that the survey response rate variable is defined for all observations in the regression, we restrict in columns 5 and 6 to observations after 52 weeks of tenure. Columns 5-6 have 2,172 observed driver-weeks of overconfidence and 5,664 driver-weeks where overconfidence is not observed. All regressions include week of tenure dummies and work type controls. All columns restrict to driver-weeks with positive miles in the current week, as well as positive miles in the subsequent week.

Weighting. There are two stages in estimation with Inverse Probability Weighting. In the first stage, we fit a probit model of whether a driver ever responds to the productivity belief survey as a function of time-invariant demographics (race, gender, years of schooling, and age at start of work), as well as dummies for having above-median IQ and above-median experimentally-measured patience.[2] None of these covariates are otherwise used in the estimation of the structural model. In the second stage, we use the inverse of the first stage predicted values to weight each driver's contribution to the likelihood, and then perform our main maximum likelihood estimation. For standard errors, we ignore estimation error from the first stage probit; doing so leads to conservative standard errors (Wooldridge, 2002). As seen in column 3 of Table F1 (the table containing the various robustness checks for the structural estimates), our main structural results are quite robust to Inverse Probability Weighting. Thus, although certain types of people are more likely to respond to the survey than others, this appears to have little impact on the structural estimates.

The identifying assumption for Inverse Probability Weighting is that survey response is missing at random conditional on the observable characteristics used in the first stage probit model. Beyond standard demographics, it is possible that unobserved characteristics could affect non-response. However, a large advantage of our data is that we have a great deal of additional information about people that could potentially affect their response beyond standard demographics, including cognitive ability, non-cognitive ability (personality traits), and experimental measures of preferences. For example, one might think that people would less likely to respond if they have a low IQ, are generally impatient, or have low numeracy; our data allow us to convert what are usually unobserved characteristics into observed ones. We checked that our results are robust to Inverse Probability Weighting using different combinations of first-stage variables, including several of these richer characteristics.

Even controlling for these very rich characteristics, one could still be concerned that non-response is being driven by unobservables. As mentioned before, we would overstate the amount of overconfidence if people who were more overconfident were more likely to respond to the survey. It is difficult to assess this argument directly given that overconfidence is not observed when people do not respond to the survey. However, we can examine whether there is any correlation between lagged average beliefs to date (or lagged average overprediction to date) and response to the survey. There is no significant relationship between average productivity beliefs to date and survey response (column 3 of Table A1), or between average overprediction to date and survey response (column 4 of Table A1). These are precisely estimated zero coefficients.[3]

To formally test whether belief response is occurring based on unobservables, we analyze a Heckman (1979) ("Heckit") model. We use whether drivers respond to prior surveys on topics other than productivity beliefs as a basis for a plausible exclusion restriction. (This strategy, of using response on other surveys to estimate a Heckman (1979) selection model of response on a different survey, is used in prominent papers such as Choi et al. (2014).) The identifying assumption is that whether a driver responds to other surveys influences whether the driver responds to the productivity beliefs survey, but not the outcome variable of interest. Columns 5-6 of Table A1 estimate a Heckman (1979) model for overconfidence, using a driver's average response rate on the 6-month and 12-month continuing driver surveys as the exclusion restriction variable.[4] Column 5 shows that drivers who

[2] Burks, Carpenter, Goette, Monaco, Porter, and Rustichini (2008) detail the data collection effort at Firm A.

[3] For example, in column 3, the 95% CI for the coefficient on lagged average beliefs to date (in hundreds of miles) is [-0.0025, 0.0040]. Thus, we can rule out that a 300 mile change in average productivity beliefs would decrease survey response by more than 0.75 percentage points or increase survey response by more than 1.2 percentage points.

[4] Thus, our survey response rate variable is 0, 0.5, or 1 for each driver (the average is 0.44). Because we are estimating pooled cross-sectional models (without individual fixed effects), it is fine that our exclusion restriction variable does not vary within person. For the survey during training, almost all (over 90%) of trainees invited to participate chose to participate; those who chose not to participate were not sent productivity belief surveys. Thus, we cannot use the survey during training in the Heckit model. We also do not use the exit survey in the Heckit model, as the exit survey

respond to the continuing driver surveys are substantially more likely in subsequent weeks to respond to the productivity belief survey. However, our estimate of the error correlation coefficient, $\rho = -0.06$, is economically small and is statistically indistinguishable from zero. We interpret this as evidence that non-response bias does not drive simple models of overconfidence.[5]

A final piece of support that non-response is not driving our results on overconfidence comes from Figure 1. Figure 1 shows that patterns of overprediction and learning are similar in a sample of all subjects and in a sample of workers responding to the survey.

A fully structural approach to address non-response would be to explicitly model the choice every week or whether to respond to the survey. Given our various "reduced form" tests suggesting that non-response bias is limited, we conjecture that structurally modeling the response decision would have little impact on our findings. In addition, doing so would substantially increase the computational complexity of the model.

A.2 Further Discussion on Differential Overconfidence

We believe the assumption of differential overconfidence is reasonable in our setting. What would happen to our paper's results, however, if a worker was substantially overconfident about both his inside and outside options?

Regarding model fit, the possibility of workers being overconfident about the outside option can be accommodated in the model by making r represent someone's perceived outside option instead of their actual outside option. Column 6 of Table F1 shows our estimates are very similar assuming a higher outside option.[6]

In terms of counterfactuals, it becomes useful to distinguish two exercises: eliminating inside overconfidence vs. eliminating both inside and outside overconfidence. When workers exhibit substantial overconfidence about the outside option, eliminating inside overconfidence will still make workers more likely to quit and would presumably increase profits from training. Unlike in the paper, it has the potential to reduce worker welfare. In contrast, eliminating both inside and outside overconfidence may have little impact on quits, profits, and welfare.

A.3 More Details on Data and Sample Construction

Data Subset. As described in footnote 9 in the main text, we restrict our sample to drivers with a code denoting no prior trucking experience or training. Beyond eliminating drivers with prior experience or training, it also eliminates drivers who did our survey, but then failed to complete the Firm A training.

Teams. In a modest share of driver-weeks, drivers work in two-person teams. For example, one worker sleeps while the other one drives. In the data subset, about 13% of worker-weeks feature a worker driving with another driver. For team drivers, in the payroll data provided to us, Firm A equally splits total miles driven between the two workers. As part of our work type controls, we control for whether a worker is a team driver. About 40% of driver-weeks with productivity beliefs at or above 4,000 miles are from team drivers. Excluding team driving weeks, and re-doing the results

was administered after the worker left the company (and thus was not prior to any productivity belief surveys).

[5]The standard error on ρ is relatively large at 0.26, meaning we cannot rule out from the Heckit model alone that there could be positive or negative selection based on overconfidence into survey response. Thus, while the Heckit model results are *suggestive* of limited non-response bias on their own, they buttress the multiple pieces of evidence in Appendix A.1 that non-response is not biasing our main results.

[6]It is important to note that mean bias, η_b, in the prior has a tenure-varying impact on quitting and subjective beliefs.

on beliefs/miles over time (Figure 1) and the relationship between beliefs and quitting (Table 3), the results are qualitatively similar.

Missing Indicators. A small number of drivers have missing data for race (1%), gender (1%), and marital status (0.2%). We set missing values to 0, and we include dummy variables for the variable being missing as part of the demographic controls.

A.4 Worker Credit Scores

As described in Section 2.2, Firm A drivers have very low average credit scores. The credit score is the FICO-98 and ranges from 300 to 850. 53% of drivers have a credit score below 600, compared to only 15% of the US general population. What credit score indicates a "subprime" borrower is not absolute and has varied, but the cutoff is often a value somewhere in between roughly 600 to 650 such as 620 (e.g., Rustichini et al., 2016). Thus, the majority of drivers in the sample would likely be considered subprime borrowers. Furthermore, 43% of drivers have scores below 550 compared to only 7% of the US population. The credit score statistics on the US general population are from the "Report to the Congress on Credit Scoring and Its Effects on the Availability and Affordability of Credit" issued by the Federal Reserve Board of Governors (August 2007).

A.5 Exit Survey

There were 8 possible responses in the survey: Over-the-Road long haul, Over-the-Road regional, Driving locally, Nondriving job, Unemployed, Disabled, Retired, or Other. We provide our numbers in the text ignoring the 7% of responses given as Disabled, Retired, or Other. We ignore these categories because they may be different from other types of exits, but the percentages in the text are similar if these categories are included. Regional drivers deliver loads in a particular region. Like long-haul drivers, regional drivers are also usually paid by the mile, so ability to get miles may transfer from long-haul to regional. Still, drivers who are best at long-haul need not be the same one who are best at regional work.

One concern about doing an exit survey is that workers may lie about where they went next. For our survey, however, it was repeatedly emphasized to drivers that their responses were anonymous and would never be seen by the company, presumably eliminating incentives to lie. Another concern is that drivers who respond to the exit survey may be non-representative, as the response rate on the exit survey was only about 25%. However, whether a driver responded to the exit survey is uncorrelated with average productivity and most demographics, suggesting the results from the exit survey are unlikely to be biased by non-response. The one significant predictor of response is that older drivers were more likely to respond, with an additional 10 years at age of hire associated with a 6 percentage point increase in the probability of responding. We do not think this will bias our findings, as age is not significantly correlated with whether a driver reports moving to a long-haul job (or to either a long-haul or regional job).

Our purpose in the exit survey is to examine the share of drivers who are leaving for some different type of work. One limitation we face in using the exit survey for this purpose is that while most drivers in our data period at Firm A are doing long-haul work, a small share are doing work that is more correctly thought of as regional. Because of this, the share of workers who are moving to the same type of work is probably higher than 12%. However, based on understanding of Firm A, a considerable majority of Firm A drivers in our data are leaving for other types of work, and we control for work type in the various regressions.

A.6 Predicting Beliefs using Average Productivity to Date

Column 3 of Appendix Table E4 shows that, in predicting beliefs, the weight on lagged average productivity to date increases with tenure. Section 3.1 of the main text states that this is consistent with simple Bayesian updating, and this section provides a simple proof. Assume a worker behaves as in the model in Section 4 of the main text, but the worker has no belief bias (i.e., $\eta_b = 0$ and $\widetilde{\sigma_y} = \sigma_y$) and there is no learning by doing.[7] Consider regressions of the form:

$$b_{it} = \alpha_t + \beta_t \overline{y}_{it-1} + \varepsilon_{it}$$

which are run separately for each week of worker tenure t. Note further that:

$$\begin{aligned} \beta_t &= \frac{cov\left(b_{it}, \overline{y}_{it-1}\right)}{var\left(\overline{y}_{it-1}\right)} \\ &= \frac{cov\left(\left(1-\gamma_{t-1}\right)\eta_0 + \gamma_{t-1}\overline{y}_{it-1} + \epsilon^b_{it}, \overline{y}_{it-1}\right)}{var\left(\overline{y}_{it-1}\right)} \\ &= \gamma_{t-1} \end{aligned}$$

where $\gamma_{t-1} = \frac{(t-1)\sigma_0^2}{(t-1)\sigma_0^2+\sigma_y^2}$ and ϵ^b_{it} is classical measurement error in stated beliefs. Thus, $\frac{\partial \beta_t}{\partial t} > 0$.

A.7 Further Discussion on Firm A Belief Elicitation

Day of Week for Survey. Other than Tuesday, the day was Monday in 7% of driver-weeks, Wednesday in 4% of driver-weeks, and Thursday in 3% of driver-weeks. These percentages are among the dates when there is a date in the data indicating when the survey was sent. In 24% of driver-weeks in our sample, the date of survey sending is missing.

High Belief Values. Subjective belief predictions contain a small number (128 observations) of very high values. We surmise that many of these may be driver typos, where a driver accidentally added an extra 0 to the end of their prediction. Thus, for predictions greater than 10,000 miles and less than 50,000 miles (100 observations), we divide these numbers by 10. All other belief predictions at 10,000 miles and higher are trimmed and converted to missing. Appendix Table F1 show that our main structural estimates are qualitatively robust to Winsorizing high belief values at 4,000 miles.[8]

Lumpy Beliefs. As is common in data on subjective beliefs, the responses given by subjects exhibit lumpiness (e.g., Zafar, 2011). Specifically, as suggested by part (b) of Figure E3, drivers' subjective beliefs are usually multiples of 100 miles, and are often multiple of 500 miles (about 60% of responses are multiples of 500 miles). A possible concern is that subjects could be "rounding up" and this could be contributing toward observed overconfidence.

We do not have any particular reason to believe that subjects would be more likely to round up than to round down, especially given it was emphasized to the subjects that only the researchers (and not the company) would observe their participation. Further, we do not believe that belief lumpiness is driven by lack of incentives, as lumpiness is also observed in the incentivized beliefs data from Firm B. Still, to avoid concerns that subject could be rounding up to the nearest 500 miles, we re-did our main results on beliefs (in Figure 1, Table 2, and Table 3) excluding observations where predicted miles are a multiple of 500 miles, and all conclusions are robust (in fact, the results in Tables 2 and

[7] The derivation is very similar if workers have biased beliefs, with $\beta_t = \gamma^b_{t-t} = \frac{(t-1)\sigma_0^2}{(t-1)\sigma_0^2+\widetilde{\sigma_y}^2}$ and $\frac{\partial \beta_t}{\partial t} > 0$.

[8] This Winsorization is done using the data as processed above, where high values are dropped and likely typos corrected.

3 become stronger). We do not have power to exclude cases where beliefs are multiples of 100, but even if subjects rounded up to the nearest 100 miles, this would only explain a modest portion of the substantial observed overprediction. Thus, we do not believe that lumpiness of beliefs is driving our findings.

In the structural model, we model reported subjective beliefs as equal to true subjective beliefs plus normally distributed error. Given the lumpiness in reported subjective beliefs, this assumption may be violated. However, we not think this is important for our main findings. The mean bias term, η_b, will be identified by average differences between predicted and actual miles. It seems that as long as average reported beliefs do not differ from average underlying subjective beliefs, mis-specfication of the error term in reported beliefs will not affect the conclusions of a counterfactual where we eliminate average underlying overconfidence from the population.

Alternatives to Belief Elicitation Method. Instead of asking for a point-estimate, another method of belief elicitation would to have asked truckers for their subjective productivity distribution at every week (e.g., "what is the chance you will run between X and Y miles next week," varying X and Y to span the whole distribution), as has been done in the pioneering work of Charles Manski and colleagues (Manski, 2004). We chose our approach of asking for point estimates because of our desire to reduce survey time burden for the two-year weekly study and out of desire to keep questions simple for drivers (many of whom have only a high-school degree).

A.8 No Effort in the Model

As mentioned in footnote 18, we speculate that including effort in the model would not qualitatively affect our main conclusions or would actually strengthen them. For example, suppose that there was complementarity between effort and perceived ability. Under this assumption, our main counterfactual of eliminating worker overconfidence in Section 5 would have an additional downside for firms of reducing worker effort. One potential modification of our conclusions would come if overconfidence was useful for agents in setting goals or overcoming self-control problems (Benabou and Tirole, 2002). To the extent that effort and overcoming self-control via overconfidence were important, this would seem likely to reduce the calculated worker welfare gain from debiasing in Section 5.

A.9 Evidence from Psychology and Behavioral Economics for Assumption of Differential Overconfidence

As discussed in Section 4.2, a central assumption in the structural model is that the worker exhibits differential overconfidence. That is, the worker must be more overconfident about his inside option than his outside option. We describe here how this assumption is consistent with work in psychology and behavioral economics. In the psychology literature, Moore and Swift (2010) and Moore and Healy (2008) show that different measures of overconfidence are only weakly related to each other and to various individual characteristics. In addition, differential overconfidence is consistent with cognitive dissonance, the tendency of people to receive discomfort when holding contradictory beliefs (Festinger, 1957). Cognitive dissonance theory (Festinger, 1957; Akerlof and Dickens, 1982) suggests that workers may be averse to beliefs that are inconsistent with having made good decisions. For a worker who has invested substantial time and effort to train with Firm A (and who has also incurred a financial obligation to stay), it may be mentally difficult to believe one is a bad match with Firm A. In cognitive dissonance models (e.g., Mayraz, 2011), as well as in related models with taste for consistency preferences (e.g., Eyster, 2002), a worker may come to believe that being with Firm A gives him the highest earnings.

A.10 Additional Structural Robustness Checks

Beyond the robustness checks described in Section 4.5, we have also performed additional robustness checks that we omitted from the main text for ease of exposition.

Heterogeneity in Overconfidence. In an earlier version of the paper, to allow for heterogeneity in overconfidence, we estimated the model allowing for mass point heterogeneity in η_b. Allowing for such heterogeneity, the impacts of our debiasing counterfactual on profits and attrition were somewhat smaller, though the conclusions from the simulation were substantively unchanged. In addition, adding heterogeneity in η_b tended to increase our estimate of τ, as well as our estimate of $\widetilde{\sigma_y}$, which amplifies our conclusion that $\widetilde{\sigma_y}$ differs from σ_y (i.e., learning is slower than predicted by Bayes' Rule). Our preferred model without overconfidence heterogeneity matches key data patterns and is computationally simpler.

Further Robustness Checks. In addition, we have (1) Estimated with finer and coarser discretizations of miles (as suggested by Rust, 1987); (2) Eliminated subjective beliefs greater than 4,000 miles instead of Winsorizing them; and (3) Assumed the taste heterogeneity is normally distributed instead of mass point distributed. The estimates are generally robust to these checks. Eliminating subjective beliefs greater than 4,000 miles decreases the estimated mean belief bias to 538 miles (instead of 674 miles in the baseline and 614 miles when Winsorized at 4,000 miles).

A.11 Out-of-Sample Fit

In Hoffman and Burks (2017), the structural model and parameters used for simulation are slightly different from those in the present paper. However, the baseline model and parameters in the present paper can also predict some basic retention patterns under out-of-sample contractual regimes.

A.12 Additional Information on the Counterfactual Simulation

We further discuss our definition of profits, as well as our assumption on contract enforcement.

A.12.1 Profits

We make a number of simplifications in our calculation of profits. In particular, beyond not including firing decisions, we ignore a number of components of profits, including vacancy costs, hiring costs (including recruiting costs and any hiring bonuses), employee referral bonuses, trucking accident costs, non-mileage driver pay (including driver bonuses), and driver benefits. Instead, we simply make an assumption on the overall fixed cost per week. It would be difficult and taxing for the model to try to model all of these different components, some of which we have only limited data on. Not separately modeling these different components should not affect the conclusions of our counterfactual analyses unless these interact in some way with the counterfactuals. The general conclusions of the counterfactuals, however, seem robust to different assumptions.

For the weeks of on-the-job training, we assume that firm profits are -$375 (i.e., minus one times assumed flat salary training pay).

A.12.2 Worker Welfare

In the week that a driver quits, we assume that worker utility is determined by the outside option instead of miles run on the job, following the model. We do this even though many drivers have

substantial miles in the week of quitting.[9] Calculated worker welfare is very similar if instead we assume that worker welfare is determined by miles and taste for the job in the week of quitting.

A.12.3 Contract Enforcement

As discussed in the text, roughly 30% of quit penalties were collected at the firm (Hoffman and Burks, 2017). In terms of how the collection rate matters for the paper's results, the collection rate, θ, does not enter into the estimates of our structural parameters or affect worker welfare (given the assumption that the worker experiences the full utility cost of the contract upon quitting, as discussed in footnote 23 in the main text). For our main counterfactual of debiasing, we see qualitatively similar impacts on profits with $\theta = 0.1$ and $\theta = 0.5$.

A.13 Other Papers Estimating Structural Models using Subjective Beliefs Data

In the introduction, for brevity, we provided only a very limited discussion of the small, but growing literature using subjective beliefs to estimate dynamic structural models. Here, we describe additional papers. For example, in pioneering papers, van der Klaauw and Wolpin (2008) and Chan, Hamilton, and Makler (2008) estimate structural models of worker retirement and managerial decision-making, respectively. Pantano and Zheng (2010) use subjective beliefs to relax assumptions about unobserved heterogeneity. van der Klaauw (2012) provides a general analysis of using subjective expectations to estimate structural models, which he illustrates using teacher career decisions. Stinebrickner and Stinebrickner (2014) use beliefs about grades to estimate a structural model of the college dropout decision. Wang (2014) estimates a structural model of smoking. Arcidiacono et al. (2012) and Wiswall and Zafar (2015) use subjective beliefs to estimate structural models of college major choice. Arcidiacono, Hotz, Maurel, and Romano (2014) use subjective beliefs to analyze a 3-stage model of occupational choice. There are also papers using subjective beliefs to estimate static structural models, e.g., Bellemare, Kroger, and van Soest (2008); Delavande (2008); Hendren (2013). Zafar (2011) collects subjective beliefs regarding majors, grades, and earnings to analyze whether beliefs data should be used in choice models.

A.14 Other Papers Using the Firm A Data Subset

As mentioned in Section 2.2, several other papers have used data from the Firm A data subset (also called the "New Hire Panel" in the other papers) to study various topics. Burks et al. (2009) examine whether worker cognitive skills predict experimental measures of worker preferences, worker strategies in experimental games, and worker retention.[10] Rustichini et al. (2016) examine whether measures of trainee personality predict experimental measures of worker preferences, worker strategies in experimental games, health behavior, worker retention, and worker accidents. Anderson et al. (2013) compare measures of social preferences between truckers, students, and non-trucker adults.

In an unrelated paper written after ours, we combined the entire Firm A data with data from 8 other firms to study differences across workers in terms of whether they were hired through a referral from an incumbent employee (Burks, Cowgill, Hoffman, and Housman, 2015). Additionally controlling for referral status does not affect the findings of our main reduced form tables (i.e., on whether productivity beliefs predict productivity (Table 2) or whether productivity beliefs predict quitting (Table 3). While Burks, Cowgill, Hoffman, and Housman (2015) primarily use the entire

[9] This is in contrast to the simplified timing structure in the model, where miles are not observed in t if a driver quits in t.

[10] Including cognitive skills in our tables on whether productivity beliefs predict productivity (Table 2) or whether productivity beliefs predict quitting (Table 3) has almost no effect on our estimates.

Firm A dataset for their analyses of truckers, they also use the data subset at times, e.g., for comparing referred and non-referred drivers in terms of cognitive skill.

B Field Experiments with Firm B, Further Information

B.1 Background

The first goal of the field experiment was to examine if using incentives for accurate guessing would have any effect on productivity beliefs, given that our main beliefs data from Firm A are non-incentivized. In addition, we sought to "test" our main counterfactual of debiasing (that is, of eliminating overconfidence) by providing information. On the first goal, we find that incentives do not seem to affect beliefs. On the second goal, we find that reducing overconfidence through information does seem potentially feasible (at least in the short-run), but we lack the statistical power to examine whether information-induced changes in beliefs affected quitting (though we do observe a statistically insignificant uptick in quitting).[11]

Firm B is a large trucking firm. Unlike Firm A, Firm B does not operate CDL training schools. All drivers in the study had already received a commercial driver's license before starting with Firm B, and Firm B did not use training contracts on the workers. We attempted to contact all workers who had started at the company in the last several months prior to the start of the experiment. The experiment was conducted via weekly phone surveys. 272 workers participated in the experiment. Phone calls were made by one of the authors (Hoffman) and by undergraduate research assistants. The experiments and data-gathering from Firm B were conducted for this paper and have never been used in any other research.

As seen in Figure B1, the first randomization was whether or not a worker would receive incentives for accurately guessing. We asked for guesses about both mileage and earnings, but we focus on the mileage predictions (see Section B.4 below). Of the 272 workers, 134 were in the $10 incentive condition and 138 were in the No Incentives condition. We divided workers into 5 groups based on how recently they joined Firm B. Randomization occurred within each of these 5 groups. There is a slight difference in worker counts between the Incentives and No Incentives conditions because for logistical reasons, workers were randomized into $10 Incentive and No Incentives conditions before they agreed to participate in the experiment.[12] Drivers assigned to the No Incentives condition were asked every week to predict their miles and earnings for the next week without incentives for accuracy. Drivers assigned to the Incentives condition were asked to predict their miles and earnings for next week, with accurate guesses rewarded under a quadratic scoring rule. Quadratic scoring rules are a common means by which experimental economists elicit agents' expectations in an incentive-compatible manner.[13]

Under a quadratic scoring rule, the agent's payoff is $A - B(x-b)^2$, where b is the agent's stated belief, x is the realization of the outcome of interest, and A and B are constants chosen by the researcher, with $B > 0$. This mechanism is incentive-compatible if the subject is risk-neutral.[14] The

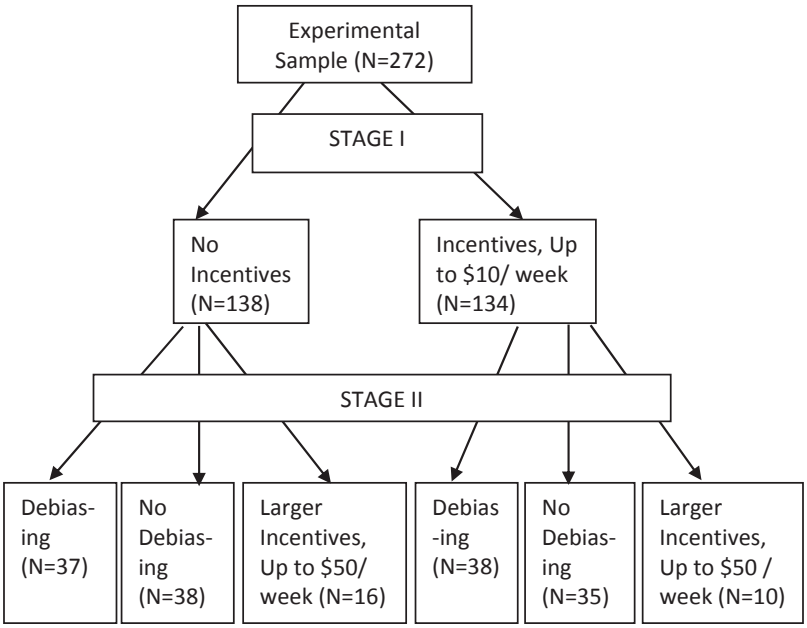
[11] Our result that having quadratic scoring rule incentives does not affect beliefs is consistent with lab experiments by Friedman and Massaro (1998) and Sonnemans and Offerman (2001).

[12] Specifically, we did the randomization before workers agreed to participate because the phone interviewers explained the incentive system to workers in the Incentives condition immediately after the worker accepted.

[13] Several recent papers using quadratic scoring rules include Holt and Smith (2009), Radzevick and Moore (2011), and Hoffman (2016). While there is an active debate about when they should be used and in what form (Hossain and Okui, 2013; Offerman et al., 2009; Schlag and van der Weeley, 2009), quadratic scoring rules are a standard and established tool in experimental economics (Selten, 1998).

[14] To see why, consider the problem of choosing b in order to maximize one's payoff. Let $f(x)$ denote the agent's subjective assessment of the distribution of x. The problem is: $argmax_b \int A - B(x-b)^2 f(x)dx = argmin_b \int (x -$

Figure B1: Firm B Experiment: Experimental Design



Notes: The number of workers participating in Stage II is smaller than in Stage I due to attrition from the survey (where we were unable to have phone contact with a driver for an interview) and attrition from the company.

payment for guessing miles was $10 - 10 * ((x_m - b_m)/1000)^2$ and the payment for guessing earnings was $10 - 40 * ((x_e - b_e)/1000)^2$, where x_m and x_e are actual miles and earnings and b_m and b_e are predicted miles and earnings. To preserve incentive-compatibility, drivers were paid for either their prediction of earnings or miles, with which one determined randomly. All subjects were paid a $5 participation fee for each survey taken.

Significant care was taken so that the incentives for accuracy would be understood by drivers. We told drivers that they would be rewarded for their accuracy and explained the payment amounts through examples. We explained to drivers that the reward rule was incentive-compatible, that is, "that you maximize your reward by stating your true beliefs." If drivers had further questions or wanted to know more, we walked them through additional examples and provided them with the exact mathematical formula for the reward. Our approach of telling subjects the quadratic scoring rule is incentive-compatible follows Radzevick and Moore (2011) and Hoffman (2016). The experimental instructions and survey wording are given below in Section B.6.

Workers made predictions for about 2-6 weeks. The number of weeks for workers in Stage I varied based on the week on which workers were first contacted and the number of weeks for which we were unable to contact them for an interview. While there is substantial subject attrition throughout the experiment, there are no significant differences between the Incentives and No Incentives conditions in attrition.[15]

After about 2-6 weeks, workers were assigned to a different treatment. For four-fifths of the workers (specifically, for 4 of the 5 groups of workers based on date of hire), they were randomly

$b)^2 f(x)dx$. This leads to a first-order condition of $\int \frac{d}{db}(x-b)^2 f(x)dx = 0$, which simplifies to $b = \int x f(x)dx$.

[15] Specifically, we analyzed whether a driver was in the study for at least X weeks, where X was a number from 2-6, and regressed it on whether the driver was assigned to the Incentive condition or not, and there was no significant correlation.

Table B1: Field Experiment at Firm B: Covariate Balance

Panel A: Stage I	No Incentives	$10 Incentive	t-test of (1) vs (2)			
	(1)	(2)	(3)			
Female	0.07	0.08	0.59			
Age	41.87	41.24	0.64			
Experience in years	8.4	7.52	0.43			
West	0.43	0.43	0.93			
South	0.35	0.4	0.35			
Midwest	0.21	0.15	0.19			
Northeast	0.01	0.01	0.98			
Number of drivers	138	134				
Panel B: Stage II	Debiasing	No Debiasing	$50 Incentive	t-test of (1) vs (2)	t-test of (1) vs (3)	t-test of (2) vs (3)
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.01	0.13	0.12	0.01	0.02	0.88
Age	41.7	41.3	43.3	0.84	0.54	0.46
Experience in years	8.55	8.24	5.89	0.84	0.20	0.26
West	0.41	0.48	0.42	0.42	0.93	0.62
South	0.33	0.36	0.46	0.77	0.25	0.35
Midwest	0.25	0.15	0.08	0.12	0.06	0.34
Northeast	0.00	0.01	0.04	0.31	0.09	0.45
Number of drivers	75	73	26			

Notes: Columns marked "t-test" display p-values calculated using a two-sided t-test. The number of workers participating in Stage II is smaller than in Stage I due to attrition from the survey (where we were unable to have phone contact with a driver for an interview) and attrition from the company. Regions are defined based on US Census regions. Experience is a driver's total years of trucking experience and we measure it once (it does not vary across weeks). Six drivers have gender missing and seven drivers have experience missing.

assigned to receive debiasing (75 workers) or no debiasing (73 workers). The fifth group of workers was assigned to receive a larger incentive (26 workers).[16] For this group, we randomized the order in which drivers would receive the larger incentive.[17]

In the larger incentives treatment, drivers were paid up to \$50 per guess and faced sharper penalties for mistakes. These drivers were paid according to the rules $50 - 200 * ((x_m - b_m)/1000)^2$ and $50 - 800 * ((x_e - b_e)/1000)^2$. The debiasing treatment consisted of telling workers at Firm B about the existence of overconfidence in the workers at Firm A, as well as reminding the Firm B workers of their average prediction to date. The no debiasing treatment consisted of simply reminding drivers of their average miles prediction to date (further details below in Section B.3).

B.2 Further Discussion on Stake Size

We put significant thought into designing appropriate financial stakes for the experiment. We designed the experiment's incentive system in consultation with Firm B managers. The bonus amount of up to \$10 was chosen so as to be large enough to be salient for drivers, but small enough to be unlikely to influence their driving behavior. A \$10 incentive is significant relative to drivers' value of time—the experiment each week was quite brief (usually about five minutes or less), whereas drivers often make around \$10-\$25 per hour.

We chose to also incorporate a larger incentive treatment (up to \$50 per week) to test the robustness of the main results. We wanted to know, even if there was no difference in overconfidence in workers with no incentive and an incentive of up to \$10, might there be one when stakes were made larger? Firm B believed that both the smaller (up to \$10) and larger (up to \$50) incentive systems would be salient for drivers and would make drivers put effort into their guesses. Had we paid workers hundreds or thousands of dollars for guessing accurately, the incentive system could have affected worker behavior, invalidating the incentive-compatibility of the scoring rule. For example, workers might have chosen to stop driving exactly when they reached their guess, or have engaged in excessive speeding in order to reach their guess. In addition, paying hundreds or thousands of dollars to some workers but not others could have caused workplace equity problems.

To our knowledge, there is limited prior work on the impact of stake size on the effectiveness of quadratic scoring rules.[18] In experimental economics as a whole, there is no general evidence that experimental results with smaller stakes are undone by using larger stakes.[19] In light of this, we would speculate that our conclusions would not be undermined had we used incentives beyond up to \$50.

[16] The group assigned to receive the big incentives had the longest tenure out of the 5 groups, but they still had not been employed at Firm B for very long. In addition, drivers in the bigger incentive group actually have lower average total trucking experience than drivers assigned to debiasing or no debiasing, but the difference is not statistically significant.

[17] Thus, whether drivers received the big incentive vs. something else in Stage II was not randomly assigned (though the three groups do look relatively balanced on covariates, as seen in Table B1). However, we can exploit the fact that the weeks (i.e., order) in which drivers received the larger incentive was randomly assigned. We regress mileage predictions on a dummy for having the larger incentive, driver fixed effects, dummies for the number of days that a driver predicts not working during the upcoming week (including a dummy for this being missing), and week of interview controls. We restrict the sample to workers who eventually get the larger incentive and trim the lower and upper 5% of predictions to limit the effect of outliers. Though standard errors are fairly large, we see no evidence that the larger incentive reduces prediction. Specifically, the coefficient is +124 miles (se=157 miles), leading to a sizable 95% confidence interval of -200 to +449 miles, but one where we can rule out very large negative impacts on beliefs.

[18] See Armantier and Treich (2013) for an exception.

[19] See Camerer and Hogarth (1999) for discussion. For example, Roth et al. (1991) and Cameron (1999) find that most aspects of play in the ultimatum game are similar even when stakes are made very large, as do Cherry et al. (2002) for the dictator game. See also the discussion in Levitt and List (2007), which contains a few examples where large incentives do seem to matter.

B.3 Debiasing

Psychologists have long been interested in whether overconfidence and other behavioral biases can be eliminated, focusing primarily on laboratory settings.[20] After discussion with a psychologist on different methods of debiasing, we chose to inform the workers at Firm B about our findings on overconfidence with the workers at Firm A. Workers were either administered the debiasing treatment, where they received information about our findings about overconfidence, a suggestion to reflect on past predictions, and information about their average prediction in their first several weeks of the experiment; or the control treatment, where they received information about their average prediction in the first several weeks of the experiment. Our debiasing treatment is deliberately somewhat heavy-handed, as we wanted to avoid a treatment that seemed too weak to affect anyone's beliefs. At the same time, however, we wanted our debiasing not to require a lot of individual information (as would, say, an alternative treatment of providing individual-level feedback on over-prediction), both for logistical ease and in recognition that an actual debiasing policy may not be able to provide extensive individual information.

Drivers selected for debiasing were read the following script. (After the first paragraph, drivers were asked if they had any questions or comments.)

Before we get started, we'd like to share with you some of our findings so far. At another trucking company we studied, workers over-estimated their next week's miles by around 500 miles per week during their first few months with the company. That is, they thought they were going to drive 500 more miles per week than their actual average miles. Even after more than one year with the company, people were still over-predicting their miles by 300 miles per week; for example, many people thought they would average 2,400 miles per week, but they ended up only driving 2,100 miles per week.

Please think for a moment about the last few weeks. Were your predictions of your mileage high or low? Also think about the week ahead. Are there any factors that might decrease you mileage, for example, bad weather, bad traffic, or a late unloading?

In our survey, your average prediction per week has been [INSERT MILES NUMBER] miles.

Drivers selected not to receive debiasing were simply told:

In our survey, your average prediction per week has been [INSERT MILES NUMBER] miles.

B.4 Miscellaneous Data Issues

Earnings. Before the experiment, our impression was that we would be able to obtain precise data from Firm B on driver earnings. While we obtained precise data on driver miles,[21] we have not been able to obtain precise data on driver earnings. An important difficulty is that drivers at Firm B get paid different rates per mile on different loads, and we do not have load-level data. Lacking precise data on earnings, we focus our analysis on drivers' predictions of their miles.[22] For purposes of driver payment, we calculated our best guess of driver earnings.

[20] Fischoff (1982) provides an excellent early summary of the literature. Many papers provide support for the feasibility of laboratory debiasing (e.g., Arkes et al., 1987; Lau and Coiera, 2009), but many also do not (e.g., Sanna et al., 2002; Fleisig, 2011). In economics, there is limited work that explicitly attempts to debias overconfidence (an exception being the lab experiment by Larkin and Leider (2012)), though there is growing work on debiasing in other contexts (e.g., debiasing misperceptions about financial investments or schooling decisions).

[21] While the mileage data are precise, the mileage data are still not perfect for our purposes. Specifically, we encountered some challenges in matching miles to the precise time window drivers are forecasting over, but we do not think this affects any of our results.

[22] We have also done analysis of the impact of incentives and debiasing on earnings predictions. We found no significant impact of incentives on earnings predictions. Like for mileage predictions, we found that debiasing significant reduced earnings predictions, and that the effects on earnings were a bit more persistent than those on miles.

Quitting. We measure a worker as having quit the company if the worker is missing miles or has zero miles in the final two weeks of our data. This is a proxy for having left the firm instead of an actual record of it. Further, unlike Firm A where we have data codes to distinguish quits and fires, we cannot do so at Firm B.

B.5 Results

Table B2 estimates the impact of incentives and information on people's beliefs:

$$b_{it} = \alpha_0 + \alpha_1 10INCENT_{it} + \alpha_2 50INCENT_{it} + \alpha_3 DEBIAS_{it} + \beta t + X_{it}\delta + \epsilon_{it}$$

where b_{it} is agent i's subjective belief at tenure t; $10INCENT_{it}$ and $50INCENT_{it}$ are dummies for having up to a \$10 or \$50 incentive for guessing about productivity in week t; $DEBIAS_{it}$ is a dummy for having received the debiasing treatment at or before week of tenure t; X_{it} is other control variables; and ϵ_{it} is an error. α_1 and α_2 are the impact of financial incentives for accuracy on worker beliefs. α_3 is the impact of information on worker beliefs.

Table B2 shows that incentives had little impact on beliefs, but that debiasing seems to reduce beliefs. Debiasing reduces miles beliefs by 113 miles (column 2) in our preferred specifications with controls. However, effects vary substantially by time since the week of debiasing. In the week of debiasing, miles beliefs decline by 207 miles. Given that the average miles overprediction in the Firm B data is 253 miles, the experiment eliminated nearly 80% of miles overconfidence in the first week. The coefficients decline as more weeks pass, remaining sizable, but become statistically insignificant.[23]

Column 1 of Table B3 shows that the debiasing experiment led to a 8 percentage point increase in actual quitting. This is statistically insignificant, but sizable relative to the mean of 29% for drivers without debiasing. While we designed the experiment to estimate impacts on beliefs in Table B2 with reasonable precision, the impacts on quits are much less precisely estimated. The 95% confidence interval for the impact on quitting is -10 to +23 percentage points. In our counterfactual simulation in Table 6, by the start of the 9th week of tenure, eliminating 50% of overconfidence increases quitting by 12 percentage points.[24] Thus, while statistically insignificant, our experimental estimate is relatively close to that from the structural model. Table B3 also shows that debiasing had no impact on surveyed intention to search for a new job or surveyed job satisfaction.

We show the main results are robust in two robustness checks where we focus on either debiasing or incentives for accuracy. In Table B5, we restrict the sample to weeks where drivers have already

[23] While it is possible that the experimental impacts on beliefs could be driven by an "experimental demand" effect, as is the case for many lab and field experimental findings, impacts on beliefs are similar whether or not beliefs are incentivized. The reduction is 106 (standard error=88) miles with an incentive and 123 (se=97) miles without an incentive, when we repeat column 2 of Table B2, splitting the sample by incentive for guessing or not. If the debiasing impacts occurred merely because subjects wanted to tell the surveyors "what they wanted to hear," one would think that this may not occur when subjects are incentivized to guess correctly.

[24] We look at the impact of debiasing on survival to the start of the 9th week of tenure in our counterfactual simulation since this time period corresponds roughly to debiasing in the randomized experiment. In the experiment, workers are tracked after debiasing for about 2 months to calculate their quitting percentage. In addition, we look at 50% debiasing since our randomized experiment did not permanently eliminate all of worker overconfidence and is best thought of as reducing some of worker overconfidence. One reason why our experiment may have failed to eliminate all overconfidence permanently is that it was a one-time intervention. Given that the experiment's impacts on beliefs appear to fade somewhat after a few weeks, it is not particularly surprising that the experiment did not affect real outcomes like quitting. To more permanently eliminate worker overconfidence, it may instead be necessary to provide debiasing information on a more frequent basis. Finally, while the experimental impacts seeming to diminish over time could potentially reflect experimental demand effects, they seem more likely to us to be a manifestation of limited memory.

Table B2: Do Incentives for Accuracy or Information Reduce Worker Overconfidence? The Field Experiment with Firm B

Dep var:	Miles Prediction (in miles)		
	(1)	(2)	(3)
Incentives for accuracy (up to $10/wk)	-32.5	-56.1	-56.4
	(50.5)	(46.8)	(46.8)
Larger incentives (up to $50/wk)	-4.7	-45.0	-42.7
	(89.9)	(84.9)	(85.2)
Debiasing	-95.7	-112.9	
	(70.7)	(65.4)	
Debiasing X 0wk post-treat			-207.7
			(80.2)
Debiasing X 1wk post-treat			-91.5
			(89.4)
Debiasing X 2wk post-treat			-111.1
			(82.1)
Debiasing X 3wk post-treat			-81.7
			(95.2)
Debiasing X 4-6wks post-treat			-61.5
			(73.1)
Joint sig of two incentive treatments incentive treatments (p-value)	0.809	0.478	0.477
Demographic Controls	No	Yes	Yes
Observations	1,097	1,072	1,072
Mean dep var	2316	2314	2314
Subjects (clusters)	254	243	243

Notes: OLS regressions with standard errors clustered by driver in parentheses. An observation is a worker-week. The mean over-prediction in miles is 253 miles in the column 1 sample. All regressions include worker tenure in weeks and dummies for the number of days that a driver predicts not working during the upcoming week (including a dummy for this being missing). The variable "Debiasing" equals one if the driver had received the Debiasing information treatment in the current week or a past week. All regressions also include a dummy for assignment to Debiasing (irrespective of whether the worker is debiased in a future week) and a variable indicating whether the worker had received either the Debiasing or No Debiasing information treatment in a current or past week. Demographic controls are controls for gender, age, trucking experience (measured once), and region of home residence. To limit the effect of outliers, we trim the lower and upper 5% on the dependent variable. This trimming leads the number of subjects to be less than 272.

Table B3: Impacts of Debiasing on Quitting, Intention to Search for a New Job, and Job Satisfaction

Dep Var:	Actual Quitting (0-1)	Intention to Search for a New Job (1-3)	Job Satisfaction (1-4)
Method:	OLS	Ordered Probit	Ordered Probit
	(1)	(2)	(3)
Debiasing	0.07	-0.05	0.01
	(0.08)	(0.27)	(0.22)
Incentives for accuracy (up to $10/wk)	-0.01	-0.19	0.01
	(0.08)	(0.27)	(0.22)
Observations	117	99	319

Notes: Standard errors clustered by driver in parentheses. "Actual Quitting" is whether the worker quits during the time frame of the study. The question about search intention was asked only once, coming 1-3 weeks after debiasing. The question about job satisfaction was asked in multiple weeks. The sample is restricted to people assigned to receive debiasing or not. Intention to Search for a New Job is the worker's intention to look for a new job during the next 6 months and is measured on a 1-3 Scale (Not at all likely, Somewhat likely, Very likely). Job Satisfaction is overall current job satisfaction and is measured on a 1-4 Scale (Not at all satisfied, Not too satisfied, Somewhat satisfied, Very satisfied). In column 1, we control for driver experience and driver tenure in months when the Firm B RCT first began. In columns 2 and 3, we control for driver experience, dummies for current week in the study, and dummies for the number of days that a driver predicts not working during the upcoming week (as in Table B2).

Table B4: Field Experiment, Incentive Robustness: Sample Restricted to Stage I of Experiment

	Miles Prediction		Miles Overconfidence	
	(1)	(2)	(3)	(4)
Incentives for accuracy (up to $10/wk)	-20.6	-34.3	10.9	28.7
	(52.7)	(50.0)	(69.9)	(68.8)
Demographic Controls	No	Yes	No	Yes
Observations	573	557	472	467
Mean dep var	2320	2320	300.3	297.4
Subjects (clusters)	252	241	227	222

Notes: OLS regressions with standard errors clustered by driver in parentheses. An observation is a worker-week. All regressions include worker tenure in weeks and dummies for the number of days that a driver predicts not working during the upcoming week (as in Table B2). The demographic controls are the same as in Table B2. To limit the effect of outliers, we trim the lower and upper 5% on each dependent variable.

Table B5: Field Experiment, Debiasing Robustness: Sample Restricted to Stage II of Experiment

Panel A: Impact on Mileage Prediction

	0-6 weeks after debiasing (1)	Week of debiasing (2)	Week after debiasing (3)	2 weeks after debiasing (4)	3 weeks after debiasing (5)	4-6 weeks after debiasing (6)
Debiasing	-127.0	-135.1	-177.8	-56.5	-92.4	-170.6
	(59.9)	(91.7)	(104.3)	(104.5)	(122.7)	(82.5)
Observations	474	110	84	77	67	136

Panel B: Impact on Prediction - Avg Pre-Debias Productivity

	0-6 weeks after debiasing (1)	Week of debiasing (2)	Week after debiasing (3)	2 weeks after debiasing (4)	3 weeks after debiasing (5)	4-6 weeks after debiasing (6)
Debiasing	-196.7	-151.2	-277.1	-68.9	-105.6	-254.7
	(52.1)	(79.7)	(89.3)	(106.8)	(101.7)	(82.5)
Observations	447	105	79	68	65	130

Notes: OLS regressions with standard errors clustered by driver in parentheses. An observation is a worker-week. All regressions include a dummy for having the $10 incentive in a given week; the average number of miles in pre-debiasing prediction that was shared to the driver as part of the Debiasing or No Debiasing treatments; worker tenure in weeks; dummies for the number of days that a driver predicts not working during the upcoming week (as in Table B2); and demographic controls (same as in Table B2). To limit the effect of outliers, we trim the lower and upper 5% on each dependent variable.

received the Debiasing or No Debiasing treatment in the current or a past week. In Table B4, we restrict to workers in Stage I of the experiment, where they are receiving either incentives or no incentives for accurate guessing. In this sample, as well, we again see no evidence of the incentives on mileage predictions or overconfidence (mileage prediction minus actual productivity that week).

B.6 Experiment Wording

B.6.1 Incentivized Version, $10

[For subsequent surveys] Hi this is [FULL NAME] from the University of California, with the trucking survey. Might you like to participate again?[25]

[First survey] In the next two questions, we're going to ask you to estimate your miles and earnings for next week, if you're willing. We're going to give you a small reward (in addition to the $5) for predicting accurately. For example, if you run exactly the number of miles you predict, you get $10. For each mile you're off, the reward will go down, with larger reductions the further you are off. If you're off by 500 miles, you get $7.50. And if you're off by 1,000 or more miles, you get $0. We'll use a similar reward system for your prediction on how much you will earn. This might sound complicated, but this system has been used in other research, and is specially designed so that you maximize your reward by stating your true beliefs. We'll pay you either for your miles or your

[25] Note that we made small changes to survey wording over the course of the experiment. As an example, earlier on, we had asked drivers to predict their miles starting on Tuesday, but we later shifted to asking about Monday through Sunday after discussion with a Firm B manager. Our main result on incentives not affecting beliefs is robust to restricting to the time period before or after the question shift.

earnings guess, with which one chosen randomly. Does this make sense? [If not, explain to them. Also, go through payment system on back if want to know more.]

[For subsequent surveys] Do you happen to remember how the reward system works, where you get rewarded for guessing close to the actual number of miles you run? [If not, refresh their memory.]

- How many miles do you expect to run next week, that is, from Tuesday until next Tuesday?
- How many dollars do you expect to earn before taxes next week, that is, from Tuesday until next Tuesday? [If someone is a team driver, ask them if the miles and earnings they reported are for themselves or for the two of them. If they give for the two of them, ask them to report miles and earnings for themselves.]
- Next week, then, are there any days when you will not be working?

B.6.1.1 Further Information on Payment System to Give Respondents

[This information was given to respondents when they had further questions about the quadratic scoring rule.] Here are some further examples of how you will be paid for the miles prediction:

Distance Between Actual and Predicted Miles	Your Payment
Your guess equals the actual	$10
Your guess is 250 miles from the actual	$9.38
Your guess is 500 miles from the actual	$7.50
Your guess is 750 miles from the actual	$4.38
Your guess is 1,000 or more miles from the actual	$0.00

Specifically, your payment will be given by the equation Payment = $10 - $10*(Actual Miles in Thousands *minus* Predicted Miles in Thousands)^2.

Here are some further examples of how you will be paid for the earnings prediction:

Distance Between Actual and Predicted Miles	Your Payment
Your guess equals the actual	$10
Your guess is 100 dollars from the actual	$9.60
Your guess is 200 dollars from the actual	$8.40
Your guess is 300 dollars from the actual	$6.40
Your guess is 400 dollars from the actual	$3.60
Your guess is 500 or more dollars from the actual	$0.00

Specifically, your payment will be given by the equation Payment = $10 - $40*(Actual Earnings in Thousands *minus* Predicted Earnings in Thousands)^2.

B.6.2 Unincentivized Version

[For subsequent surveys] Hi this is [FULL NAME] from the University of California, with the trucking survey. Might you like to participate again?

- How many miles do you expect to run next week, that is, from Tuesday until next Tuesday?
- How many dollars do you expect to earn before taxes next week, that is, from Tuesday until next Tuesday? [If someone is a team driver, ask them if the miles and earnings they reported are for themselves or for the two of them. If they give for the two of them, ask them to report miles and earnings for themselves.]
- Next week, then, are there any days when you will not be working?

B.6.3 Increasing the Incentive to $50 per Week

This week, we're going to do something a little different. You will earn up to $50 for predicting accurately instead of $10. For example, if you run exactly the number of miles you predict, you get $50. For each mile you're off, the reward will go down, with larger reductions the further you are off. If you're off by 250 miles, you get $37.50. And if you're off by 500 or more miles, you get $0. We'll use a similar reward system for your prediction on how much you will earn. This might sound complicated, but this system has been used in other research, and is specially designed so that you maximize your reward by stating your true beliefs. We'll pay you either for your miles or your earnings guess, with which one chosen randomly. Does this make sense? [If not, explain to them. Also, go through payment system on back if want to know more.]

B.6.3.1 Further Information on Payment System to Give Respondents, Incentive up to $50 per week [This information was given to respondents when they had further questions about the quadratic scoring rule.] Here are some further examples of how you will be paid for the miles prediction:

Distance Between Actual and Predicted Miles	Your Payment
Your guess equals the actual	$50.00
Your guess is 125 miles from the actual	$46.88
Your guess is 250 miles from the actual	$37.50
Your guess is 375 miles from the actual	$21.88
Your guess is 500 or more miles from the actual	$0.00

Specifically, your payment will be given by the equation Payment = $50 - $200*(Actual Miles in Thousands *minus* Predicted Miles in Thousands)^2, so long as it is greater than 0.

Here are some further examples of how you will be paid for the earnings prediction:

Distance Between Actual and Predicted Earnings	Your Payment
Your guess equals the actual	$50
Your guess is 50 dollars from the actual	$48
Your guess is 100 dollars from the actual	$42
Your guess is 150 dollars from the actual	$32
Your guess is 200 dollars from the actual	$18
Your guess is 250 or more dollars from the actual	$0

Specifically, your payment will be given by the equation Payment = $50 - $800*(Actual Earnings in Thousands *minus* Predicted Earnings in Thousands)^2, so long as it is greater than 0.

B.7 Additional Questions asked in the Weeks After Debiasing or No Debiasing

- **Job Search.** Taking everything into consideration, how likely is it you will make a genuine effort to find a new job within the next 6 months? Not at all likely, Somewhat likely, or Very likely?
- **Job Satisfaction.** All in all, how satisfied are you with your job? Not at all satisfied, Not too satisfied, Somewhat satisfied, or Very satisfied?

C One Period Model

In this section, we present a very simple one-period model to show formally that differential overconfidence (i.e., being more overconfident about the inside option compared to the outside option) will make a worker less likely to quit after training.

Consider a firm that trains its workers. The worker's post-training productivity and earnings are uncertain. Let W be the worker's true post-training earnings. Let $\overline{W}$ be the worker's post-training outside option. This outside option is utility inclusive of any quit penalties paid. Workers have some non-pecuniary taste for the job ε, which they learn after training. We assume that ε has a distribution function F and has support over the entire real line. A worker decides to quit by comparing $W + \varepsilon$ compared to $\overline{W}$. If a worker is overconfident, we let $B(W)$ denote be his belief about his earnings in the inside option, and $B(\overline{W})$ be his belief about his earnings in the outside option. The following proposition is easy to see:

Proposition 1 *Consider two workers with the same ability, training contract, and piece rate, one worker who is overconfident and one who is not. Then the overconfident worker will be less likely to quit than the rational worker if and only if he is more overconfident about his inside than his outside option.*

Proof. The probability of staying for the rational worker is $1 - F(\overline{W} - W - k)$, whereas it is $1 - F(B(\overline{W}) - B(W) - k)$ for the overconfident worker. The probability of staying is higher for the overconfident worker when $B(W) - B(\overline{W}) > W - \overline{W}$ or when $B(W) - W > B(\overline{W}) - \overline{W}$. ■

We test this proposition in Table 3. Specifically, we regress quitting on worker subjective beliefs and average productivity to date. We want to use Beliefs instead of (Beliefs - Productivity) as the main regressor since the probability of staying, $1 - F(B(\overline{W}) - B(W) - k)$, depends only on beliefs, and not the difference between beliefs and productivity. Empirically, we find that workers with higher beliefs are less likely to quit.[26]

[26]Instead of comparing an overconfident worker with a rational worker in Proposition 1, we could alternatively examine the impact on retention of slightly raising a worker's overconfidence (that is, his belief about his inside option). This will increase retention if and only if $\frac{\partial B(W)}{\partial B(\overline{W})} < 1$. So, when the problem is re-phrased this way, the required assumption is not differential overconfidence, but rather that beliefs about the outside option rise less than one-for-one with beliefs about the inside option. The assumption that outside beliefs rise less than one for one with inside beliefs is closely related to the assumption of differential overconfidence, and also seems quite plausible in our setting, for many of the same reasons that we give in Section 4.2.

D Structural Model and Estimation Details

Estimation Sample. For the sample for the structural analysis, we start with our baseline data subset sample of 895 drivers. Next, we drop any drivers who are ever seen working at non-piece rate trucking jobs at Firm A where they are paid based on their activities or on salary (e.g., this drops drivers who ever go to work themselves as driver trainers at the training schools). We also drop a small number of drivers with a missing individual characteristic, leaving an estimation sample of 699 drivers.

Probability of Staying. Let $\Lambda(x) = \frac{exp(x)}{1+exp(x)}$ and let fixed non-pecuniary taste for the job be $\alpha + X\overline{\alpha}$, where α has a mass point distribution and $\overline{\alpha}$ are the utility coefficients associated with different worker characteristics. At time T, the probability of staying, given the state variables, is:[27]

$$\begin{aligned} Pr(STAY_T|\mathbf{x}_T) &= Pr(V_T^S > V_T^Q|y_1,...,y_{T-1},X,\alpha,\eta_b) \\ &= Pr(\alpha + X\overline{\alpha} + E^b(w_T y_T|y_1,...,y_{T-1}) + \delta E^b(V(\mathbf{x})|\mathbf{x}_T) + \varepsilon_T^S > -k_T + \frac{r_T}{1-\delta} + \varepsilon_T^Q) \\ &= \Lambda\left(\frac{\alpha + X\overline{\alpha} + w_T E^b(y_T|y_1,...,y_{T-1}) + \delta E^b(V(\mathbf{x})|\mathbf{x}_T) + k_T - \frac{r_T}{1-\delta}}{\tau}\right) \end{aligned}$$

To evaluate this probability, we need to calculate both $E^b(y_T|y_1,...,y_{T-1})$ and $E^b(V(\mathbf{x})|\mathbf{x}_T)$. The former depends on $y_1,...,y_{T-1}$, which would imply that the state space has dimensionality of order K^{T-1} when y_t is discretized with K values. The key to avoiding a very high dimensional problem is that in a normal learning model (both a model with standard beliefs and our generalized learning model), the worker's expectation of future productivity depends only on his prior and his de-trended average of past productivity. That is, *the average of past productivity is a sufficient statistic for the sequence* $y_1,...,y_{t-1}$ (DeGroot, 1970).

For a general period t, the probability of staying is:

$$Pr(STAY_t|\mathbf{x}_t) = Pr(V_t^S > V_t^Q|\mathbf{x}_t) = \Lambda\left(\frac{\alpha + X\overline{\alpha} + w_t E^b(y_t|y_1,...,y_{t-1}) + \delta E^b(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t) + k_t - \frac{r_t}{1-\delta}}{\tau}\right)$$

Calculating $E^b(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t)$ requires integrating expectations of future miles and ε shocks:

$$\begin{aligned} E^b(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t) &= E^b_{y_t} E^b_{\varepsilon|y_t}(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t) && (9) \\ &= E^b_{y_t} E^b_{\varepsilon}(max\{\overline{V}^S_{t+1}(\mathbf{x}_{t+1}) + \varepsilon^S_{t+1}, \overline{V}^Q_{t+1} + \varepsilon^Q_{t+1}\}|\mathbf{x}_t) && (10) \\ &= \int \tau \log\left(\exp(\frac{\overline{V}^S_{t+1}(\mathbf{x}_{t+1})}{\tau}) + \exp(\frac{\overline{V}^Q_{t+1}}{\tau})\right) f^b(y_t|y_1,...,y_{t-1})dy_t && (11) \\ &= \sum_k \tau \log\left(\exp(\frac{\overline{V}^S_{t+1}(\mathbf{x}_{t+1})}{\tau}) + \exp(\frac{\overline{V}^Q_{t+1}}{\tau})\right) P^b(y_t^k|y_1,...,y_{t-1}) && (12) \end{aligned}$$

Equation (9) expresses that the value function involves expectations over unknown miles and idiosyncratic shocks. Equation (10) uses the definition of V and that the idiosyncratic shocks are independent of miles. We write $\overline{V}^Q_{t+1}$ instead of $\overline{V}^Q_{t+1}(\mathbf{x}_{t+1})$ because k_{t+1} and r_{t+1} in $\overline{V}^Q_{t+1}(\mathbf{x}_{t+1})$ only depend on tenure in our data. Equation (11) integrates out y_t, which is not yet observed when the driver makes his period t quit decision (and where $f^b(y_t|y_1,\ldots,y_{t-1})$ is a perceived probability

[27] After time T, quitting is governed by the asymptotic value functions in (5) using $E^b(\cdot)$ instead of $E(\cdot)$.

density), and also integrates out the idiosyncratic shocks. Equation (12) follows because, in implementation, miles will be discretized into K possible values. The perceived transition probability $P^b(y_t^k|y_1,\ldots,y_{t-1})$ is expressed below in Equation (19). A related derivation can be found in Stange (2012).

Likelihood Function. Let $L_i = L(d_{i1},...,d_{it},y_{i1},...,y_{it},b_{i1},...,b_{it})$ be the likelihood of driver i for an observed sequence of quitting decisions, miles realizations, and subjective beliefs. We show how to derive the likelihood function.

$$
\begin{aligned}
L_i &= \int L(d_{i1},...,d_{it},y_{i1},...,y_{it},b_{i1},...,b_{it}|\alpha,\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b && (13)\\
&= \int \{L(d_{i1},...,d_{it}|y_{i1},...,y_{it},b_{i1},...,b_{it},\alpha,\eta_b) * L(b_{i1},...,b_{it}|y_{i1},...,y_{it},\alpha,\eta_b) && (14)\\
&\quad *L(y_{i1},...,y_{it}|\alpha,\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b\}\\
&= \left[\int L(d_{i1},...,d_{it}|y_{i1},...,y_{it},\alpha,\eta_b) * L(b_{i1},...,b_{it}|y_{i1},...,y_{it},\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b\right] L(y_{i1},...,y_{it}) && (15)\\
&= \left[\int \prod_{s=1}^{t} L(d_{is}|d_{i1},...,d_{is-1},y_{i1},...,y_{it},\alpha,\eta_b) * \prod_{s=1}^{t} L(b_{is}|b_{i1},...,b_{is-1},y_{i1},...,y_{it},\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b\right] && (16)\\
&\quad *L(y_{i1},...,y_{it})\\
&= \left[\int \prod_{s=1}^{t} L(d_{is}|y_{i1},...,y_{is-1},\alpha,\eta_b) \prod_{s=1}^{t} L(b_{is}|y_{i1}...,y_{is-1},\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b\right]\left(\prod_{s=1}^{t} L(y_{is}|y_{i1}...,y_{is-1})\right) && (17)\\
&\equiv \left[\int L_i^1(\alpha,\eta_b)L_i^3(\eta_b)f(\alpha,\eta_b)\,d\alpha d\eta_b\right] L_i^2 && (18)
\end{aligned}
$$

Equations (13), (14), and (16) follow from rules of probability. Equation (15) holds because (a) quit decisions are independent of reported subjective beliefs conditional on the overconfidence heterogeneity and miles realizations; (b) beliefs are unaffected by the taste heterogeneity; and (c) miles are unaffected by the taste and overconfidence heterogeneity. Equation (17) follows because (a) future miles are not observed when a worker decides to quit or forms his current subjective beliefs; (b) the ε shocks are i.i.d.; and (c) reported subjective beliefs are independent of past reported subjective beliefs conditional on miles realizations and the belief heterogeneity. In Equation (18), we define the part of the likelihood due to the quitting decisions as $L_i^1(\alpha,\eta_b)$, the part due to the miles realizations as L_i^2, and the part due to subjective beliefs as $L_i^3(\eta_b)$.

For a driver who quits in period t, $L_i^1(\alpha,\eta_b)$, L_i^2, and $L_i^3(\eta_b)$ can be written as

$$
\begin{aligned}
L_i^1(\alpha,\eta_b) &= \left(\prod_{s=1}^{t-1} Pr\left(STAY_{is}|\mathbf{x}_{is}\right)\right)\left(1-\Pr(STAY_{it}|\mathbf{x}_{it})\right)\\
L_i^2 &= f(y_{i1}) * \prod_{s=2}^{t} f(y_{is}|y_{i1},...,y_{is-1})\\
L_i^3(\eta_b) &= f(b_{i1}|\eta_b) * \prod_{s=2}^{t} f(b_{is}|y_{i1},...,y_{is-1},\eta_b)
\end{aligned}
$$

with

$$\begin{aligned}
f(y_{i1}) &\sim N(\eta_0 + a(1), \sigma_0^2 + \sigma_y^2) \\
f(y_{is}|y_{i1}, ..., y_{is-1}) &\sim N\left((1-\gamma_{s-1})\,\eta_0 + \gamma_{s-1}\frac{\sum_{n=1}^{s-1}(y_n - a(n))}{s-1} + a(s), \Omega_{s-1}\right) \text{ for } s > 1 \\
f(b_{i1}|\eta_b) &\sim N(\eta_0 + \eta_b + a(2), \sigma_b^2) \\
f(b_{is}|y_{i1}, ..., y_{is-1}, \eta_b) &\sim N\left((1-\gamma_{s-1}^b)(\eta_0 + \eta_b) + \gamma_{s-1}^b\frac{\sum_{n=1}^{s-1}(y_n - a(n))}{s-1} + a(s+1), \sigma_b^2\right) \text{ for } s > 1
\end{aligned}$$

where $\gamma_s = \frac{s\sigma_0^2}{s\sigma_0^2+\sigma_y^2}$, $\Omega_s = \frac{\sigma_0^2\sigma_y^2}{s\sigma_0^2+\sigma_y^2} + \sigma_y^2$, and $\gamma_s^b = \frac{s\sigma_0^2}{s\sigma_0^2+\widetilde{\sigma_y}^2}$.[28]

The overall likelihood is computed, first, by integrating over the unobserved heterogeneity for each individual's likelihood, and then by taking the product over all people. Since the unobserved heterogeneity is mass-point distributed, the integral becomes a sum.

$$\begin{aligned}
L &= \prod_i \left(\int L_i^1(\alpha, \eta_b) L_i^3(\eta_b) f(\alpha, \eta_b) d\alpha d\eta_b\right) L_i^2. \\
\log(L) &= \sum_i \log\left(\sum_{\alpha,\eta_b} L_i^1(\alpha, \eta_b) L_i^3(\eta_b) f(\alpha, \eta_b)\right) + \sum_i \log\left(L_i^2\right)
\end{aligned}$$

Perceived Transitions between Miles. In solving the dynamic programming problem, expected future mileage is governed by a perceived transition matrix. As mentioned above, we discretize productivity into K values. In our baseline estimation, we let productivity range in increments of 300 from 100 to 4,000 miles per week (that is, $K = 14$). Perceived transitions between miles states are given by:

$$P^b(y_s^k|y_1, ..., y_{s-1}) = \Phi\left(\frac{y_s^k + .5 * kstep - E^b(y_s^k|y_1, ..., y_{s-1})}{\sqrt{\Omega_{s-1}^b}}\right) - \Phi\left(\frac{y_s^k - .5 * kstep - E^b(y_s^k|y_1, ..., y_{s-1})}{\sqrt{\Omega_{s-1}^b}}\right) \quad (19)$$

where $\Omega_s^b = \frac{\sigma_0^2\widetilde{\sigma_y}^2}{s\sigma_0^2+\widetilde{\sigma_y}^2} + \widetilde{\sigma_y}^2$, y_s^k is the value of y_s at the kth grid point, and where $kstep$ is the distance between grid points. See Stange (2012) for a similar formula. Our estimates are similar using a finer grid with increments of 100 from 100 to 4,000 miles ($K = 40$) as seen in column 7 of Table F1. We can also derive perceived transition probabilities between levels of average productivity to date.

$$\begin{aligned}
P^b(\overline{y}_s^k|\overline{y}_{s-1}) = \Phi&\left(\frac{s\left(\overline{y}_s^k + .5 * kstep\right) - (s-1)\,\overline{y}_{s-1} - E^b\left(y_s|y_1, ..., y_{s-1}\right)}{\sqrt{\Omega_{s-1}^b}}\right) \\
-\Phi&\left(\frac{s\left(\overline{y}_s^k - .5 * kstep\right) - (s-1)\,\overline{y}_{s-1} - E^b\left(y_s|y_1, ..., y_{s-1}\right)}{\sqrt{\Omega_{s-1}^b}}\right)
\end{aligned}$$

For the parts of the likelihood on miles (L_i^2) or on subjective beliefs (L_i^3), we use the mileage data in continuous form instead of discretized.[29]

[28] This follows by applying the standard formula for the conditional density for a multivariate normal distribution: $X_1|(X_2 = x_2) \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

[29] That is, we assume that perceived transition probabilities are based on miles in a discrete form, whereas actual miles and beliefs are not. This can be justified on the grounds that perceived transition probabilities may be conceptually difficult for drivers, and may be naturally thought of according to a discrete grid.

Estimation Procedure. The model is estimated by maximum likelihood using an extension of the canonical nested fixed point algorithm (Rust, 1987). For every parameter guess, we first use value function iteration to solve for the asymptotic value functions (V_S and V_Q). With these in hand, we use backwards recursion to solve for the choice-specific value functions V_t^S and V_t^Q for $t = 1, ..., T$.

χ^2 test. χ^2 is calculated as $\sum_t$ (the number of drivers at risk during week of tenure t) * [(actual quit hazard(t) - predicted quiz hazard(t))^2 / predicted quiz hazard(t)].

12-Month Contract in Structural Model. The quit penalties under the training contracts varied slightly by training school at the firm. Furthermore, if drivers could not pay the money owed upon a quit, a significant interest rate may also have been assessed. For the structural estimation, we assume a penalty of $3,750 for the 12-month contract.

Zero Mile Weeks. The data contain a significant number of zero mile weeks for drivers. These often reflect weeks where the driver is not working. These weeks are not counted toward the miles component of the likelihood, and average miles to date (in terms of the quit decision) is given by the prior week's average miles to date. Also excluded from the likelihood are a small number of driver-weeks with predictions of 0 miles (estimates are similar when they are instead included).

Compensation and Additional Bonuses. At Firm A, drivers may receive small quarterly bonuses (based on customer/shipper satisfaction, good fuel economy, and other factors).[30] In addition, for low-mileage loads, drivers may receive "premiums" in cents per mile above their regular cents per mile. For computational simplicity, we ignore bonuses and premiums in our analysis. Further, at some points in the past, the firm has provided a guaranteed minimum earnings level for new inexperienced drivers when starting out (e.g., up through week 12), and we ignore this as well. For the piece rate-tenure profile in the structural model, we use data from an internal firm document in 2004. It provides the profile for the region where the training school in the data subset is located. We use the profile for the most common work type. Although actual pay per mile continues to increase with tenure, for simplicity in our model, we assume that pay per mile does not increase beyond the rate paid when drivers have 2–3 years of tenure. This is the rate paid when $T = 130$ weeks is reached.

Maximization. For the inner loop in the Rust (1987) procedure, we use a tight tolerance of 1e-15. For our baseline estimates, we maximize the likelihood function using "fminunc" in Matlab. We use a Quasi-Newton algorithm and a function tolerance of 2e-5. We verified that another Matlab algorithm, "fminsearch" (Nelder-Mead), yields convergence to the same parameter estimates. For the estimates with learning by doing, we first maximize using "fminunc" and then use the estimates as starting values for performing "fminsearch" (doing "fminsearch" with a function tolerance of 1e-5). We perform the maximization while restricting that the levels of the taste unobserved heterogeneity mass points (α_1, α_2, α_3) are greater than or equal to -$375 for the baseline case and greater than or equal to -$2,000 for the case with learning by doing.[31]

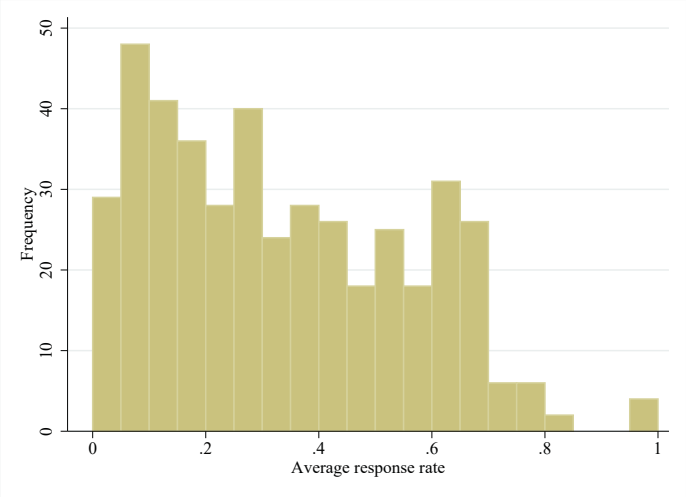
[30] At some points in the past, new inexperienced drivers only became eligible to receive a quarterly bonus after 1 year of tenure.

[31] If this restriction is not made, for the models with belief bias, maximization will sometimes yield parameter estimates where one of the taste mass points tends toward $-\infty$. In the baseline model (column 2 of Table 4), this can lead to a point with very low α for one unobserved type and high τ. In the model with learning by doing (column 2 of Table 5), it can lead to a point with very low α for one unobserved type and very high θ_1. We view such parameter estimates as less economically plausible, leading us to impose a restriction on the taste mass points. However, even for such parameter estimates, the key belief parameters, as well as the impact of debiasing on worker retention, firm profits, and worker welfare, are qualitatively similar to those in the main results.

Following Knittel and Metaxoglou (2014), we perform a number of checks on our optimization procedure for all our four main models in Tables 4 and 5. First, for the identified optima, we checked that our exit code indicates successful convergence; that $\|g\|_\infty$ and $g'H^{-1}g$ are small, where g is the gradient and H is the Hessian; and that H is positive-definite. Second, in line with Knittel and Metaxoglou (2014) and DellaVigna et al. (2017), for each model, we randomly generate a variety of starting values. We use uniform distributions for each parameter, drawing values over roughly economically plausible ranges. We verify that our reported parameter values yield the best fit out of the various estimates achieved.

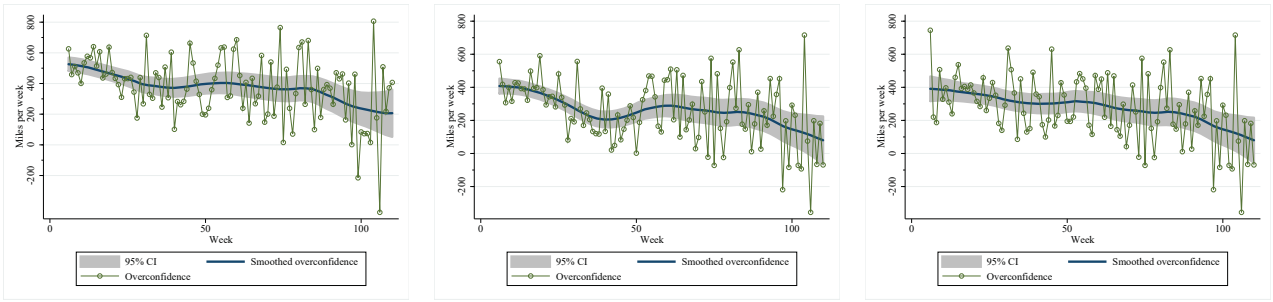
E Further Reduced-Form Results

Figure E1: Heterogeneity in Response Rates to the Firm A Subjective Productivity Beliefs Survey



Notes: This figure plots the distribution of driver-level average response rate to the survey (averaged over a driver's weeks in the data), excluding drivers who never respond. On the y-axis is the number of drivers in each bin. Observations are excluded from the sample if weekly miles are 0, if weekly predicted miles are 0, or if the driver is ever observed in the dataset receiving activity-based pay or salary pay instead of being paid by the mile.

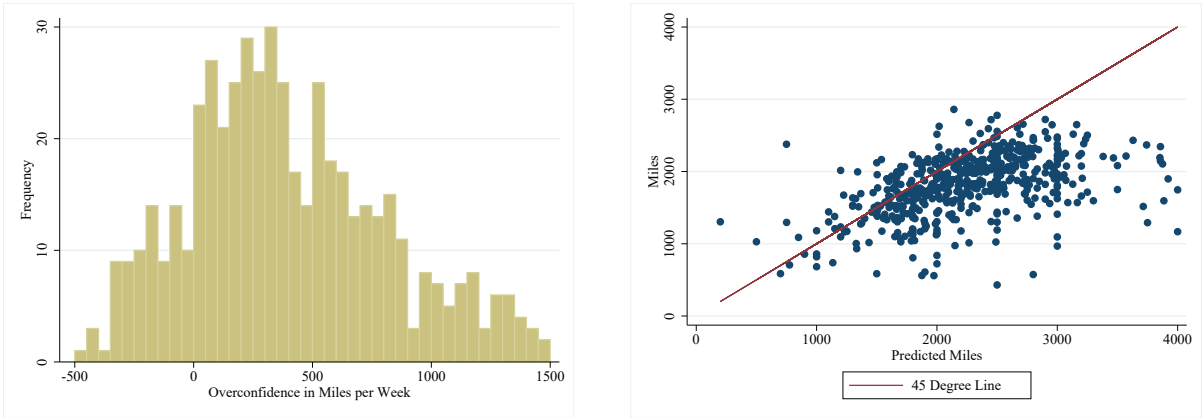
Figure E2: Tenure and Overconfidence (Productivity Beliefs *minus* Productivity)



(a) Means (b) Medians (c) Medians, ≥75 Wks, Responder

Notes: This figure analyzes the evolution of average driver overconfidence as a function of driver tenure. Overprediction, defined as productivity beliefs *minus* realized productivity, is collapsed (across all drivers) by week of tenure. Week t on the graph is corresponded to the driver's prediction in week t about productivity in week $t+1$, as well as to the driver's actual productivity in week $t+1$. The dots correspond to the collapsed means or medians. The smoothed curve is plotted using a local polynomial regression and a bandwidth of 7 weeks. In panel (a) beliefs minus actual productivity across drivers is collapsed into weekly means before local polynomial smoothing. In panels (b) and (c), beliefs minus actual productivity across drivers is collapsed into weekly medians before smoothing. In panel (c), we restrict to workers who stay at least 75 weeks and who respond to the beliefs survey that week. Results are similar if instead we look at workers who ever respond. Observations are excluded from the sample if weekly miles (one week ahead) are 0, if weekly predicted miles are 0, or if the driver is ever observed in the dataset receiving activity-based pay or salary pay instead of being paid by the mile. By looking at overprediction (instead of productivity and beliefs separately), we restrict to realized mile observations where there is a corresponding prediction. We restrict attention to weeks of tenure between 6 and 110 (early weeks involve training and the sample becomes relatively scant after around two years).

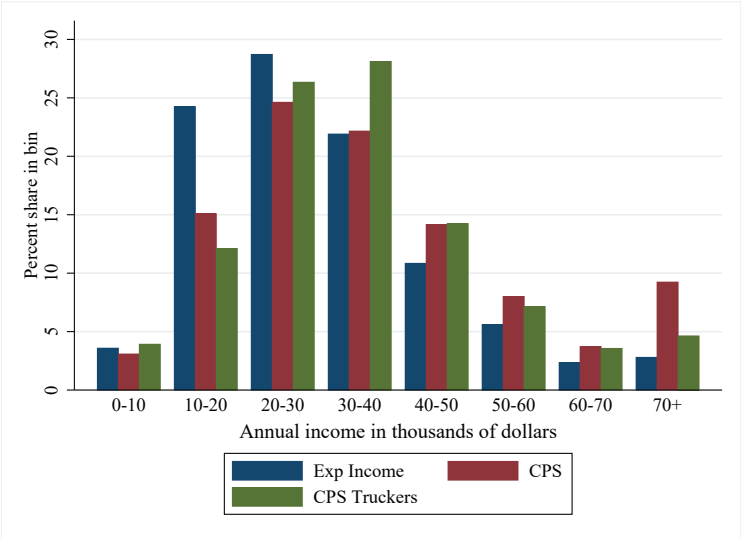
Figure E3: Distribution of Overconfidence Across Drivers



(a) Dist of Individual Overconfidence (b) Predicted Miles and Actual Miles

Notes: This figure presents reduced-form evidence on the distribution of overconfidence across drivers. Panel (a) plots a histogram of driver-level overconfidence, where overconfidence is defined as the difference between average beliefs and average productivity. In panel (b), each driver is represented by a dot located at their average productivity and average beliefs. For both panels, we calculate average productivity by averaging over all the driver's weeks (excluding weeks with 0 miles), and we calculate average beliefs by averaging over all the driver's weeks (excluding weeks with predictions of 0 miles). In panel (a), we restrict attention to drivers with driver-level overconfidence between -500 and 1,500 miles. In Panel (b), the figure is made while dropping any drivers for whom average beliefs or average productivity is greater than 4,000 miles.

Figure E4: Are Workers Overconfident About their Outside Option? A Comparison of Firm A Workers' Believed Outside Option with Earnings of Similar Workers in the CPS



Notes: This figure analyzes worker beliefs about their outside option. During driver training, workers at Firm A were asked "Which range best describes the annual earnings you would normally have expected from your usual jobs (regular and part-time together), if you had not started driver training with [Firm A], and your usual jobs had continued without interruption?" Answers were given in eight intervals: $\$0 - \$10,000$, $\$10,000 - \$20,000$, $\$20,000 - \$30,000$, $\$30,000 - \$40,000$, $\$40,000 - \$50,000$, $\$50,000 - \$60,000$, $\$60,000 - \$70,000$, $\$70,000+$. "Exp Income" is the expected income answer to this question, which is present for drivers in our data. The CPS comparison data are from the 2007 March CPS (also known as the Annual Social and Economic Supplement to the CPS). "CPS" is the income and earnings for 35-year old male workers with a high school degree who worked full-time last year and had positive income and earnings. "CPS Truckers" is the same as "CPS" except it is for "Driver/sales workers and truck drivers" ("Occ"=9130) and uses the age range of 30-40 instead of 35. *Provided we can compare our truckers to the workers in the CPS, there is no evidence that drivers overestimate their outside option.* **Further, in a weekly regression of perceived outside option in dollars on driver beliefs about their inside option in dollars & Table 3 full controls, the coefficient on beliefs about the inside option is only** 0.07 **($p - val = 0.16$), suggesting that perceived inside and outside options are weakly correlated.**

Table E1: Do Productivity Beliefs Predict Quitting? Robustness Check Comparing Above-Median and Below-Median Beliefs

	(1)	(2)	(3)	(4)	(5)
Predicted miles are above their median level (0 or 1)	-0.762			-0.689	-0.940
	(0.274)			(0.305)	(0.338)
Avg miles to date		-0.081	-0.118	-0.023	-0.078
		(0.013)	(0.038)	(0.035)	(0.042)
Demographic Controls	No	Yes	Yes	No	Yes
Work Type Controls	No	Yes	Yes	No	Yes
Observations	8,509	33,374	8,509	8,509	8,509

Notes: This table is similar to Table 3. It differs in that the main regressor is a dummy for whether predicted miles is above its median level (instead of predicted miles in continuous form). The odds-ratios are 0.47, 0.50, and 0.39 for columns 1, 4, and 5, respectively, indicating reductions in quitting of 53%, 50%, and 61% from having above-median subjective beliefs vs. below median beliefs.

Table E2: Do Productivity Beliefs Predict Quitting? Robustness Check with Lagged Values

	(1)	(2)	(3)	(4)	(5)	(6)
L. Predicted Miles	-0.028	-0.036			-0.030	-0.032
	(0.017)	(0.019)			(0.019)	(0.020)
L. Avg miles to date			-0.053	-0.039	0.006	-0.020
			(0.013)	(0.039)	(0.034)	(0.041)
Demographic Controls	No	Yes	Yes	Yes	No	Yes
Work Type Controls	No	Yes	Yes	Yes	No	Yes
Observations	8,343	8,343	32,649	8,343	8,343	8,343

Notes: This table is a robustness check to Table 3, where predicted miles and average miles to date are lagged (instead of un-lagged). We also add an additional column, column 2, which is the column 1 specification plus additional controls. In columns 1, 2, 4, 5 and 6, the sample is restricted to observations with non-missing lagged average miles to date, positive lagged miles beliefs, and lagged miles beliefs less than or equal to 5,000 miles.

Table E3: Do Productivity Beliefs Predict Quitting? Robustness Check with a Person's Average Subjective Belief to Date

	(1)	(2)	(3)	(4)	(5)	(6)
Avg predicted miles to date	-0.031	-0.054			-0.031	-0.038
	(0.016)	(0.025)			(0.017)	(0.027)
L. Avg miles to date			-0.053	-0.075	-0.001	-0.052
			(0.013)	(0.034)	(0.032)	(0.039)
Demographic Controls	No	Yes	Yes	Yes	No	Yes
Work Type Controls	No	Yes	Yes	Yes	No	Yes
Observations	8,493	8,493	32,649	8,493	8,493	8,493

Notes: This table is a robustness check to Table 3, where predicted miles is replaced by average predicted miles to date. We also add an additional column, column 2, which is the column 1 specification plus additional controls. In columns 1, 2, 4, 5 and 6, the sample is restricted to observations with non-missing miles, non-missing lagged average miles to date, non-missing mile beliefs, and positive average mile beliefs to date.

Table E4: Do Workers Update their Subjective Productivity Beliefs?

	(1)	(2)	(3)	(4)	(5)
L. Avg miles to date	0.878	0.622	0.507		0.403
	(0.083)	(0.076)	(0.078)		(0.067)
Tenure X L. Avg miles to date			0.0032		
			(0.0016)		
L^2. Avg miles to date				0.551	
				(0.073)	
L. Miles				0.086	
				(0.015)	
Demographic Controls	No	Yes	Yes	Yes	No
Work Type Controls	No	Yes	Yes	Yes	No
Individual FE	No	No	No	No	Yes
Observations	8,624	8,624	8,624	8,317	8,624
R-squared	0.162	0.335	0.337	0.337	0.614

Notes: Using data from Firm A, this table presents OLS regressions of subjective productivity beliefs on lagged average productivity to date. Standard errors clustered by worker in parentheses. Columns 1-2 show that workers increase their subjective beliefs in response to increases in lagged average productivity to date, as predicted in a normal learning model. Column 3 shows that, as predicted in a normal learning model, workers increase the weight on lagged average productivity to date as worker tenure increases. Column 4 shows that, while agents weigh recent productivity shocks, they place most of the weight on accumulated average productivity to date. Column 5 confirms that updating occurs within worker. All columns include week of tenure dummies. Demographic controls are as in Table 2.

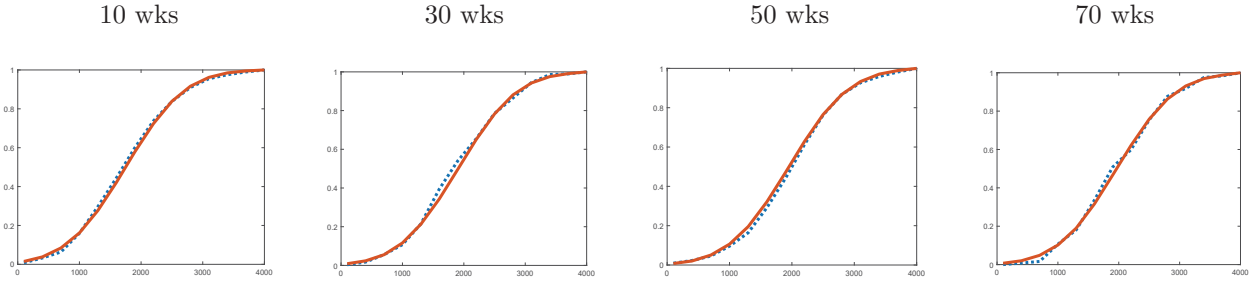
Table E5: Do Productivity Beliefs Predict Productivity? OLS Regressions at Firm B

	(1)	(2)	(3)	(4)	(5)	(6)
L. Pred miles	0.299	0.298	0.264	0.140	0.146	0.056
	(0.051)	(0.053)	(0.064)	(0.052)	(0.053)	(0.053)
L. Avg miles to date				0.571	0.583	
				(0.077)	(0.075)	
\$10 Incentive			-1.955			
			(2.775)			
\$10 Incentive X L. Pred miles			0.085			
			(0.116)			
\$50 Incentive			-4.476			
			(6.860)			
\$50 Incentive X L. Pred miles			0.174			
			(0.309)			
Demographic Controls	No	Yes	Yes	No	Yes	No
Subject FE	No	No	No	No	No	Yes
Observations	803	803	803	695	695	803

Notes: The dependent variable is miles driven per week (in hundreds). An observation is a driver-week. Standard errors clustered by driver in parentheses. All regressions include worker tenure in weeks, dummies for the number of days not worked in a week, and dummies for a worker's week in the study. The demographic controls are gender, age, trucking experience, and region of home residence. These drivers are all from Firm B where we collected subjective productivity forecasts similar to as at Firm A, but randomizing financial incentives for accurate guessing to some workers. As at Firm A, the data here show that productivity beliefs are moderately predictive of actual productivity across workers, but only weakly so within workers. This finding is consistent with our model in Section 4. In addition, we see that there are no statistically significant differences as to whether productivity beliefs are more predictive of actual productivity when they are financially incentivized.

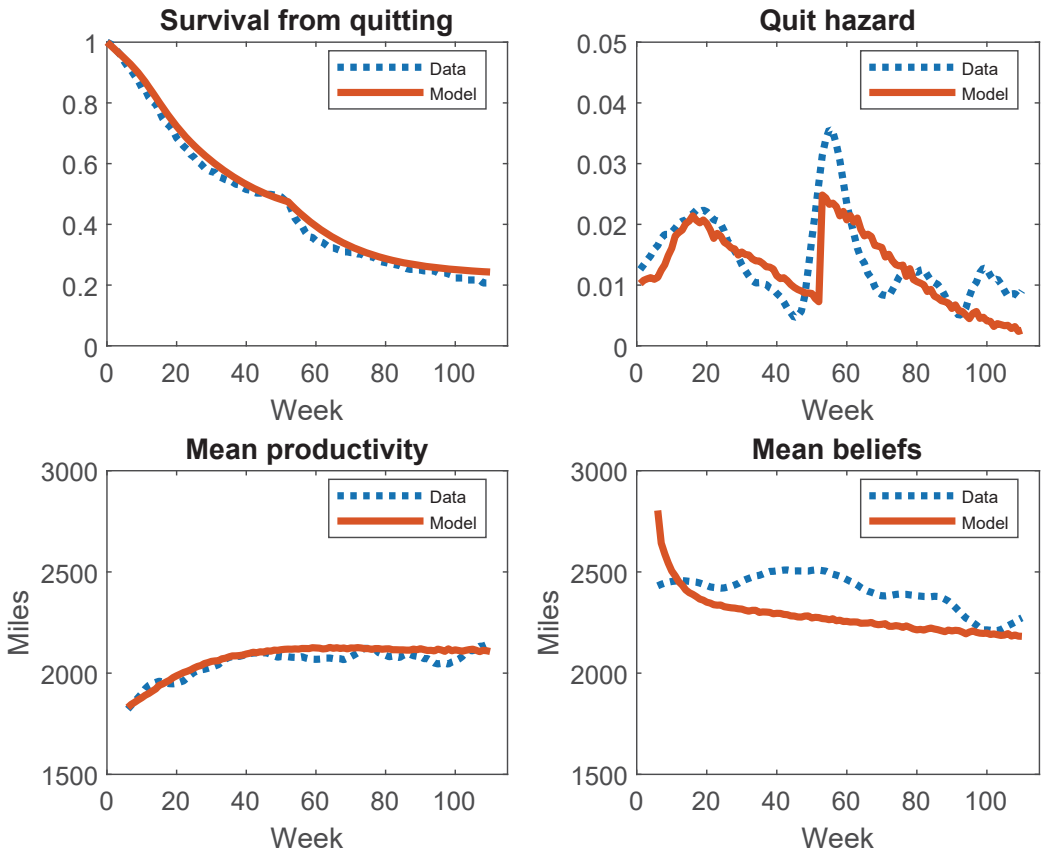
F Further Structural Results

Figure F1: Other Aspects of Model Fit: Tenure and the Distribution of Productivity



Notes: This figure plots the cumulative distribution of productivity at different tenure levels, both in the data and as simulated by the model (from column 2 of Table 5) with 200,000 simulated drivers.

Figure F2: Model Fit: Model Estimated With Overconfidence and Standard Learning



Notes: The notes are the same as for Figure 2 except the underlying model is different. The model is similar to that in Column 2 in Table 5 except that it imposes standard learning, that is, where the perceived variance of the productivity signals equals the actual variance, $\widetilde{\sigma_y} = \sigma_y$. In other words, the model here assumes no variance bias, but allows for mean bias.

Table F1: Robustness Tests of Alternative Model Specifications

		Baseline	Annual $\delta = .90$	IPW	Winsorize beliefs at 4k mi	T=200	Higher outside option	Finer grid
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Productivity and Skill Parameters							
η_0	Mean of prior productivity dist	1993	1993	1990	1993	1993	1993	1993
		(15)	(15)	(15)	(15)	(15)	(15)	(15)
σ_0	Std dev of prior productivity dist	292	289	292	292	292	292	290
		(11)	(11)	(11)	(11)	(11)	(11)	(11)
σ_y	Std dev of productivity shocks	708	708	707	708	708	708	708
		(3.5)	(3.5)	(3.5)	(3.5)	(3.5)	(3.5)	(3.5)
s_0	Value of skilled gain wks 1-5	4.1	2.0	4.8	4.0	4.2	7.5	3.3
		(4.0)	(2.7)	(3.9)	(4.7)	(4.0)	(4.0)	(3.4)
	Taste UH Parameters							
μ_1	Mass point 1 of taste UH	-290	-367	-288	-279	-286	-130	-243
		(20)	(21)	(21)	(25)	(20)	(20)	(13)
μ_2	Mass point 2 of taste UH	-138	-170	-140	-122	-139	23	-127
		(12)	(12)	(13)	(14)	(12)	(12)	(21)
μ_3	Mass point 3 of taste UH	145	132	150	168	124	305	168
		(40)	(40)	(42)	(43)	(37)	(40)	(43)
p_1	Probability type 1	0.31	0.30	0.32	0.29	0.32	0.31	0.48
		(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.08)
p_2	Probability type 2	0.46	0.48	0.46	0.49	0.46	0.46	0.31
		(0.06)	(0.05)	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
	Belief Parameters							
η_b	Belief bias	674	683	661	614	673	674	685
		(32)	(33)	(32)	(26)	(32)	(32)	(30)
$\widetilde{\sigma_y}$	Believed std dev of productivity shocks	1673	1612	1714	1507	1683	1673	1618
		(128)	(121)	(136)	(95)	(130)	(128)	(113)
σ_b	Std dev in beliefs	877	877	870	662	877	877	877
		(8.0)	(8.0)	(8.0)	(6.0)	(8.0)	(8.0)	(8.0)
	Scalar Parameter							
τ	Scale param of idiosyncratic shock	2553	3356	2514	3004	2529	2554	2223
		(450)	(638)	(445)	(563)	(452)	(450)	(306)
	Log-likelihood	-94127	-94132	-93586	-92278	-94127	-94127	-94118
	Number of workers	699	699	699	699	699	699	699

Notes: This table presents a number of robustness checks for our main estimates. Standard errors are in parentheses and are calculated by inverting the Hessian. Column 1 repeats the baseline estimates from column 2 of Table 4. Column 2 sets the discount factor equal to 0.9980, corresponding to an annual discount factor of 0.90. Column 3 uses inverse probability weighting to correct for survey non-response (see Appendix A.1). Column 4 eliminates all the subjective belief observations where the stated belief is greater than 4,000 miles in a week. Column 5 increases the period during which learning about productivity may occur from 130 weeks to 200 weeks. Column 6 raises the outside option r by 25% from \$640 per week to \$800 per week. Column 7 uses a finer grid with increments of 100 miles from 100 miles to 4,000 miles. In columns 2, 3, and 5, we first estimate using "fmincon" in Matlab (imposing α_1, α_2, and α_3 are greater than or equal to -\$375) before running "fminunc" as described in Appendix D for the baseline models without learning by doing. Running "fminunc" at the end ensures that our method of calculating the Hessian (using "fminunc") is comparable across columns.

G Measuring Productivity

General. Drivers at Firm A are primarily paid by the mile. Drivers also receive small additional payments for non-miles related tasks such as going through customs, loading and unloading, scales weighing, working on trailers, and training other drivers. Some drivers are paid based on their activities or on salary instead of by the mile (e.g., drivers who work full-time as instructors at the training schools).

Beyond tenure with the firm, the driver's rate per mile increases with experience outside the firm. However, all the driver we study are new to the industry, so the distinction is not relevant.

Per the federal hours-of-service regulations, truckers in firms like Firms A and B are allowed to work 70 hours over an 8 day period. Per calendar week, this translates to a federal limit of roughly 60 hours per week. See http://www.fmcsa.dot.gov/rules-regulations/topics/hos/index.htm, accessed in October 2010.

To our understanding, good loads are not systematically assigned to good drivers. In addition, there is no scope for boss-worker favoritism, since the driver's boss, with whom he interacts with over the week, does not assign him loads.[32] Firm A is a leading firm with a large number of available loads. During the time period we study, the firm had basic on-board computers (Hubbard, 2003), but drivers were responsible for all route planning and time management.

Measurement Error. Observed miles per week has a small amount of measurement error. We explain the source of the measurement error; describe how we can correct for it; and show that correcting for it has little impact on our paper's estimates and conclusions.[33]

Because miles are only recorded once a worker reaches his destination, miles are imperfectly observed each week. If the driver is in the middle of a load at the end of the week, the miles on that load performed during the concluding week will be counted toward miles on the week just beginning. To correct for measurement error and address how much measurement error affects our results, we requested and analyzed new *load-level data* from Firm A, covering most drivers over a 9-week period. With the new data, we assign half of the mileage from a driver's first load each week to the current week and half to the previous week for any loads spilling over weeks.[34]

Using the new load-level data, we develop a simple algorithm to correct week-level data for measurement error. The load-level data provides "true miles" in a week. We create week-level data, as in the main Firm A dataset, by aggregating loads by week and adding the measurement error. The basic idea for the algorithm is when we observe a low-mileage week followed by a high-mileage week, we transfer some miles from the high to the low mileage week because a small portion of the difference is likely measurement error.

Formally, note that the observed miles in week-level data, y_t^m, is equal to:

$$y_t^m = y_t + \alpha_{t-1} A(t-1, t) - \alpha_t A(t, t+1)$$

where y_t is true miles; $A(t-1, t)$ is the number of miles from a load that started in week $t-1$

[32] We note that even if there were various forms of systematic assignment of loads to drivers, this would not affect the main message or conclusions of the paper, only the interpretation of what drivers are overconfident about. Whether drivers are overconfident about how quick will be at delivering loads or whether they are overconfident about what type of loads they will be assigned, they will still be more likely to sign training contracts and less likely to quit after training, if they are overconfident relative to their outside option.

[33] The measurement error discussed here is also present in the data from Firm B, but we focus the discussion on Firm A. We do this because most of the analysis in the paper is with Firm A data and because we only have load-level data from Firm A.

[34] While most of the Firm A data are from payroll records, the load level data are created from operational records. It should be noted that even for the newer load-level data, we are still not observing actual miles within the week-time exactly. However, we can come much closer to a driver's true productivity in a given week.

and ended on week t; and α_{t-1} is the share of those miles that were completed in week $t-1$. That is, observed miles are true miles, plus spillover miles from the past week to the current week, minus spillover miles from the current week to the next week. We re-arrange this equation to get $y_t = y_t^m - \alpha_{t-1}A(t-1,t) + \alpha_t A(t,t+1)$. To empirically implement our best guess of y, we consider the regression equation:

$$y_{i,t} = y_{i,t}^m - \beta(y_{i,t}^m - y_{i,t-1}^m)) + \beta(y_{i,t+1}^m - y_{i,t}^m) + \epsilon_{i,t}$$

We wish to find the β that leads to the smallest sum of squared deviations between our "corrected" productivity measure (the right-hand side) and the "true" productivity measure (the left-hand side). By moving $y_{i,t}^m$ to the left-hand side, we can estimate this equation by OLS, obtaining $\hat{\beta} = 0.091$, with a standard error of 0.001.

Having developed the algorithm with the load-level data, we can now apply the algorithm to the main Firm A data and do a robustness check on the impact of controlling for measurement error. Re-doing the results in Table 2 on whether productivity beliefs predict productivity, we find very similar results with the measurement error correction as before without it. Further, we re-did the baseline structural results from column 2 of Table 4; the structural estimates are robust to correcting for measurement error.

Appendix References

Akerlof, George A and William T Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, 1982, *72* (3), 307–19.

Anderson, Jon, Stephen V. Burks, Jeffrey Carpenter, Lorenz Goette, Karsten Maurer, Daniele Nosenzo, Ruth Potter, Kim Rocha, and Aldo Rustichini, "Self-selection and variations in the laboratory measurement of other-regarding preferences across subject pools," *Experimental Economics*, 2013, *16* (2), 170–189.

Arcidiacono, Peter, V. Joseph Hotz, and Songman Kang, "Modeling college major choices using elicited measures of expectations and counterfactuals," *Journal of Econometrics*, 2012, *166* (1), 3 – 16.

_ , _ , Arnaud Maurel, and Teresa Romano, "Recovering Ex Ante Returns and Preferences for Occupations using Subjective Expectations Data," Working Paper 20626, National Bureau of Economic Research October 2014.

Arkes, Hal et al., "Two Methods of Reducing Overconfidence," *Org. Behavior & Human Decision Processes*, 1987, *39*, 133–144.

Armantier, Olivier and Nicolas Treich, "Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging," *European Economic Review*, 2013, *62*, 17–40.

Bellemare, Charles, Sabine Kroger, and Arthur van Soest, "Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities," *Econometrica*, 2008, *76* (4), 815–839.

Benabou, Roland and Jean Tirole, "Self-Confidence And Personal Motivation," *QJE*, 2002, *117* (3), 871–915.

Burks, Stephen V., Bo Cowgill, Mitchell Hoffman, and Michael Housman, "The Value of Hiring through Employee Referrals," *Quarterly Journal of Economics*, 2015, *130* (2), 805–839.

_ , Jeffrey Carpenter, Lorenz Goette, and Aldo Rustichini, "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment," *PNAS*, 2009, *106* (19), 7745–7750.

_ , _ , _ , Kristen Monaco, Kay Porter, and Aldo Rustichini, "Using Behavioral Economic Field Experiments at a Firm: The Context and Design of the Truckers and Turnover Project," in "The Analysis of Firms and Employees: Quantitative and Qualitative Approaches" 2008.

Camerer, Colin F and Robin M Hogarth, "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework," *Journal of Risk and Uncertainty*, 1999, *19* (1-3), 7–42.

Cameron, Lisa A, "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Economic Inquiry*, 1999, *37* (1), 47–59.

Chan, Tat, Barton Hamilton, and Christopher Makler, "Using Expectations Data to Infer Managerial Objectives and Choices," 2008. Mimeo, Washington University in St. Louis.

Cherry, Todd, Peter Frykblom, and Jason Shogren, "Hardnose the Dictator," *AER*, 2002, *92* (4), 1218–1221.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman, "Who is (more) rational?," *American Economic Review*, 2014, *104* (6), 1518–1550.

Delavande, Adeline, "Pill, Patch, Or Shot? Subjective Expectations And Birth Control Choice," *International Economic Review*, 2008, *49* (3), 999–1042.

DellaVigna, Stefano, Attila Lindner, Balázs Reizer, and Johannes F Schmieder, "Reference-dependent job search: Evidence from Hungary," *Quarterly Journal of Economics*, 2017, *132* (4), 1969–2018.

Eyster, Erik, "Rationalizing the Past: A Taste for Consistency," 2002. Mimeo, Oxford.

Festinger, Leon, *A Theory of Cognitive Dissonance*, Stanford, CA: Stanford University Press, 1957.

Fischoff, Baruch, "Debiasing," in Paul Slovic Daniel Kahneman and Amos Tversky, eds., *Judgment under Uncertainty: Heuristics and Biases*, University of Chicago Press, 1982, pp. 422–444.

Fleisig, Dida, "Adding Information May Increase Overconfidence In Accuracy Of Knowledge Retrieval," *Psychological Reports*, 2011, *108* (2), 379–392.

Friedman, Daniel and Dominic Massaro, "Understanding Variability in Binary and Continuous Choice," *Psychonomic Bulletin and Review*, 1998, *5* (819), 370389.

Heckman, James J, "Sample Selection Bias as a Specification Error," *Econometrica*, 1979, *47* (1), 153–61.

Hendren, Nathaniel, "Private information and insurance rejections," *Econometrica*, 2013, *81* (5), 1713–1762.

Hoffman, Mitchell, "How is Information Valued? Evidence from Framed Field Experiments," *The Economic Journal*, 2016, *126* (595), 1884–1911.

__ **and Stephen V. Burks**, "Training Contracts, Employee Turnover, and the Returns from Firm-sponsored General Training," 2017. NBER Working Paper 23247.

Holt, Charles A. and Angela M. Smith, "An Update on Bayesian Updating," *JEBO*, 2009, *69* (2), 125 – 134.

Hossain, Tanjim and Ryo Okui, "The Binarized Scoring Rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

Hubbard, Thomas N., "Information, Decisions, and Productivity: On-Board Computers and Capacity Utilization in Trucking," *American Economic Review*, 2003, *93* (4), 1328–1353.

Knittel, Christopher R and Konstantinos Metaxoglou, "Estimation of random-coefficient demand models: Two empiricists' perspective," *Review of Economics and Statistics*, 2014, *96* (1), 34–59.

Larkin, Ian and Stephen Leider, "Incentive Schemes, Sorting and Behavorial Biases of Employees: Experimental Evidence," *American Economic Journal: Microeconomics*, 2012, *4* (2), 184–214.

Lau, Annie Y.S. and Enrico W. Coiera, "Can Cognitive Biases during Consumer Health Information Searches Be Reduced to Improve Decision Making?," *J. of the American Medical Informatics Association*, 2009, *16* (1), 54 – 65.

Levitt, Steven D. and John A. List, "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

Manski, Charles F., "Measuring Expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

Mayraz, Guy, "Priors and Desires: a Model of Payoff Dependent Beliefs," 2011. Mimeo.

Moore, Don A. and Paul J. Healy, "The Trouble With Overconfidence," *Psychological Review*, 2008, *115* (2), 502 – 517.

__ **and Samuel Swift**, "The Three Faces of Overconfidence in Organizations," in Rolf Van Dick and J Keith Murnighan, eds., *Social Psychology and Organizations*, 2010.

Offerman, Theo, Joep Sonnemans, Gijs Van De Kuilen, and Peter Wakker, "A Truth Serum for Non-Bayesians: Correcting Proper Scoring Rules for Risk Attitudes," *Review of Economic Studies*, 2009, *76* (4), 1461–1489.

Pantano, Juan and Yu Zheng, "Using Subjective Expectations Data to Allow for Unobserved Heterogeneity in Hotz-Miller Estimation Strategies," 2010. Mimeo.

Radzevick, Joseph R. and Don A. Moore, "Competing to Be Certain (But Wrong): Market Dynamics and Excessive Confidence in Judgment," *Management Science*, 2011, *57* (1), 93–106.

Roth, Alvin, Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *AER*, 1991, *81* (5), 1068–95.

Rust, John, "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher," *Econometrica*, 1987, *55* (5), 999–1033.

Rustichini, Aldo, Colin G. DeYoung, Jon E. Anderson, and Stephen V. Burks, "Toward the integration of personality theory and decision theory in explaining economic behavior: An experimental investigation," *Journal of Behavioral and Experimental Economics*, 2016, *64*, 122–137.

Sanna, Lawrence J. et al., "When Debiasing Backfires: Accessible Content and Accessibility Experiences in Debiasing Hindsight," *J. of Experimental Psychology: Learning, Memory, and Cognition*, 2002, *28* (3), 497 – 502.

Schlag, Karl H. and Joel van der Weeley, "Eliciting Probabilities, Means, Medians, Variances and Covariances without assuming Risk Neutrality," 2009. Mimeo, Universitat Pompeu Fabra.

Selten, Reinhard, "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1998, *1* (1), 43–61.

Sonnemans, Joep and Theo Offerman, "Is the Quadratic Scoring Rule Really Incentive Compatible?," 2001. Mimeo, CREED, University of Amsterdam.

Stange, Kevin M., "An Empirical Investigation of the Option Value of College Enrollment," *American Economic Journal: Applied Economics*, 2012, *4* (1), 49–84.

Stinebrickner, Ralph and Todd Stinebrickner, "Academic performance and college dropout: Using longitudinal expectations data to estimate a learning model," *Journal of Labor Economics*, 2014, *32* (3), 601–644.

van der Klaauw, Wilbert, "On the Use of Expectations Data in Estimating Structural Dynamic Choice Models," *Journal of Labor Economics*, 2012, *30* (3), 521 – 554.

__ **and Kenneth I. Wolpin**, "Social Security and the Retirement and Savings Behavior of Low-income Households," *Journal of Econometrics*, 2008, *145* (1-2), 21–42.

Wang, Yang, "Dynamic Implications of Subjective Expectations: Evidence from Adult Smokers," *American Economic Journal: Applied Economics*, 2014, *6* (1), 1–37.

Wiswall, Matthew and Basit Zafar, "Determinants of College Major Choice: Identification using an Information Experiment," *Review of Economic Studies*, 2015, *82* (2), 791–824.

Wooldridge, Jeffrey, *Econometric Analysis of Cross Section and Panel Data*, New York, NY: MIT Press, 2002.

Zafar, Basit, "Can subjective expectations data be used in choice models? evidence on cognitive biases," *Journal of Applied Econometrics*, 04 2011, *26* (3), 520–544.