Worker Overconfidence: Field Evidence, Dynamics, and Implications for Employee Turnover and Returns from Training*

Mitchell Hoffman†
Stephen V. Burks‡

November 2016

Abstract

Combining weekly productivity data with weekly productivity beliefs for a large sample of truckers over two years, we show that workers tend to systematically over-predict their productivity and that overprediction is slow to diminish. If workers are overconfident about their own productivity at the current firm relative to their outside option, they should be less likely to quit. Empirically, all else equal, having higher productivity beliefs is associated with an employee being less likely to quit. To study the implications of overconfidence for worker welfare and firm profits, we estimate a structural learning model with biased beliefs that accounts for many key features of the data. While worker overconfidence moderately decreases worker welfare, it also substantially increases firm profits. This may be critical for firms (such as the main one we study) that make large initial investments in worker training.

JEL Classifications: J24, D03, M53, J41
Keywords: Overconfidence; biased learning; turnover; firm-sponsored training

*A previous paper appeared as "Training Contracts, Worker Overconfidence, and the Return from Firm-Sponsored General Training." That previous paper has been divided in two, with the present paper focusing on overconfidence. We are deeply indebted to David Card, Stefano DellaVigna, John Morgan, and Steve Tadelis for their advice and encouragement. For especially detailed comments, we also thank Ben Handel, Ben Hermalin, Ken Judd, Pat Kline, Botond Koszegi, Don Moore, Matthew Rabin, and Kenneth Train. We thank seminar participants at Arizona, Berkeley, Berkeley Haas, Carnegie Mellon, Chicago, CIPREE Conference on Subjective Expectations, Columbia Business School, Econometric Society, Harvard Business School, LSE, MIT, NBER (Labor / Personnel Economics), New York Fed, Northwestern Kellogg, Ohio State, RAND, Rochester Simon, SITE, SED, SOLE, Stanford GSB, Toronto, Toronto Rotman, UBC Sauder, WashU Olin, and Yale SOM. We especially thank the many trucking industry managers and drivers who shared their insights with us. We thank managers at Firms A and B for sharing their data, for facilitating on-site data collection (Firm A), and for helping with the field experiment (Firm B). Graham Beattie, Christina Chew, Dan Ershov, Sandrena Frischer, Will Kuffel, Amol Lingnurkar, Kristjan Sigurdson, and Irina Titova provided outstanding research assistance. Hoffman acknowledges financial support from the National Science Foundation IGERT Fellowship, the Kauffman Foundation, and the Social Science and Humanities Research Council of Canada. Burks and the Truckers & Turnover Project acknowledge financial support from Firm A, the MacArthur Foundation, the Sloan Foundation, the Trucking Industry Program at Georgia Tech, and University of Minnesota, Morris.

†University of Toronto Rotman School of Management and NBER; mitchell.hoffman@rotman.utoronto.ca
‡University of Minnesota, Morris and IZA; svburks@morris.umn.edu

1 Introduction

Going back over 200 years to Adam Smith, it has long been argued that employees have a tendency to be overconfident about their abilities at work. Decades of research in psychology (and growing research in economics) support the idea that people are overconfident. However, this research is primarily based on short-term student lab experiments. Much less is known about overconfidence in the field,[1] and even less so in the context of employee productivity in the workplace. Are workers overconfident about their productivity in an actual workplace setting? How does overconfidence about ability change over time; that is, is overconfidence persistent or does it quickly disappear over time? What are the implications of worker overconfidence for employee behavior, employee welfare, and the profits of firms?

We address these questions using unique data from the trucking industry. While it is one specific industry, trucking is ideal for our study because it is large (see Section 2) and productivity (miles driven per week) is easy to measure. At a leading trucking firm (which we call Firm A), 895 workers were asked to predict their weekly productivity for two years. We show that workers who expect higher productivity end up achieving higher productivity, so subjective beliefs are predictive. However, the data also reveal a pattern where workers tend to systematically overpredict their productivity. Overprediction fades only very slowly. The overprediction we observe without financial incentives remains even when belief-elicitation is made incentive-compatible using randomized financial incentives for accurate prediction at a second large trucking firm (Firm B). We refer to this overprediction as "overconfidence" and say more about the term below.

Having documented this overprediction, we next seek to model it quantitatively, as well as to understand its implications. We turn to Jovanovic's (1979) canonical model of turnover, where quitting decisions reflect the evolution of worker beliefs about job match or productivity. We document that, consistent with theory, workers who expect higher productivity are less likely to quit. From the standpoint of the firm, this may be especially important because the firm is providing the workers with firm-sponsored general training at no direct cost. Turnover is costly for the firm, leading the firm to lose the individuals that they recently provided training to. While potentially useful for firms, if workers are overconfident about their ability at the firm relative to their outside option, this may

[1] For exceptions in economics, see, e.g., Malmendier and Tate (2005), who study CEO overconfidence and investment, and Hoffman (2016), who studies how overconfidence affects businesspeople's demand for information.

distort worker quitting decisions, reducing worker welfare.

To evaluate the importance of overconfidence for worker welfare and firm profits, we develop a structural model of worker turnover. Similar to Jovanovic (1979), workers learn about their underlying productivity through weekly productivity realizations, and decide when, if ever, to quit. However, we do not impose that workers are fully rational. Workers may hold biased priors, or learn faster or slower than predicted by Bayes' rule, nesting the standard model as a special case. Using our rich subjective belief data for identification, we estimate that workers have mean bias of 30-35% of underlying productivity, as well as substantial variance bias. Our model fits the data quite well, whereas a standard model performs far worse. In a counterfactual simulation, we show that eliminating worker overconfidence would moderately increase worker welfare (because workers make better decisions), but would substantially reduce firm profits.

Our study makes three main contributions to the literature. First, we provide long-term high-frequency field evidence on overconfidence, the longest high-frequency evidence we are aware of in any field (psychology or economics). Moore and Healy (2008) provide an excellent survey of recent work and divide overconfidence into three types: relative overconfidence (thinking you are better than others), absolute overconfidence (thinking you are better than you actually are), and overprecision (thinking your beliefs are more precise than they actually are). Our paper's largest focus is on absolute overconfidence, which we refer to hereafter simply as overconfidence. Overconfidence research has mostly focused on short-term laboratory tasks, e.g., trivia games. This paper analyzes overconfidence using weekly data over two years on forecasts about individual productivity in an actual work setting.

Second, we quantify the worker welfare impacts of overconfidence by developing a structural learning model with biased beliefs. We present one of the first papers in economics to estimate a learning model with biased beliefs.[2] More generally, we contribute to a small but growing literature using subjective beliefs in various ways to estimate structural models (for pioneer papers, see, e.g.,

[2]While several recent papers in labor and personnel economics analyze learning using a structural approach (e.g., Arcidiacono, 2004; Bojilov, 2013; Sanders, 2014; Stange, 2012), we allow for both generalized and non-rational learning. Two papers in industrial organization, Goettler and Clay (2011) and Grubb and Osborne (2015), estimate biased learning models of plan choice for online groceries and cell phone service, respectively. A main difference in our paper is that belief biases are identified using high-frequency subjective belief data, whereas in Goettler and Clay (2011) and Grubb and Osborne (2015), biases are identified through contractual choices. There are advantages of each approach. An advantage of using contracts relative to using subjective beliefs is that economists are more trusting of "what people do" compared to "what people say." A virtue of using beliefs is that repeated sub-optimal *ex post* contractual choices may reflect factors other than biased beliefs, including inertia or switching costs.

Chan et al., 2008; van der Klaauw and Wolpin, 2008)[3] as well as to an even smaller literature combining structural estimation with author-conducted field experiments; see Bellemare et al. (2008) and Wiswall and Zafar (2015) for noteworthy examples of both.

Third, we demonstrate that worker overconfidence benefits firms by increasing the profitability of training. Counterfactual simulations suggest biased beliefs are quantitatively important in facilitating training, i.e., training would be substantially less profitable for firms if workers were not overconfident. While a number of field studies analyze how firms may benefit from consumer biases,[4] ours is the first (to our knowledge) to analyze how firms may benefit from biases of their workers.

Section 2 gives background on trucking and describes the data. Section 3 analyzes subjective productivity belief data, both from Firm A (without incentives) and from Firm B (with randomized financial incentives). Section 4 develops the model and structurally estimates it. Section 5 performs the counterfactual simulations. Section 6 concludes.

2 Background and Data

2.1 Institutional Background

Truckdriving in the US. Truckdriving is a large occupation, with roughly 1.8 million US workers operating heavy trucks such as those used by the firms we study (BLS, 2010). Firms A and B are in the long-distance truckload segment of the for-hire trucking industry, which is the largest employment setting for this occupation. An important distinction is between long-haul and short-haul trucking. Long-haul truckload drivers are usually paid by the mile (a piece rate) (Belzer, 2000) and drive long distances from home. In contrast, short-haul truckload drivers generally spend fewer nights away from home and are not usually paid by the mile.[5]

The main training for heavy truckdrivers is that needed to obtain a commercial driver's license (CDL). Most new drivers take a formal CDL training course, and in some states it is required by

[3] See van der Klaauw (2012) for a general discussion on incorporating subjective beliefs into dynamic structural models; see Appendix A.11 for a description of additional papers.

[4] See Koszegi (2014) for a survey.

[5] We highlight a few more institutional details. Truckload is the segment that hauls full trailer loads. Truckload has employee turnover rates, often over 100% per year (Burks et al., 2008), as well as low unionization, and most drivers do not own their own trucks. Around 10% of trucks in 1992 were driven by drivers who own their own truck (owner-operators), with the remainder driven by drivers driving company-owned trucks (company drivers) (Baker and Hubbard, 2004). All the drivers we study are non-union company drivers. For an analysis of productivity in trucking, see Hubbard (2003).

law (BLS, 2010). CDL training can be obtained at truck driving schools run by trucking companies, at private truck driving schools, and at some community colleges. At Firm A, the CDL training drivers received lasted about 2-3 weeks, and included classroom lectures, simulator driving, and actual behind-the-wheel truck driving. The market price for CDL training at private training schools varies, but is often several thousand dollars.

The drivers we study in this paper received training under a 12-month training contract. Under this contract, Firm A pays for the training and in return the driver commits to stay with the firm for a year. If they driver left early, they were fined between $3,500 and $4,000. Drivers did not post a bond, and the firm seemed to collect only about 30% of the penalties owed (despite the firm making very strong efforts to collect the penalties owed); further details on the contracts are provided in a companion paper (Hoffman and Burks, 2016), which studies the contracts in detail.[6]

Production. Truckload drivers haul full loads between a wide variety of locations. While our data do not contain driver hours, drivers are constrained by the federal legal limit of about 60 hrs/week, and managers informed us that drivers often work up to the limit. Firm A loads are assigned via a central dispatching system and are assigned primarily by proximity (as well as hours left up to the federal limit). Once a load is finished, a driver may start a new one.

Productivity in long-haul trucking is measured in miles per week. There are significant cross-driver differences in average productivity, as well as substantial idiosyncratic variation in productivity within drivers. When asked the reason for significant cross-driver differences, managers described various factors including speed, skill at avoiding traffic, route planning (miles are calculated according to a pre-specified distance between two points, not by distance traveled), not getting lost, and coordinating with people to unload the truck. For example, drivers who arrive late to a location may have to wait a long time for their truck to be unloaded, which can be highly detrimental to weekly miles. As for the sources of week-to-week variation, managers emphasized weather, traffic, variable loading/unloading time, and disadvantageous load assignments. Thus, weekly miles, our measure of productivity, reflect both driver performance and effort, as well as factors that drivers do not control and may be difficult to predict *ex ante*. See Appendix G.1 for more on measuring productivity.

[6]The Appendix of Hoffman and Burks (2016) explains how it is common for large truckload firms to provide CDL training.

2.2 Firm A Data

Data Information. To create our datset, we collected subjective beliefs about next week's productivity for a subset of 895 new drivers trained at one of the firm's training schools in late 2005-2006. Records from the firm provide weekly data on miles and earnings. Beyond the productivity beliefs survey, subset drivers did various tests (e.g., IQ, personality) during training, and were invited to do other surveys during their first two years of work (see Appendix A.1). We will sometimes refer to drivers in our data as the "data subset,"[7] and several other papers by the author analyze this subset of drivers in other work.[8] However, the productivity belief data we collect have never been analyzed previously.

Every week around Tuesday,[9] drivers in the data subset were asked to predict their miles for the following pay week (Sunday-Saturday, starting on the Sunday in 5 days). Drivers responded by typing an answer to the below question, which we sent over the truck's computer system: *About how many paid miles do you expect to run during your next pay week?* We interpret this question as asking drivers for their subjective mean.[10] There are several potential concerns with using our beliefs question to predict behavior and study overconfidence:

1. Researchers might worry that beliefs are stated to please others, e.g., drivers exaggerate their productivity beliefs to please their boss. However, in our setting, drivers were informed repeatedly that their responses and participation were never to be shared with the company. That is, driver supervisors would never even know whether a worker participated in the survey, let

[7] We use the term "data subset" to distinguish it from the full sample of drivers at Firm A (for whom there is regular personnel data, but no beliefs data) who are studied in Hoffman and Burks (2016) and Burks et al. (2015).

[8] Appendix A.12 describes several unrelated papers using the data subset (e.g., comparing social preferences of truckers, students, and non-trucker adults). Burks et al. (2013) analyze new truckers predicting their quintile on an IQ test to test between different theories of relative overconfidence (people tending to overestimate how well they do compared to other people). Our paper differs from Burks et al. (2013) in that we study absolute overconfidence instead of relative overconfidence; we study beliefs about productivity instead of about performance on an IQ test; and we study beliefs over time instead of at a single point in time. In addition, Burks et al. (2013) is focused on testing between different theories of the *causes* of relative overconfidence across people, whereas our paper focuses on the *consequences* of absolute overconfidence for worker behavior and contract design. Although the papers deal with quite different issues, we view the contributions as complementary. Burks et al. (2008) describe the Firm A data collection in detail. 1,069 drivers took part in data collection during training. We restrict our sample to drivers with a code denoting no prior trucking experience or training, giving us 895 drivers whom we are confident are brand-new to trucking.

[9] The question was sent to drivers on Tuesday in 85% of driver-weeks, with the remainder on nearby days (details in Appendix A.6).

[10] Another possible interpretation is that it is asking drivers for the median of their subjective mile distribution for next week. In the data subset, mean and median miles are almost identical (the median of worker miles per week is 1% less than the mean miles per week). Thus, whether workers reported their mean or median expected miles seems unlikely to matter for the reduced-form or structural estimation. See Appendix A.6 for further discussion of belief elicitation methods, as well as the issue of lumpy beliefs /possible rounding.

alone what his responses were.

2. No incentives were used to incentivize accurate belief responses. However, as we discuss below in our field experiment with Firm B (see Section 3.2), we find no evidence that beliefs are different when workers are rewarded for accurate beliefs.

3. There is substantial non-response: the average response rate to the weekly beliefs survey is 28%. A 28% response rate may seem low, but is comparable to that in many nongovernmental surveys. For example, in an influential recent study, Card et al. (2012) find a response rate of 20% in a survey of UC Berkeley employees. In Appendix A.1, using Inverse Probability Weighting and a Heckit selection model (using the response rate to prior surveys other than the productivity beliefs survey to form an exclusion restriction), we provide evidence that non-response bias is limited and is not an important driver of our results.

One limitation of personnel data is we generally do not see where drivers go when they terminate. Fortunately, we did an "exit survey" by mail for drivers in the data subset. In drivers leaving the firm, the vast majority are not moving to long-haul trucking jobs. Specifically, about 48% of drivers report moving to a non-trucking job or unemployment, and 25% went to a local driving job. Only 12% report moving to a long-haul trucking job, and 15% to a regional trucking job. See Appendix A.5 for more on the exit survey.

Summary Statistics. Panel A of Table 1 present sample means on driver characteristics. The median data subset driver is male, white, and 35 years old. Drivers have very low average credit scores. Of the 88% of drivers with credit scores (12% lack a sufficient credit history to have a score), the mean and median credit scores are 586 and 564, respectively, compared to a median of 723 for the US general population at the time of data collection; further, 53% of drivers have a credit score below 600 (i.e., "subprime"), compared to 15% of the US population (Appendix A.4).

Panel B provides quantiles of productivity and productivity beliefs for our main sample, as well as for the sample of 699 drivers used to estimate the structural model. That productivity beliefs exceed productivity on average is easily observed in these simple statistics. In our estimation sample, the median productivity belief corresponds with roughly the 75th percentile in the distribution of actual productivity.

3 Reduced Form Analysis

In this section, we show that, while subjective beliefs are predictive about actual productivity and employee turnover, workers also exhibit a tendency to overpredict productivity. We first present our main results from about two years of non-incentivized belief data from Firm A, and then present the Firm B incentivized data to show the results are robust to incentives.

3.1 Firm A Data

Predicting Productivity. Table 2 shows that beliefs help predict productivity beyond other predictors. We estimate:

$$y_{i,t} = \alpha + \beta b_{i,t-1} + \gamma \bar{y}_{i,t-1} + X_i \delta + \epsilon_{i,t} \tag{1}$$

where $y_{i,t}$ is driver i's productivity in his t^{th} week with the company; $b_{i,t-1}$ is his subjective belief about his productivity in week t stated in week $t-1$; $\bar{y}_{i,t-1}$ is lagged average productivity to date; and X_i are controls. Column 2 estimates $\hat{\beta} = 0.15$, meaning, a driver whose expectation is 100 miles higher than another driver will end up driving an average of 15 miles more. Once average productivity to date or driver fixed effects are added, the coefficient drops to between 0.07 and 0.08. That is, *within person*, subjective beliefs have some predictive power, but less so. Overall, the results suggest that subjective beliefs have informational content, being somewhat predictive across individuals, and mildly predictive within individuals. The relatively low coefficients likely reflect attenuation bias due to measurement error in subjective beliefs (which we account for in the structural model).

Predicting Quitting. Table 3 shows that quitting decisions reflect subjective beliefs outside of predictors in a standard Bayesian model. We estimate Cox proportional hazard models of quitting of the form:

$$log(h_{i,t}) = \alpha_t + \beta b_{i,t} + \gamma \bar{y}_{i,t} + X_i \delta, \tag{2}$$

where h_{it} is the quit hazard of driver i with t weeks of tenure and α_t is the log baseline hazard. Average productivity to date, $\bar{y}_{i,t}$ is a sufficient statistic for beliefs about productivity in a standard Bayesian normal learning model. However, a 100 mile increase in subjective miles predicts a 6% decrease in the probability a worker quits. The true effects are likely higher, with observed estimates biased downward due to measurement error. The coefficient on beliefs does not change very much

as controls are added.

The finding that having higher productivity beliefs is associated with a lower chance of quitting is robust. To show that the finding is not driven by outliers, Appendix Table E2 repeats Table 3 using a dummy for beliefs being above the median (as opposed to a continuous measure) and finds sizeable impacts. Our result remains when we use lagged beliefs (Appendix Table E3) or a worker's average belief to date (Appendix Table E4), the latter which aims to measure beliefs more of as a stable worker characteristic. These two checks help assuage the concern of reverse causality (e.g., one concern is that people who expect to quit in the future might believe that they will slow down and drive fewer miles). These two checks yield the same result.[11] In contrast to a "standard" setup where workers hold the same beliefs given their productivity signals, workers' heterogeneous subjective beliefs predict quitting.

Overprediction. Although beliefs are predictive, Figure 1 shows that average beliefs consistently exceed productivity. Productivity and beliefs are collapsed by week of tenure and then smoothed using a local polynomial regression. Workers initially overpredict by roughly 500 miles per week, about 25% of average productivity. Overprediction declines over time, though it is persistent and decreases slowly. Even after 100 weeks, workers still overpredict by around 150 miles per week. Two concerns with Panels (a) and (b) are (i) The sample changes over time (due to quits) and (ii) The productivity line is based on all workers whereas the belief line is based on workers who respond to the survey in a given week. To address (i), we restrict the sample to workers who are there for most of the sample period (at least 75 weeks) in Panel (c). To address (ii), in Panel (d), we re-make the picture dropping the 38% of workers who never respond to the survey. We restrict to workers who are there at least 75 weeks, and look at medians instead of means (to verify results are not driven by outliers). In both cases, the overall pattern of overconfidence is similar, though standard errors are larger.[12]

[11] In additional unreported robustness checks with squared beliefs or dividing beliefs into various quantiles, the relationship between beliefs and quitting is relatively linear. Despite the different checks and despite the fact that the coefficient on beliefs does not change much by including observable variables, it is possible that there may be selection on unobserved variables. Thus, we interpret the results here as evidence that overconfidence correlates with fewer quits instead of that overconfidence causes fewer quits. In addition to these checks, one might also think to include some version of (Beliefs - Productivity) instead of Beliefs as a regressor. However, as we discuss after Prop. 1 in Appendix C, it is the level of a person's perceived inside and outside options that affects quitting in theory, not overprediction. Further, including (Beliefs-Productivity) as a regressor imposes the restriction that coefficients on Beliefs and Productivity are the same.

[12] In Panels (c) and (d), we stop at 75 weeks instead of the full sample to increase sample size. However, results are similar if we restrict to workers who are there for 100 weeks. More generally, we have made the basic graph comparing

The average results mask that beliefs are sensible in several ways, and there is a lot of heterogeneity within and across drivers. First, beliefs exhibit aspects of Bayesian updating. Increases in average productivity to date are associated with substantial increases in future beliefs, both across and within drivers, and the weight on average productivity to date increases with tenure (Appendix Table E5). Second, although beliefs exceed miles in almost every week when averaged over all drivers, individual beliefs exceed miles only 65% of the time; in 35% of driver-weeks, drivers underpredict, so it's not the case that each driver overpredicts each week. Third, drivers differ substantially in average overprediction. Appendix Figure E4 shows that many drivers are moderately overconfident, some are well-calibrated, and some are very overconfident. Fourth, as mentioned above in Section 2.1, there is a lot of week-to-week variation in productivity that drivers don't control.

That workers' beliefs are sensible in several ways suggests to us that their beliefs are plausible (and not a mark simply of people not taking the survey seriously). In addition, Huffman and Shvets (2016) have recently collected data that replicate our results. In particular, after our paper first appeared, Huffman and Shvets (2016) worked with a firm where stores managers were asked to predict their quintile in a tournament where store managers competed against one another in terms of quarterly store performance. In their paper-in-progress, Huffman and Shvets (2016) also find evidence of persistent overconfidence (with overprediction even among managers who had been there for two years or more).[13] Thus, our long-term field evidence on overconfidence in the workplace has recently been replicated in a very different context.

There are several interpretations of our result that workers tend to systematically over-predict productivity and that this over-prediction is slow to decline. For example, workers may report aspirations instead of true expectations. Or, workers may report expected miles supposing that "everything goes well" and there are no unexpected hiccups. For both these explanations, one might

productivity and productivity beliefs a number of different ways, varying means vs. medians, restricting to workers with different tenure levels, and restricting based on survey response (all subjects, excluding subjects who never respond, and restricting only to weeks where both subject productivity and productivity belief are available). The basic path of overprediction is similar across these different specifications. In Appendix Figure E3, we also plot Beliefs - Productivity as a function of tenure, and observe a similar pattern of slowly decreasing overconfidence.

[13] About 48% of managers predicted a better quintile than they actually achieved. Huffman and Shvets (2016) interpret their evidence in support of selective memory as a mechanism for overconfidence, e.g., people tend to selectively forget the times where they don't do well (an interpretation bolstered by asking managers to recall performance in a past tournament, and observing that those who did poorly tend to be more forgetful). This is broadly related to, but different from our modeling in the structural model, where we allow people's perception of signal precision to potentially differ from true signal precision. In our setting, because a majority of people exhibit overprediction, they often get signals that are worse than their beliefs; thus, ignoring or forgetting bad signals is similar to thinking the signal is less precise than it actually is. While this evidence replicates our finding of persistence workplace overconfidence, it is also different because it studies relative (instead of absolute) overconfidence.

imagine that misprediction could be eliminated if workers were incentivized to state the mean of their subjective productivity distribution. Alternatively, overprediction may reflect workers' true beliefs and instead reflect a persistent behavioral bias that would be hard to eliminate with an incentive; overprediction may fade only slowly, given both substantial idiosyncratic variation in miles and given potential variance bias. To distinguish between these explanations, we turn to incentivized data.

3.2 Incentivized belief data from Firm B

To distinguish between these different explanations and to overcome other concerns with non-incentivized data (e.g., that non-incentivized subjects do not "think hard" enough about their forecasts), we randomized financial incentives for truckers at another large trucking company, Firm B, to accurately guess about their productivity. 272 workers were randomly assigned to guess without financial incentives or to receive up to $10 per week for guessing about their productivity.[14] Subjects did this for about 2-6 weeks before being re-assigned to another treatment: control (nothing changes), increased incentive (up to $50 per week), or "debiasing."[15] See Appendix B for further information (e.g., how stake size was chosen).

Appendix Table B2 shows that neither the $10 incentive nor the $50 incentive had a significant impact on productivity beliefs. Given that the standard errors are moderately sized, we can't rule out moderate-sized effects in either direction. For example, using column 1 of Table B2, the 95% confidence interval on the impact of the $10 incentive was -130 miles to +68 miles, and the 95% confidence interval on the impact of the $50 incentive was -181 miles to +171 miles. However, given a mean overconfidence level of 253 miles in the column 1 sample, we are able to reject the hypothesis that *all* the observed overconfidence would disappear if workers were given either $10 or $50 incentives.

[14] While we focus on productivity, we also had subjects guess about their weekly earnings (see Appendix B).

[15] "Debiasing" refers to an additional experiment treatment where we provided information about the existence of overconfidence in truckers so as to see if overconfidence could be reduced. Providing information about the existence of overconfidence led to some decreases in productivity beliefs, but impacts seemed to fade with time since treatment, giving us limited power to examine whether randomized changes in beliefs affected quitting. Discussion is left to Appendix B.

4 Model and Structural Estimation

Our reduced-form analysis suggests that workers overpredict their productivity, as well as that greater beliefs are correlated with a lower chance of quitting. We now develop a structural model of quitting and belief formation to help understand these results and to do counterfactuals of changing beliefs. With the estimated model, we can look at how much overconfidences affect training profits and welfare, and look how overconfidence affects the optimal training contract.

The model is a discrete time extension of the model in Jovanovic (1979) allowing for biased beliefs. A worker decides each week whether to quit his job. It is an optimal stopping problem; once he quits, he cannot return.[16] Quitting is the only decision to make—in particular, there is no effort decision.[17] Workers have different underlying productivities, but productivity is initially unknown, both to the worker and the firm. The worker is forward-looking in his quitting decision and each week's miles provides him a noisy signal from which he learns about his underlying productivity. However, workers may be subject to belief biases. The worker's priors need not be accurate, e.g., he may believe that the job is on average quite lucrative. Further, as new productivity information arrives, he may over- or underweight his prior relative to pure Bayesian updating. In addition to reflecting productivity beliefs, quitting decisions will also reflect a driver's underlying taste for the job or career (e.g., how much a driver dislikes being away from home) as well as idiosyncratic shocks (e.g., a fight with the boss). In the model, the piece rate and training contracts are taken as given; both are endogenized later in Section 5 when we consider optimal contract design for the firm.[18]

4.1 Model Setup

The time horizon is infinite and given in weeks 1, 2, ... Workers have baseline productivity η, which is distributed $N\left(\eta_0, \sigma_0^2\right)$. Workers are paid by a piece rate, w_t, that depends on their tenure.

[16] While workers who quit Firm A are allowed to re-apply, relatively few return. Of inexperienced workers starting in 2002-2003 who quit during 2002-2003, less than 8% return by the end of 2009.

[17] Effort decisions are not included in most related structural learning models (e.g., Arcidiacono, 2004; Stange, 2012). Our data do not contain exogenous variation in the piece rate that would be needed to plausibly identify the cost of effort function. We speculate, however, that including effort in the model would not qualitatively affect our main conclusions or would actually strengthen them (see Appendix A.7 for further discussion).

[18] In the model we estimate, the firm makes no decisions. In addition to taking the piece rate and training contract as fixed, the firm is assumed not to fire workers. In the data subset, quitting is over 3 times more common than firing, and ignoring firing enormously simplifies the model by preventing us from having to estimate a dynamic game. In Section 5, the firm is assumed to offer a training contract and piece rate, which the worker must be willing to accept before work begins.

Workers know the piece rate-tenure profile, and believe that this profile will not be changed by the company at some future date.[19] A worker's weekly miles, y_t, are distributed $N\left(a\left(t\right)+\eta,\sigma_y^2\right)$,[20] and weekly earnings are thus $Y_t = w_t y_t$. $a(t)$ is a known learn-by-doing process, which we specify below. The worker's outside option is r_t and also depends on his tenure. Every period t, the worker makes a decision, d_t, whether to stay ($d_t = 1$) or to quit ($d_t = 0$). Workers make the decision to quit in t having observed their past miles $y_1, y_2, ..., y_{t-1}$, but not their current week miles, y_t. Workers and firms are assumed to be risk-neutral and to have a discount factor given by δ.[21]

Stay-or-Quit Decisions. Workers make their stay-or-quit decisions every period to maximize perceived expected utility:

$$V_t(\mathbf{x}_t) = \max_{d_t, d_{t+1}, \ldots} E_t\left(\sum_{s=t}^{\infty} \delta^{s-t} u_s\left(d_s, \mathbf{x}_s\right) | d_t, \mathbf{x}_t\right). \tag{3}$$

where $\mathbf{x}_t$ is the vector of state variables ($\mathbf{x}_t$ includes past miles, $y_1, ..., y_{t-1}$, and is detailed further below). (3) can be written as a Bellman Equation: $V_t(\mathbf{x}_t) = \max_{d_t} E_t\left(u_t\left(d_t, \mathbf{x}_t\right) + \delta V_{t+1}(\mathbf{x}_{t+1}) | d_t, \mathbf{x}_t\right)$.

The per-period utility from staying at the job is equal to the sum of the worker's non-pecuniary taste for the job, earnings, and an idiosyncratic shock:

$$u_t\left(1, \mathbf{x}_t\right) = \alpha + w_t y_t + \varepsilon_t^S,$$

where α is the worker's non-pecuniary taste for the job, and ε_t^S is an i.i.d. idiosyncratic error unobserved to the econometrician (but observed by the worker) with an Extreme Value-Type 1 distribution and scale parameter τ. Since workers likely differ unobservedly in taste for the job, we assume there is unobserved heterogeneity in non-pecuniary taste for the job, α, with α drawn from a mass-point distribution (Heckman and Singer, 1984).

If the worker quits, he may have to pay a fine associated with the training contract. Let the vector $\mathbf{k}$ denote the training contract, with k_t the penalty for quitting at tenure t. The utility from

[19] Assumptions of this form are standard in structural labor and personnel economics, and allows us to avoid having to specify beliefs over possible future firm policy changes. We believe the assumption is reasonable in our setting, given it is not common for the firm to make large changes in the pay schedule.

[20] Assuming that signals are normally distributed is standard in structural learning models (see the survey by Ching et al. (2013)). Visually, the distribution of signals (miles) among all workers has a bell shape centered close to around 2,000 miles, suggesting this assumption is reasonable (and that the distribution is closer to normal than to log-normal or uniform).

[21] Risk neutrality is assumed in many dynamic learning models (e.g., Crawford and Shum, 2005; Nagypal, 2007; Stange, 2012; Goettler and Clay, 2011), though not in all (for examples with risk aversion, see the survey by Ching et al. (2013)). Coscelli and Shum (2004) show that risk parameters are not identified in certain classes of learning models.

quitting is the fine, plus the discounted value of his outside option, plus an idiosyncratic shock:

$$u_t\left(0,\mathbf{x}_t\right)=-k_t+\frac{r_t}{1-\delta}+\varepsilon_t^Q,$$

where ε_t^Q is an i.i.d. unobserved idiosyncratic error with the same distribution as ε_t^S.[22] Let $V_t^S \equiv E_t\left(u_t\left(1,\mathbf{x}_t\right)+\delta V_t(\mathbf{x}_{t+1})|1,\mathbf{x}_t\right)$ and $V_t^Q \equiv E_t\left(u_t\left(0,\mathbf{x}_t\right)+\delta V_t(\mathbf{x}_{t+1})|0,\mathbf{x}_t\right)$ be the choice-specific value functions for staying and quitting, respectively. Plugging in for $u_t\left(1,\mathbf{x}_t\right)$ and $u_t\left(0,\mathbf{x}_t\right)$, the choice-specific value functions are given by:

$$\begin{aligned} V_t^Q &= -k_t+\frac{r_t}{1-\delta}+\varepsilon_t^Q \equiv \overline{V}_t^Q+\varepsilon_t^Q \\ V_t^S &= \alpha+E_t(w_t y_t|\mathbf{x}_t)+\delta E(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t)+\varepsilon_t^S \equiv \overline{V}_t^S+\varepsilon_t^S, \end{aligned}$$

and the Bellman Equation can be re-written as $V_t(\mathbf{x}_t)=\max_{d_t\in\{0,1\}}\left(V_t^S\left(\mathbf{x}_t\right),V_t^Q\left(\mathbf{x}_t\right)\right)$.

Agents gradually learn their productivity as more and more productivity signals are observed. Thus, after a sufficiently large number of periods, T, the value function can be approximated by the following asymptotic value functions:

$$\begin{aligned} V^Q &= \frac{r_T}{1-\delta}+\varepsilon^Q \equiv \overline{V}^Q+\varepsilon^Q \\ V^S &= \alpha+w_T\eta+\delta E(V(\mathbf{x}')|\mathbf{x})+\varepsilon^S \equiv \overline{V}^S+\varepsilon^S \\ V(\mathbf{x}) &= \max_{d\in\{0,1\}}\left(V^S\left(\mathbf{x}\right),V^Q\left(\mathbf{x}\right)\right) \end{aligned}$$

Belief Formation. In a standard normal learning model, a worker's beliefs about his period t productivity equals the weighted sum of his prior and his demeaned average productivity to date:

$$E(y_t|y_1,...,y_{t-1})=\frac{\sigma_y^2}{(t-1)\,\sigma_0^2+\sigma_y^2}\eta_0+\frac{(t-1)\,\sigma_0^2}{(t-1)\,\sigma_0^2+\sigma_y^2}\frac{\sum_{s=1}^{t-1}y_s-a(s)}{s-1}+a(t) \tag{4}$$

As t increases, the agent eventually shifts all the weight from his prior to his average productivity signals. We augment the standard learning model in two ways. First, we allow for agents to be overconfident: instead of believing that their productivity, η, is drawn from a distribution $N\left(\eta_0,\sigma_0^2\right)$, agents believe η is drawn from a distribution $N\left(\eta_0+\eta_b,\sigma_0^2\right)$. Second, we allow for agents to have a perception of signal noise that may be different from the true signal noise: workers perceive the standard deviation of weekly productivity signals to be $\widetilde{\sigma_y}$ instead of σ_y. With these two assumptions,

[22] Even though only a portion of the penalties owed were collected, as described in Section 2, we assume that drivers act as if the utility cost of quitting is equivalent to the utility loss from paying the contract penalty. We believe this assumption is reasonable. Firm A was very firm with new drivers about its intention to collect money owed upon a quit. After a quit, drivers who did not pay faced aggressive collection contacts by both Firm A and collection agencies, as well as the reporting of delinquency to credit agencies. As a robustness check, we have experimented with estimating versions of the model assuming drivers act as if the utility loss from quitting is 0.3 times the penalty. Model fit tended to be less good. Indeed, our preferred model still fails to fully match the quitting spike at one year, as seen in Figure 2.

an agent's subjective expectation of his productivity, denoted by E^b (where b stands for belief), is:

$$E^b(y_t|y_1,...,y_{t-1}) = \frac{\widetilde{\sigma_y}^2}{(t-1)\,\sigma_0^2 + \widetilde{\sigma_y}^2}(\eta_0+\eta_b) + \frac{(t-1)\,\sigma_0^2}{(t-1)\,\sigma_0^2 + \widetilde{\sigma_y}^2}\frac{\sum_{s=1}^{t-1} y_s - a(s)}{s-1} + a(t) \qquad (5)$$

If η_b is greater (less) than zero, then agents exhibit positive (negative) mean bias or overconfidence (underconfidence). As more signals come in, agents will learn not to be overconfident, eventually putting zero weight on $(\eta_0+\eta_b)$. The speed at which this occurs, however, will be determined by $\widetilde{\sigma_y}$.

We allow that workers' reported subjective beliefs include some measurement error, as accurately reporting one's beliefs about productivity may be unusual or unfamiliar for a worker. We assume that reported beliefs equal underlying subjective beliefs plus a normally distributed error. The reported subjective belief, b_{it}, of driver i at tenure week t is distributed: $b_{it} \sim N\left(E^b\left(y_{it}|y_{i1},...,y_{it-1}\right),\sigma_b^2\right)$.

Summary of Within Period Timing. The within period timing in week t is as follows:

1. Workers form beliefs b_t given past miles $y_1, y_2, ..., y_{t-1}$.
2. ε_t^S and ε_t^Q are realized and workers decide whether or not to quit.
3. y_t is realized, if they do not quit.

Learning by Doing and Skill Accumulation. Productivity increases with the learning by doing function $a(t) = 2a_1 * (\Lambda(a_2 t) - .5)$, where $\Lambda(x) = \frac{e(x)}{1+e(x)}$ and t is worker tenure in weeks. $a(t)$ depends only on tenure; thus, the speed of learning by doing does not depend on the number of miles driven or on the ability of the driver. Workers fully anticipate the path of $a(t)$.[23]

We also account for skill accumulation following CDL training. After CDL training at Firm A, drivers do "on-the-job training" which includes driving with an experienced driver riding along. We use a length of 5 weeks for on-the-job training.[24] We account for the possibility that drivers may gain valuable skills during this time: we assume the outside option over time is $r_t = r - \frac{6-min\{t,6\}}{5}s_0$. We fix r using outside data, while s_0, the value of skills from on-the-job training, is estimated. (Besides allowing for skill accumulation during the first 5 weeks, we alternatively estimate the model allowing for continuous skill accumulation: $r_t = r + 2\theta_1 * (\Lambda(\theta_2 t) - .5)$, where $\Lambda(x) = \frac{e(x)}{1+e(x)}$, and θ_1 and θ_2 are parameters to estimate.)

[23] The logistic functional form is consistent with Jovanovic and Nyarko's (1996) micro-founded model of learning by doing in which the speed of learning decreases over time, as well as the empirical results on tenure and productivity in Shaw and Lazear (2008). Here, a_1 is the total amount by which productivity increases and a_2 indicates the speed of learning by doing. We believe our assumption that workers fully anticipate the learning by doing process is reasonable in our setting. In interviews, managers often referred to a steep "learning curve" for rookie drivers.

[24] During this time, drivers often are paid by flat salary instead of by mile. We use a flat salary of $375 per week during on-the-job training. We also assume drivers do not begin learning about their productivity until after 5 weeks.

Solving the Model. The state variables consist of past miles, the piece rate, the training contract, taste heterogeneity, a person's level of overconfidence, a vector of observable additional characteristics (X), and the idiosyncratic shocks: $\mathbf{x}_t = (y_1, ..., y_{t-1}, \mathbf{w}, \mathbf{k}, \alpha, \eta_b, X, \varepsilon)$. The model can allow for heterogeneity in taste for the job and/or in overconfidence. To solve the model, we first solve for the asymptotic value functions (after all learning has taken placed) using value function iteration. With the asymptotic value functions in hand, backward recursion can then be applied to solve the dynamic programming problem. We provide further details in Appendix D.

4.2 Discussion of Model Assumptions

Outside option. In our model, the outside option, r_t, depends on tenure, but not productivity. This feature differs from many models of firm-sponsored general training where the worker is paid the same share of his marginal product at both his inside and outside option (though less than his full marginal product at both), e.g., Acemoglu and Pischke (1999b). We believe our assumption is realistic in our context given that only 12% of workers who exit report moving to a long-haul trucking job, with the vast majority moving to another type of work (see Section 2.2). Having high ability in long-haul trucking does not necessarily imply that one will have high ability in non-trucking jobs or even in short-haul trucking.[25]

In addition, our assumption about the outside option is consistent with an earlier finding that, all else equal, workers with higher productivity to date are substantially less likely to quit. That is, if workers had the same productivity in their inside and outside options and were paid the same share of their marginal product at each, then high and low ability workers would be equally likely to quit. As seen in columns 2-3 of Table 3, this is not the case.

Beliefs. While our model allows for non-standard belief formation, these features are estimated from data instead of imposed. The model does not assume that people have overconfident priors or learn more slowly than would be predicted by Bayes' Rule, but rather these features are identified via the belief data (see Section 4.3); our model nests the standard model as a special case. Several aspects of our generalized normal learning model receive support from the results in Section 3. Differences

[25] As discussed in Section 2, in contrast to long-haul drivers who are usually paid per mile, short-haul drivers are usually paid by the hour. For short-haul drivers, relationship-management skills (for managing customer and client relationships) are more important than being able to do a lot of miles quickly while far from home. Some workers will be better at long-haul whereas others are will be better at short-haul.

in subjective beliefs are moderately predictive of differences in productivity across workers, but only mildly predictive within workers. This finding is consistent (broadly) with our modeling assumption that workers do not have private information about their underlying productivity. Subjective beliefs do, however, affect quitting decisions in our model, which is consistent with our earlier empirical finding that, all else equal, workers with greater belief bias are less likely to quit.

A strong assumption in the model is that workers are not overconfident about their outside option despite potentially being overconfident about their current job ability. However, for overconfidence to "lock in" workers after training, this assumption is stronger than necessary. Instead, overconfidence will reduce quitting if the worker is more overconfident about his current job earnings than his outside option (see Prop. 1 in Appendix C); i.e., he exhibits *differential overconfidence*. If the strong assumption of no outside overconfidence fails, but workers are still differentially overconfident, overconfidence will still theoretically reduce quitting, but less so than if the strong assumption held. While the assumption of differential overconfidence is difficult to test, we present 5 pieces of evidence and arguments on why it seems reasonable in our setting. Though no piece individually is foolproof, together, the 5 pieces significantly support the assumption of differential overconfidence.

1. Insofar as drivers select the job at which they believe their ability will be the highest, this may lead them toward being differentially overconfident about their ability at the current job relative to the outside option.[26]

2. We collected data on workers' *perceived outside options*. Drivers in the data subset were asked what their earnings would have been had they not started work with Firm A. First, we compare drivers' response to this question to what "similar-looking" people earned in the March 2006 Current Population Survey (CPS). As Appendix Figure E5 shows, the perceived outside option workers would have earned had they not gone through training does not appear to be higher than what people like them in the CPS are earning. Second, there is little correlation between a worker's perceived inside option and his perceived outside option (see bolded text in the notes

[26] Van den Steen (2004) provides a "winner's curse" argument on how self-selection can promote belief biases. Consider a worker choosing among several jobs. For each job, he receives a noisy signal about his productivity there. The agent will naturally choose the job with the highest signal, and will be overconfident there relative to other jobs. While it is easy to imagine that workers may have different beliefs when choosing between long-haul trucking jobs and non-long-haul trucking jobs, it is also quite possible that workers might expect to have different productivity at different long-haul trucking firms. For example, one driver may have a lot of experience driving around the South and would be less productive at a firm where most of the routes were in the Northeast. Based on where the routes are, it may be harder for a driver to get home regularly, and spending a lot of time getting home could negatively affect productivity.

of Figure E5 for details). If workers were equally overconfident about their inside and outside options, we would expect that workers who were more confident about their inside options to also be more confident about their outside options, but this is not the case.

3. Table 3 showed that, all else equal, workers with higher productivity beliefs are less likely to quit while controlling for actual productivity to date. This finding is supportive of differential overconfidence.[27]

4. While current earnings are proportional to ability in long-haul trucking, most US jobs do not pay piece rates. Thus, even if workers are overconfident about their productivity at their outside option, it is unclear how much this affects workers' perceptions of their earnings at their outside option.[28]

5. The assumption of differential overconfidence is consistent with evidence in psychology and behavioral economics (see Appendix A.8 for details).[29]

Micro-foundations of Overconfidence. Several theoretical micro-foundations for overconfidence have been proposed in the literature, including evolutionary advantages, self-signaling, and social-signaling. We remain agnostic about the source of the overconfidence (since our model and estimation do not depend on knowing the source). Our contribution is to document overconfidence and explore the implications of overconfidence for behavior and welfare, not to understand its foundation. We do, however, assume that agents do not receive psychological utility from their beliefs, consistent with our Section 3 experimental finding that incentives do not appear to reduce overconfidence. If agents received psychological utility from beliefs, we might expect them to trade off incentives for accuracy with the utility value from stating high beliefs (unless, of course, the personal benefits

[27] In Prop. 1 in Appendix C, we prove that more overconfident workers will be less likely to quit if and only if they are more overconfident about the inside than the outside option. This finding would seem unlikely if workers had the same beliefs about their inside and outside options. Moreover, if having high productivity beliefs was indicative of drivers who think the "grass is always greener" in other jobs, then high beliefs would be correlated with more quitting, not less.

[28] Performance pay is used in only 37% of U.S. jobs, comprises a median of 4% of total pay across jobs, and is less common in blue-collar jobs like trucking than white-collar jobs (Lemieux et al., 2009). Of course, other pecuniary aspects of a job (e.g., the perceived probability of being promoted to a higher wage) may be affected by overconfidence.

[29] The assumption of differential overconfidence may be more important for the interpretation of the counterfactuals than the structural estimation. In the structural estimation, overconfidence about the inside option varies over time due to learning. Thus, the model-specified overconfidence about the inside option would not be exactly offset by overconfidence about the outside option unless it varied over time in the same way (this is one important way in which the dynamic structural model differs from the one-period model in Appendix C). Appendix A.8 discusses further.

from stating optimistic beliefs are so strong that agents are unwilling to trade-off moderate-sized incentives to reduce their optimism). (To the extent that the assumption is incorrect and truckers do receive substantial psychological benefits from their beliefs, this would cause us to overstate the benefit to workers of eliminating overconfidence.) Also, we remain agnostic whether workers are overconfident about their own skills (e.g., "I have great endurance on the road") vs. whether they are overoptimistic about external events (e.g., "traffic will be better next week").

Learning about productivity. We model quitting as a product of worker learning, and provide reduced-form evidence for this assumption here. One implication of learning about productivity is that at a given level of tenure, workers who have higher productivity to date should be less likely to quit. Figure E1 supports this by comparing the average miles per week of drivers who quit that week versus drivers who make it to that week and don't quit. Selection on productivity is examined with controls in Table 3, where higher average productivity to date is associated with significantly less quitting.

4.3 Estimation and Identification

The model is estimated by maximum likelihood. In Appendix D, we derive the likelihood function and describe the estimation procedure. Although the parameters are jointly identified, we can discuss the main data features that allow us to identify particular model parameters.

Productivity and skill parameters. The productivity parameters σ_0, σ_y, and η_0 are identified primarily by the productivity data. σ_0 reflects the degree of permanent productivity differences *across* individuals. σ_y reflects differences *within* individuals in productivity. η_0 reflects the mean average ability of workers in the population. The learning by doing parameters, a_1 and a_2, are identified by how much productivity goes up (a_1) and how quickly (a_2). The skill gain parameter, s_0, is identified based on turnover levels during the first 5 weeks when workers are driving with an experienced driver. The continuous skill gain parameters, θ_1 and θ_2 are identified by how much quitting changes with tenure given the increase in measured productivity.

Taste heterogeneity. The taste for job parameters are identified from persistent differences between individual quitting behavior and the predictions of the model. Suppose that the data contained many low-productivity workers who nevertheless kept choosing not to quit. This would cause the model to estimate that there is a large amount of unobserved taste heterogeneity.

Belief parameters. The subjective beliefs data are critical for identifying the belief parameters. Prior mean bias (overconfidence), η_b, is identified by the difference between believed and actual productivity, particularly at lower tenure levels. The believed standard deviation of productivity shocks, $\widetilde{\sigma_y}$, determines the subjective speed of learning in the model. The larger $\widetilde{\sigma_y}$ is, the slower that agents' initial overprediction will disappear. The standard deviation of beliefs, σ_b, is identified by noise in beliefs unrelated to information in model-predicted subjective expectations. An increase in σ_b leads to greater week-to-week fluctuations in beliefs unrelated to actual productivity.[30]

Scale parameter. The scale parameter of the idiosyncratic shock, τ, is identified based off of how much quitting behavior in the data differs from that predicted by a model with individual unobserved heterogeneity, but not time-varying uncertainty.

4.4 Implementation

The outside option, r, is taken to be the median full-time earnings from the 2006 March CPS of workers like the data subset "median" driver (35-year old males with a high school degree), which is \$32,000 per year.[31] We convert this to a weekly wage of \$640. The weekly discount factor, δ, is set to $\delta = 0.9957$, corresponding to an annual discount factor of 0.8.[32] In our baseline estimates, we do not include demographic covariates or heterogeneity in overconfidence.[33] Learning is assumed complete after $T = 130$ periods. We use data on up to 110 weeks per driver. After presenting our baseline estimates, we discuss robustness to alternative assumptions. In our dataset, we focus on 699 workers with complete data (see Appendix D).

[30] Possible heterogeneity in η_b (discussed in Appendix A.9) is identified from differences across people in the extent of productivity overprediction. We also note that η_b and $\widetilde{\sigma_y}$, in addition to affecting subjective beliefs, will also affect quitting. For example, the faster that agents begin to rely on their average productivity to date in making quit decisions (that is, the faster their quitting decisions reflect "learning"), the smaller that $\widetilde{\sigma_y}$ will be.

[31] "High school graduate" is the modal educational category (40% of drivers) whereas "Some college" is the median category. Table F1 shows our estimates are very similar if we assume a higher outside option.

[32] It is standard in most dynamic structural models to assume rather than estimate the discount factor, as the discount factor is usually weakly identified. We have experimented with a broad range of discount factors in sensitivity analysis. Model fit appears best for annual discount factors in the range of 0.80. However, assuming an annual discount factor of 0.90 or 0.95 yields quite similar estimates (see Table F1). A discount factor of 0 yields a substantially worse fit, evidence that workers in our context are forward looking. An annual discount factor of 0.80 is "low," but is comparable or higher than discount factors used or estimated in other models analyzing dynamic choices of blue-collar or low-income workers (e.g., Paserman, 2008; Fang and Silverman, 2009; Warner and Pleeter, 2001).

[33] We have also estimated models with covariates, for example, allowing taste for the job, α, to depend on gender, education, race, and age. However, including covariates has little effect on model fit and on the estimates of the other parameters. See also Card and Hyslop (2005) for an example of a dynamic model where covariates are excluded because they do not significantly improve model fit.

4.5 Structural Results

Table 4 displays the main structural estimates and indicates substantial mean bias and variance bias. As a benchmark, column 1 provides estimates assuming no mean bias. Column 2 allows for mean bias, estimating bias, η_b, of 589 miles (or roughly 30% of 2,025 miles, the estimated mean of the true productivity distribution). The productivity parameters in column 2 also differ from those in column 1 and seem more reasonable in size.[34] In terms of variance bias, the believed standard deviation of productivity shocks is roughly 2.5 times higher than the actual standard deviation of productivity shocks. This implies that workers update beliefs considerably slower than predicted by Bayes' Rule. Recall that the weight agents place on their signals relative to their prior is $\frac{t\sigma_0^2}{t\sigma_0^2+\widetilde{\sigma_y}^2}$. After 20 weeks, the worker is estimated to place weight 0.31 on his signals (whereas if $\widetilde{\sigma_y} = \sigma_y$, the worker would place weight 0.77 on his signals).[35] Table 4 also indicates significant heterogeneity in non-pecuniary taste for the job.

Table 5 takes the baseline model and adds learning by doing and continuous skill accumulation. The fit is better in both specifications than in Table 4. Many of the parameters are qualitatively similar to before, but there are some differences. In column two, the mean prior bias is larger than before, estimated at 707 miles. The estimated taste heterogeneity is also somewhat different.

Figure 2 shows the model fits the data quite well. We simulate 40,000 drivers. The quit hazard-tenure, productivity-tenure, and beliefs-tenure profiles observed in the data are plotted using an Epanechnikov kernel. As in the data, the model-predicted quit hazard is initially increasing, reflecting learning about productivity. When workers are uncertain about their productivity, they face an incentive to wait and see how productive they will be before quitting. Also, the model predicts a large spike in quitting after 52 weeks (when drivers come off the 12-month contract).

The models with mean bias (column 2 in Tables 4 and 5) fit the data much better in terms of overall fit than the models without mean bias (column 1), according to a likelihood ratio test ($p < 0.01$). To analyze model fit on the quitting data alone, we compare the observed weekly number of drivers quitting (O_t) with the number predicted from the model (E_t) using a chi-squared

[34] For example, the mean of the prior productivity distribution is 2,025 miles per week, down roughly 20% from 2,468 miles per week in column 1. In column 1, the prior productivity distribution needs to explain the earnings data. But it also needs to explain the quitting and subjective beliefs data, which "pulls" the estimate substantially upward.

[35] This finding is consistent with the well-known psychological phenomenon of "conservatism" (Edwards, 1968), where agents update less than a rational agent would after receiving new information. For recent evidence on conservatism, see Eil and Rao (2011) and Mobius et al. (2013). For theory on conservative updating, see Schwartzstein (2014).

test. The chi-squared statistic is $\sum_t \frac{(E_t - O_t)^2}{E_t}$. Inference is conducted using T degrees of freedom, where T is the maximum number of weeks a driver is observed. For column 1 of Table 4, the chi-squared statistic is 520.2, whereas for column 2, the chi-squared statistic is 214.0 ($p < 0.01$).[36] Thus, the fit in terms of quitting is considerably better in the model with belief bias than without it.

Robustness. Table F1 shows our main estimates are quite robust to different assumptions. Increasing the discount factor does not significantly change the estimates, nor does using inverse probability weighting to address non-response. Winsorizing subjective beliefs at $4,000$ miles decreases the mean bias term by 5% to 560 miles, suggesting that very high beliefs are accounting for only a small share of the overall mean bias. Allowing learning to occur over 200 weeks (instead of 130) and increasing the outside option also do not much change the estimates. Results are also robust to allowing for heterogeneity in overconfidence. Appendix A.9 discusses this and additional robustness checks.

One can imagine alternative non-standard economic forces that affect a worker's taste for the job, e.g., feelings of commitment toward the firm providing training. However, time-invariant shifters of job taste are already accounted for via the taste heterogeneity parameters. In contrast, overconfidence provides a time-varying impact on the value of staying that fits the data well.

Out-of-Sample Fit. All drivers in our sample have the 12-month training contract described in the main text. A companion paper (Hoffman and Burks, 2016) studies worker behavior under three different contractual regime (no contract, 12-month contract, and 18-month pro-rated contract), showing that the model developed in the present paper can predict basic retention patterns under the no contract and 18-month contract regimes. Thus, our structural model also makes reasonable out-of-sample predictions, which is a more demanding test for the model.

[36] Chi-squared tests are often used to assess the fit of dynamic models (e.g., Keane and Wolpin, 1997; Card and Hyslop, 2005). Comparing the chi-squared statistic in column 1 of Table 4 with that in column 2 after one parameter is added, the difference in chi-squared is highly significant ($\chi^2_{df=1} = 520.2 - 214.0 \rightarrow p < 0.01$). (As caveated though in Card and Hyslop (2005), it is more correct to think of the calculated chi-squared statistic as an informal measure of fit, since the predicted numbers are created from the same data being used for the observed cell entries.) The model with mean bias also fits other aspects of the data besides the means in Figure 2, such as the distribution of beliefs and productivity over tenure (see Figure F1). To see the poor model fit for the model without mean bias, see Figure F2.

5 Counterfactual Simulation: Debiasing

Set-up for Counterfactuals. We use our baseline structural estimates (column 2 of Table 4) to quantify the importance of biased beliefs. Profits are defined as production profits, plus training contract penalties, minus training costs. For a worker who stays T periods before quitting, profits are:

$$\pi = \sum_{t=1}^{T} \delta^{t-1}((P - mc - w_t)y_t - FC) + \sum_{t=1}^{T} \delta^{t-1}\theta k_t q_t - TC \tag{6}$$

where P is the price the firm charges for one mile of shipment, mc is the non-wage marginal cost per mile (such as truck wear and fuel costs), y_t is a driver's productivity, FC is fixed costs per week (such as back office support for the driver), q_t is a dummy for quitting in week t, θ is the share of the training contract penalty collected by the firm, and TC is training cost per worker. Based on consultation with Firm A managers, we assume that $P - mc = \$0.70/\text{mile}$, $\theta = 0.3$, $FC = \$650/\text{week}$, and $TC = \$2,500$ for the new inexperienced workers we study. We equate the firm's weekly discount factor to the worker's, $\delta = 0.9957$, so as to avoid having results being driven by differences in discount factors; our conclusions are unchanged if we assume higher discount factors for both worker and firm. Further details on computing profits are given in the Appendix A.10.1 and Appendix D.

Although workers have biased beliefs, they have standard preferences. Worker welfare is measured simply by summing earnings, taste for trucking, and idiosyncratic shocks, as in equation (3).[37]

For the counterfactuals, we simulate the full data-generating process for 3,000 simulated workers for up to 1,300 weeks each. While workers are simulated for up to 1,300 weeks, we focus on showing profits per worker and welfare per worker numbers after 110 weeks (corresponding to the maximum number of weeks under observation in the data).[38] Throughout the counterfactuals, we focus separately on profits and worker welfare, and do not analyze total welfare.[39]

Debiasing: Reducing Worker Overconfidence. To examine quantitatively how overcon-

[37] Since workers have biased beliefs, average *experienced utility* will differ from a worker's ex ante *expected utility*. We measure welfare using experienced utility; this focus is shared by the empirical work of Grubb and Osborne (2015) and the theoretical framework in Mullainathan et al. (2012).

[38] The counterfactuals yield the same qualitative conclusions if we analyze profit and worker welfare after 1,300 weeks.

[39] While we found our conclusions on profits and worker welfare to be very robust to different assumptions, we found total welfare to depend more closely on particular assumptions made. That our counterfactuals do not lead to unambiguous changes in total welfare is not surprising, given our model allows for multiple market failures, including private information about taste for the job and shocks; biased beliefs; quitting externalities; and monopsony power in the training market.

fidence affects quitting, worker welfare, and profits, we simulate eliminating worker overconfidence, which we also refer to, following the psychology literature, as "debiasing." We also consider eliminating overconfidence by one-half, recognizing that debiasing may be incomplete in practice. This practice of analyzing the impacts of counterfactually eliminating a "behavioral" parameter also appears in work such as Handel (2013).

As seen in Table 6, full debiasing increases worker welfare by about 5% since worker quitting decisions become less distorted by overconfidence. Although the workers exhibit significant turnover in the baseline, turnover becomes even higher when workers are debiased. In the un-debaised simulation, worker retention is 43% after 60 weeks, but this falls to 33% under 50% debiasing, and to 24% under 100% debiasing. Without debiasing, workers tend to interpret low-mileage weeks as repeated instances of "bad luck" (i.e., weeks that are low relative to their prior beliefs). After debiasing, workers' quitting decisions are no longer distorted by having a rosy outlook.

In addition, firm profits substantially decline under debiasing. Under full debiasing, profits per worker decline by over $3,000. Due to the increase in quitting, the firm has less time to make profits from a given worker. Our counterfactual allows us to quantify how much overconfidence affects training profits and worker welfare in this market.[40]

6 Conclusion

Using a sample of truckers, we find that workers tend to overpredict their productivity, thereby providing robust field evidence on worker overconfidence. Overprediction declines slowly and persists (albeit in reduced form) even after two years. Higher productivity beliefs are correlated with less quitting while controlling for actual productivity to date. To quantify the importance biased beliefs for profits and welfare, we make stronger assumptions, structurally estimating a quitting model with potentially biased beliefs. We show that overconfidence is critical for the profitability of training. If worker overconfidence was eliminated, profits from training would fall substantially. Further, overconfidence moderately reduces worker welfare by distorting worker decisions.

Our results parallel several studies in behavioral industrial organization indicating that firms

[40]This counterfactual takes training contracts as fixed. It is conceivable that firms would optimally adjust contracts (wages and training contract penalties) if workers did not exhibit overconfidence. We explored this extended debiasing counterfactual in an earlier version of the paper. Under this counterfactual of debiasing with optimal contractual responses, the overall conclusion from the baseline debiasing counterfactual remain quite similar. The main difference was that debiasing tended to push firms toward decreasing optimal quit penalties.

may profitably exploit consumers' biases, focusing instead on the behavioral biases of workers. An important difference in our setting, besides the identity of the parties involved, is the possibility of a baseline market failure of underinvestment. In a second-best world with underinvestment in general training, the existence of worker overconfidence makes training and training contracts more profitable, potentially increasing the quantity of firm training. Paralleling the work in industrial organization, we find evidence that agents (in our case, workers) are harmed by a behavioral bias, but our results also raise the possibility that it may not necessarily be in workers' interests to have that bias eliminated. Unfortunately, because our data are primarily from one firm, we are unable to calculate how overconfidence affects the share of firms willing to train; thus, we are unable to weigh this benefit against the distortion from overconfidence on worker decision-making. Additional research is clearly called for.

While truckers are well-suited for examining overconfidence about productivity, it is important to highlight that we are focusing on one particular job, and the patterns we document may not necessarily hold in other settings. Future worker should examine whether worker overconfidence occurs in other settings. While piece rate compensation is not shared by most other jobs, piece rate compensation is not necessary for overconfidence to make workers less likely to quit after training (or more likely to agree to training in the first place). For example, workers may be over-optimistic about some other aspect of the job, e.g., the probability of being promoted. While we focus on firm training, overconfidence can help entice or "lock in" workers even if there is no training involved, and may be relevant for other types of human capital investment, e.g., going to college, or for occupational choice in general.[41]

Worker overconfidence may be important for many aspects of optimal job design and compensation. For example, when firms can choose to pay flat wages or piece rates, paying a piece rate may be appealing if workers are overconfident since overconfident workers perceive they may earn more than they actually will (Larkin and Leider, 2012). Future work should continue to analyze the importance of worker biases for employee behavior, worker welfare, and firm outcomes.

[41] See Stinebrickner and Stinebrickner (2012) for evidence that college students are initially overconfident about their likely performance in college. There has been recent discussion, particularly related to law schools, that students may be overly optimistic about their future job prospects when taking on student loans, e.g., David Segal, "Law Students Lose the Grant Game as Schools Win," New York Times, April 2011 and Liz Goodwin, "Law grads sue school, say degree is 'indentured servitude'," Yahoo News, August 2011. In a different application for workers, the impact of overconfidence about job-finding on optimal unemployment insurance, see Spinnewijn (2015). For work on overoptimism and stock options for nonexecutive workers, see, e.g., Oyer and Schaefer (2005).

References

Acemoglu, Daron and Jorn-Steffen Pischke, "The Structure of Wages and Investment in General Training," *Journal of Political Economy*, 1999b, *107* (3), 539–572.

Arcidiacono, Peter, "Ability Sorting and the Returns to College Major," *Journal of Econometrics*, 2004, *121* (1-2), 343–375.

Baker, George P. and Thomas N. Hubbard, "Contractibility And Asset Ownership: On-Board Computers and Governance In U. S. Trucking," *Quarterly Journal of Economics*, 2004, *119* (4), 1443–1479.

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.

Belzer, Michael, *Sweatshops on Wheels: Winners and Losers in Trucking Deregulation*, Oxford University Press, 2000.

BLS, "Bureau of Labor Statistics Occupational Outlook Handbook, 2010-11: Truck Drivers and Driver/Sales Workers," 2010.

Bojilov, Raicho, "Incentives to Work or Incentives to Quit?," 2013. Mimeo, Ecole Polytechnique.

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, "Overconfidence and Social Signalling," *Review of Economic Studies*, 2013, *80* (3), 949–983.

_ , _ , _ , 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.

Card, David and Dean R. Hyslop, "Estimating the Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers," *Econometrica*, 2005, *73* (6), 1723–1770.

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

Ching, Andrew T., Tulin Erdem, and Michael P. Keane, "Learning Models: An Assessment of Progress, Challenges, and New Developments," *Marketing Science*, 2013, *32* (6), 913–938.

Coscelli, Andrea and Matthew Shum, "An Empirical Model of Learning and Patient Spillovers in New Drug Entry," *Journal of Econometrics*, 2004, *122* (2), 213 – 246.

Crawford, Gregory S. and Matthew Shum, "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica*, 2005, *73* (4), 1137–1173.

DeGroot, Morris, *Optimal Statistical Decisions*, New York, NY: Mcgraw-Hill College, 1970.

Edwards, Ward, "Conservatism in Human Information Processing," in "*Formal Representation of Human Judgement*" 1968.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

Fang, Hanming and Dan Silverman, "Time-Inconsistency And Welfare Program Participation: Evidence From The NLSY," *International Economic Review*, 2009, *50* (4), 1043–1077.

Goettler, Ronald and Karen Clay, "Tariff Choice with Consumer Learning and Switching Costs," *Journal of Marketing Research*, 2011, *48* (4), 633–652.

Grubb, Michael D. and Matthew Osborne, "Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock," *American Economic Review*, 2015, *105* (1), 234–71.

Handel, Benjamin R., "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts," *American Economic Review*, December 2013, *103* (7), 2643–82.

Heckman, James J. and Burton Singer, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica*, 1984, *52* (2), 271–320.

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

_ **and Stephen Burks**, "Training Contracts, Employee Turnover, and the Returns from Firm-sponsored General Training," 2016. Mimeo, University of Toronto.

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

Huffman, David and Julia Shvets, "Overconfidence and Memory: Evidence from Managers," 2016. Presentation, Stanford Institute on Theoretical Economics.

Jovanovic, Boyan, "Job Matching and the Theory of Turnover," *Journal of Political Economy*, 1979, *87* (5), 972–90.

_ **and Yaw Nyarko**, "Learning by Doing and the Choice of Technology," *Econometrica*, 1996, *64* (6), 1299–1310.

Keane, Michael P. and Kenneth I. Wolpin, "The Career Decisions of Young Men," *JPE*, 1997, *105* (3), 473–522.

Koszegi, Botond, "Behavioral Contract Theory," *Journal of Economic Literature*, 2014, *52* (4), 1075–1118.

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

Lemieux, Thomas, W. Bentley MacLeod, and Daniel Parent, "Performance Pay and Wage Inequality," *Quarterly Journal of Economics*, 2009, *124* (1), 1–49.

Malmendier, Ulrike and Geoffrey Tate, "CEO Overconfidence and Corporate Investment," *Journal of Finance*, December 2005, *60* (6), 2661–2700.

Mobius, Markus M., Muriel Niederle, Paul Niehaus, and Tanya Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," 2013. Mimeo.

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

Mullainathan, Sendhil, Joshua Schwarzstein, and William J. Congdon, "A Reduced-Form Approach to Behavioral Public Finance," *Annual Review of Economics*, 2012, *4*, 511–540.

Nagypal, Eva, "Learning by Doing vs. Learning About Match Quality: Can We Tell Them Apart," *Review of Economic Studies*, 2007, *74* (1), pp. 537–566.

Oyer, Paul and Scott Schaefer, "Why Do Some Firms Give Stock Options to All Employees?: An Empirical Examination of Alternative Theories," *Journal of Financial Economics*, 2005, *76* (1), 99–133.

Paserman, M.Daniele, "Job Search and Hyperbolic Discounting: Structural Estimation and Policy Evaluation," *Economic Journal*, 2008, *118* (531), 1418–1452.

Sanders, Carl, "Skill Uncertainty, Skill Accumulation, and Occupational Choice," 2014. Mimeo, WashU in St. Louis.

Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, 2014, *12* (6), 1423–1452.

Shaw, Kathryn and Edward P. Lazear, "Tenure and Output," *Labour Economics*, 2008, *15* (4), 704–723.

Spinnewijn, Johannes, "Unemployed but Optimistic: Optimal Insurance Design with Biased Beliefs," *Journal of the European Economic Association*, 2015, *13* (1), 130–167.

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

Stinebrickner, Todd and Ralph Stinebrickner, "Learning about Academic Ability and the College Dropout Decision," *Journal of Labor Economics*, 2012, *30* (4), 707 – 748.

Van den Steen, Eric, "Rational Overoptimism (and Other Biases)," *American Economic Review*, 2004, *94* (4), 1141–1151.

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.

Warner, John T. and Saul Pleeter, "The Personal Discount Rate: Evidence from Military Downsizing Programs," *American Economic Review*, 2001, *91* (1), 33–53.

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

Table 1: Summary Statistics

Panel A: Driver Characteristics

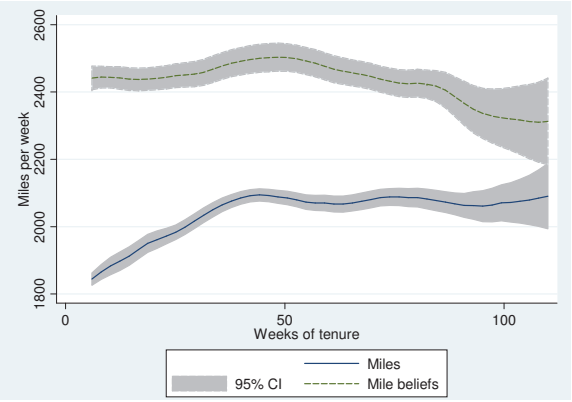
Variable	**Mean**
Female	0.10
Black	0.11
Hispanic	0.02
Age	36
Married	0.41
Number of Kids	0.96
Years of Schooling	13
Credit Score	586
No Credit Score	0.12
Number of workers	895

Panel B: Productivity and Productivity Beliefs

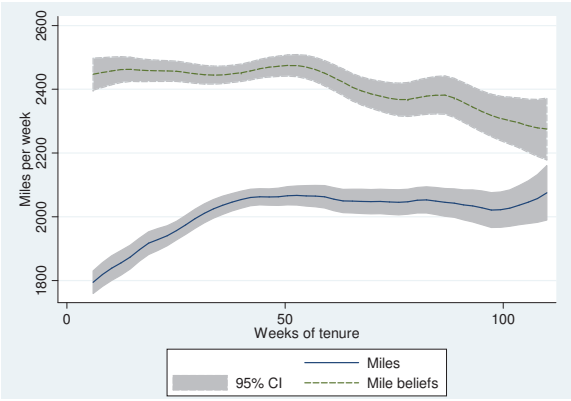
	Data Subset		Estimation Sample	
Percentile	**Miles**	**Miles Beliefs**	**Miles**	**Miles Beliefs**
10%	897	1500	1019	1500
25%	1367	1800	1475	2000
50%	1883	2300	1972	2500
75%	2427	2750	2506	2800
90%	2942	3000	3005	3050
Mean	1908	2323	1998	2423

Notes: Panel A provides summary statistics. Data subset drivers are from the same training school and were hired in late 2005 or 2006. Panel B presents quantiles and means on productivity and productivity beliefs, both for the data subset (895 drivers) and for the 699 drivers with complete data that we use in the structural estimation (see Appendix D). Summary statistics on miles are calculated restricting to weeks where miles is greater than zero. Though mean miles and beliefs are higher for the estimation sample, the difference between the two is similar in both samples. In both samples, median mile beliefs lie at roughly the 70th percentile of miles. See Appendix A.3 for more details on data and sample construction.

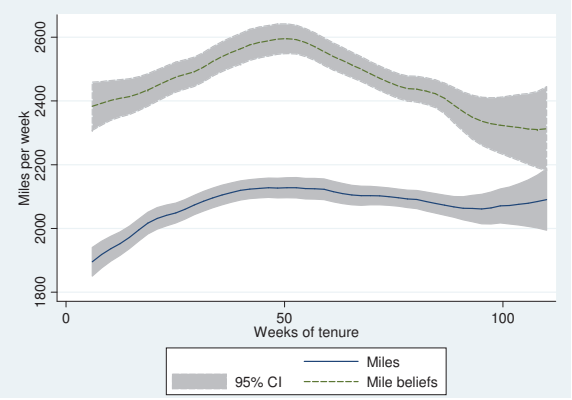
Figure 1: Overconfidence: Comparing Subjective Productivity Forecasts with Actual Worker Productivity (as a Function of Worker Tenure)



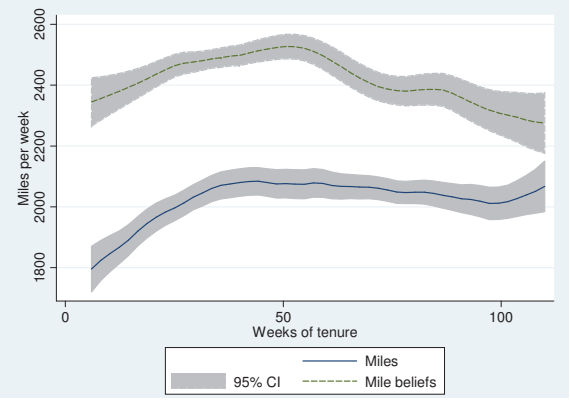
(a) Means



(b) Medians



(c) Means, Restrict to Workers Who Stay at Least 75 Weeks

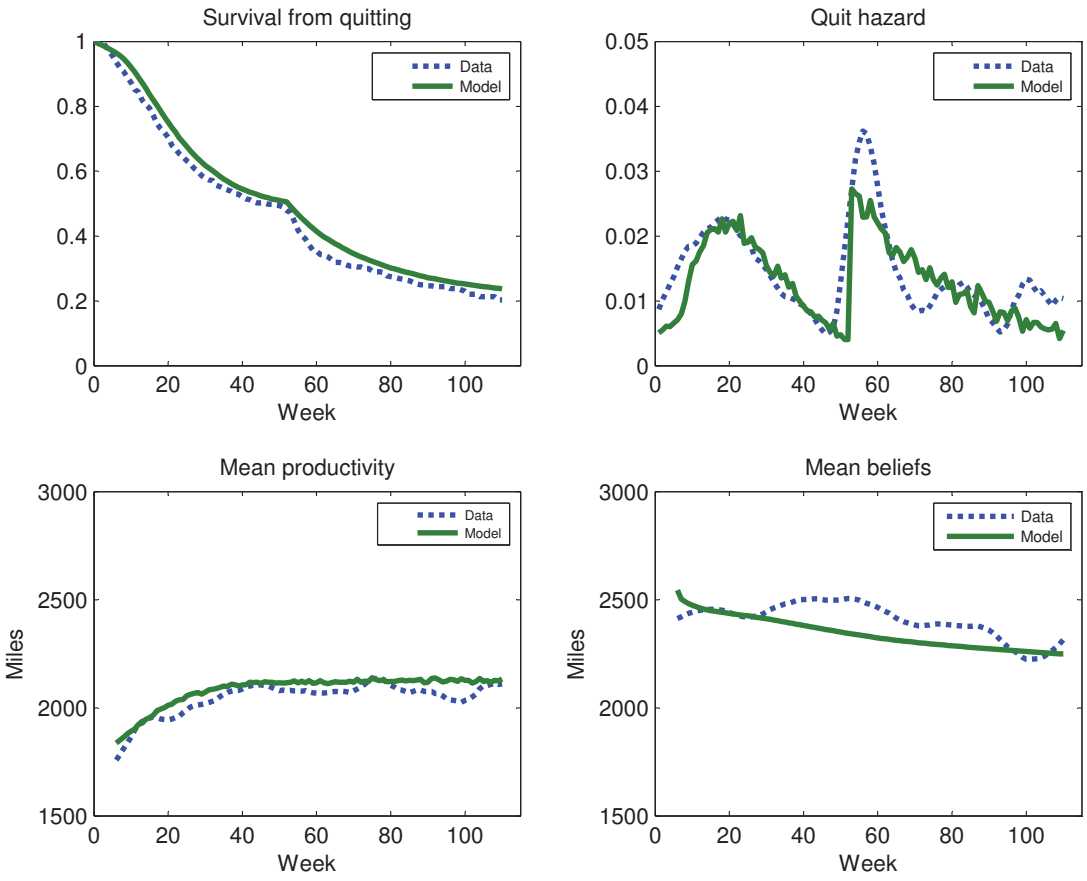


(d) Medians, Restrict to Workers Who Stay at Least 75 Weeks & Respond to Belief Survey

Notes: These figures analyze actual and believed productivity for the drivers in the data subset. The figures are plotted using a local polynomial regression with an Epanechnikov kernel. A bandwidth of 5 weeks is used for the productivity data and a bandwidth of 7 weeks is used for the belief data. In panels (a) and (c), the productivity and belief data are collapsed into weekly means before local polynomial smoothing. In panels (b) and (d), the productivity and belief data are collapsed into weekly medians before local polynomial smoothing. In panel (d), we restrict to the 62% of workers who ever respond to the survey. The figure is similar if we restrict miles to weeks where the driver responds to the survey, though standard errors are larger.

It may seem surprising that the initial amount of overconfidence is large and is slow to diminish. However, as noted in Section 3.1, this figure averages across workers. At an individual level, workers only over-predict their productivity in 65% of weeks, so they are not over-predicting every week (there is substantial idiosyncratic variation in miles in the data). Further, the slow speed of learning can be rationalized in the structural model given the substantial idiosyncratic variation in miles in the data, combined with a significant dose of conservatism / variance bias. See Figure E3 for overprediction as a function of driver tenure.

Figure 2: Structural Model: Model Fit



Notes: These figures compare model-simulated data against the actual data to assess model fit. We plot the survival curve, the quit hazard, the mean miles-tenure profile, and the mean beliefs-tenure profile. The model simulated corresponds to Column 2 in Table 5. We simulate the entire data-generating process for 40,000 drivers. The data are from 699 drivers with the 12-month contract, all of whom are from the same training school and hired in late 2005 or 2006. The paths of quits, earnings, and beliefs over tenure are plotted using an Epanechnikov kernel, with bandwidths of 6 weeks, 5 weeks, and 10 weeks, respectively. For model fit on the distribution of productivity across drivers, see Figure F1.

Table 2: Do Productivity Beliefs Predict Productivity? OLS Regressions

	(1)	(2)	(3)	(4)	(5)
L. Predicted miles	0.194***	0.147***	0.071***	0.068***	0.080***
	(0.023)	(0.019)	(0.016)	(0.015)	(0.022)
L. Avg miles to date			0.643***	0.623***	
			(0.037)	(0.036)	
Demographic Controls	No	Yes	No	Yes	No
Work Type Controls	No	Yes	No	Yes	No
Subject FE	No	No	No	No	Yes
Observations	8,449	8,449	8,445	8,445	8,449
R-squared	0.070	0.129	0.187	0.191	0.294

Notes: The dependent variable is miles driven per week. An observation is a driver-week. Standard errors clustered by driver in parentheses. Demographic controls are controls for gender, race, marital status, age, and education. Productivity is given in terms of hundreds of miles driven per week. All regressions include week of tenure dummies. All drivers are from the same training school and were hired in late 2005 or 2006. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 3: Do Productivity Beliefs Predict Quitting?

	(1)	(2)	(3)	(4)	(5)
Mile beliefs (in hundreds of miles)	-0.059***			-0.057***	-0.065***
	(0.018)			(0.020)	(0.021)
Avg miles to date		-0.081***	-0.118***	-0.011	-0.072*
		(0.013)	(0.038)	(0.034)	(0.041)
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: An observation is a driver-week. The regressions are Cox proportional hazard models, where the dependent variable is quitting. Events where the driver is fired are treated as censored. Standard errors clustered by worker in parentheses. Demographic controls are controls for gender, race, marital status, age, and education. Productivity is given in terms of hundreds of miles driven per week. Column 3 differs from column 2 in that it restricts to the sample of driver-weeks for which there is a corresponding belief expectation. All drivers are from the same training school and were hired in late 2005 or 2006. * significant at 10%; ** significant at 5%; *** significant at 1%

Table 4: Baseline Structural Estimates

		No Bias (1)	Belief Bias (2)
	Productivity and Skill Parameters		
η_0	Mean of prior productivity dist	2464	2025
		(18)	(17)
σ_0	Std dev of prior productivity dist	475	286
		(18)	(11)
σ_y	Std dev of productivity shocks	707	706
		(3.6)	(3.6)
s_0	Skill accumulation in first 5 weeks	14.9	8.6
		(3.1)	(4.4)
	Taste UH Parameters		
μ_1	Mass point 1 of taste UH	-249	-260
		(8)	(14)
μ_2	Mass point 2 of taste UH	-107	-135
		(20)	(13)
μ_3	Mass point 3 of taste UH	139	135
		(43)	(41)
p_1	Probability type 1	0.55	0.34
		(0.04)	(0.07)
p_1	Probability type 2	0.24	0.43
		(0.03)	(0.06)
	Belief Parameters		
η_b	**Belief bias**		**589**
			(28)
$\widetilde{\sigma_y}$	**Believed std dev of productivity shocks**	**3653**	**1888**
		(158)	**(136)**
σ_b	Std dev in beliefs	300	298
		(1.4)	(1.4)
	Scale Parameter		
τ	Scale param of idiosyncratic shock	1620	2208
		(178)	(350)
	Log-likelihood	-91064	-90865
	Number of workers	699	699

Notes: This table presents estimates of the structural parameters. The idiosyncratic shock, skill gain, and taste parameters are given in terms of dollars whereas the productivity and belief parameters are given in terms of miles. "Taste UH" stands for unobserved heterogeneity in taste for the job. Standard errors are in parentheses and are calculated by inverting the Hessian. A weekly discount factor of 0.9957 is assumed for workers and firms, corresponding to an annual discount factor of 0.8. The data are from 699 drivers in the data subset, all of whom face the 12-month training contract.

Table 5: Structural Estimates with Learning by Doing and Skill Accumulation

		No Bias (1)	Belief Bias (2)
	Productivity and Skill Parameters		
η_0	Mean of prior productivity dist	2401	1838
		(12)	(23)
σ_0	Std dev of prior productivity dist	508	274
		(19)	(10)
σ_y	Std dev of productivity shocks	706	706
		(1.5)	(1.6)
a_1	Learning by doing level	154	242
		(12)	(15)
a_2	Learning by doing speed	0.08	0.11
		(0.01)	(0.01)
θ_1	Skill accumulation level	56	118
		(39)	(55)
θ_2	Skill accumulation speed	0.92	0.03
		(0.33)	(0.01)
	Taste UH Parameters		
μ_1	Mass point 1 of taste UH	-219	-420
		(42)	(79)
μ_2	Mass point 2 of taste UH	-77	-162
		(44)	(56)
μ_3	Mass point 3 of taste UH	167	176
		(57)	(62)
p_1	Probability type 1	0.54	0.60
		(0.07)	(0.04)
p_1	Probability type 2	0.25	0.19
		(0.05)	(0.04)
	Belief Parameters		
η_b	**Belief bias**		**707**
			(24)
$\widetilde{\sigma_y}$	**Believed std dev of productivity shocks**	**3420**	**1329**
		(138)	**(63)**
σ_b	Std dev in beliefs	299	297
		(0.3)	(0.3)
	Scale Parameter		
τ	Scale param of idiosyncratic shock	1733	1808
		(166)	(201)
	Log-likelihood	-91020	-90767
	Number of workers	699	699

Notes: See the notes for Table 4. For the newly added parameters here, a_1 and a_2 are in miles, whereas θ_1 and θ_2 are in dollars.

Table 6: Counterfactual Simulations

Counterfactual:	Baseline	50% debias	100% debias
Profits per worker	\$4,907	\$3,247	\$1,500
Welfare per worker	\$54,840	\$56,496	\$57,612
Retention at 20wks	0.79	0.61	0.39
Retention at 40wks	0.56	0.41	0.27
Retention at 60wks	0.43	0.33	0.24

Notes: This table reports the results of the counterfactual simulations described in the text, while assuming that training contracts are not adjusted in response. Under the 50% debias and 100% debias counterfactuals, worker overconfidence is reduced by 50% or 100% (by reducing η_b by 50% or 100%). Profits and welfare are defined in Section 5 of the text. The model simulated here corresponds to column 2 of Table 4.

FOR ONLINE PUBLICATION ONLY

Web Appendix

Appendix A provides additional discussion and analysis on belief survey non-response, incentives vs. selection, and other issues. Appendix B provides more information on the Firm B field experiment. Appendix C analyzes a one-period version of the structural model and derives analytical results. Appendix D provides omitted derivations from the structural model. Appendices E-F collect additional figures and tables. Appendix G discusses two miscellaneous issues (measuring productivity and the prevalence of training contracts in trucking).

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 results.

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. The survey response rate is fairly constant over driver tenure, despite workers leaving the sample by quitting.[2] Among drivers who respond to at least one survey, there is significant variation in the average response rate, as seen in Figure E2.

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

[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.

[2] In a regression of the response rate on driver tenure (or driver tenure and drive tenure squared), there is no relationship between tenure and response rate.

message system. (A Qualcomm is a very basic computer in trucks used for sending and receiving messages.)

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			0.001			
(in hundreds of miles per week)			(0.002)			
L. Average overconfidence to date				-0.004		
(in hundreds of miles per week)				(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	2,172	2,172

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 (predicted miles minus actual miles) on characteristics. 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. Drivers are from the same training school and were hired in late 2005-2006. * significant at 10%; ** significant at 5%; *** significant at 1%

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 biasing our results, we turn now to quantitative tests. To address non-random response, we use Inverse Probability 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 observables (race, gender, years of schooling, and age at start of work). 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 the standard demographics that we have controlled for, 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, have a disagreeable personality, or are generally impatient; our data allow us to convert what are usually unobserved characteristics into observed ones. Our results are highly robust to Inverse Probability Weighting adding these richer characteristics.[3]

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 relationship between average productivity beliefs to date and survey response (column 3 of Table A1), or between average overprediction and survey response (column 4 of Table A1). These are precisely estimated zero coefficients.[4]

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.[5] Column 5 shows that drivers who 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, $\rho = -0.06$,

[3]We have estimated with different additional first-stage variables, including IQ, numeracy, and experimentally-measured patience, estimating models as in column 3 of Table F1. Our results are robust to their inclusion.

[4]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.

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

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.[6]

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.[7]

In two 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 8 in the main text, we restrict our sample to drivers with a code denoting no prior trucking experience or training. This eliminates drivers with any prior experience or training. It also eliminates drivers who did our survey, but then failed to complete the Firm A training.

Teams. In the data subset (as well as the full data), a small share of driver-weeks involve drivers working in two-person teams (e.g., one person drives while the other sleeps). In the data subset, about 13% of driver-weeks involve a driver working with another driver. For team drivers, the firm equally divides total miles driven among the two drivers in the payroll data provided to us. Excluding team driving weeks, and re-doing the results on overconfidence over time (Figure 1) and the relationship between beliefs and quitting (Table 3), the results are qualitatively similar. We control for whether a driver is a team driver as part of the work type controls.

[6] 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.

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

A.4 Worker Credit Scores

As described in Section 2.2, Firm A drivers have very low average credit scores. Firm A purchased credit scores for drivers in the data subset. 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 constitutes a "subprime" borrower varies by lender, but the cutoff is often 620 or 640. Thus, the majority of drivers in the sample would be considered subprime borrowers. Drivers are especially over-represented among those with very low credit scores, with 43% having scores below 550 compared to only 7% of the US population. The credit score statistics 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 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.

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 E4, 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 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.7 No Effort in the Model

As mentioned in footnote 17, 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.8 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.9 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. To allow for heterogeneity in overconfidence, we have estimated the model allowing for mass point heterogeneity in η_b. Using such estimates did not affect the conclusions of the counterfactual simulations. Our preferred model without overconfidence heterogeneity is computationally simpler. Furthermore, we have (1) Estimated with finer and coarser discretizations of miles (as suggested by Rust, 1987); (2) Eliminated high subjective beliefs 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 decreases the mean belief bias to 506 miles instead of 589 miles in the baseline and 560 miles when Winsorized.

A.10 Additional Information on the Counterfactual Simulation

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

A.10.1 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.

We also calculated profits per worker per week for our counterfactuals (instead of profits per worker) and the conclusions are robust.

A.10.2 Contract Enforcement

As discussed in the text, roughly 30% of quit penalties were collected at the firm (Hoffman and Burks, 2016). 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 22 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.11 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 et al. (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 illustrate using teacher career decisions. Stinebrickner and Stinebrickner (2014) use beliefs about grades to estimate a structural model of the college drop-out 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 et al. (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 et al. (2008); Delavande (2008); Hendren (2012). Zafar (2011) collects subjective beliefs regarding majors, grades, and earnings to analyze whether beliefs data should be used in choice models.

A.12 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. Rustichini et al. (2012) examine whether measures of trainee personality predict experimental measures of worker preferences, worker strategies in experimental games, health behavior, worker retention, and worker accidents.[8] 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 et al., 2015). Whether a driver was hired via referral is uncorrelated with both training contract status and average overconfidence, so excluding referral status from our present analysis should have little impact on our findings.

B Field Experiments with Firm B, Further Information

B.1 Background

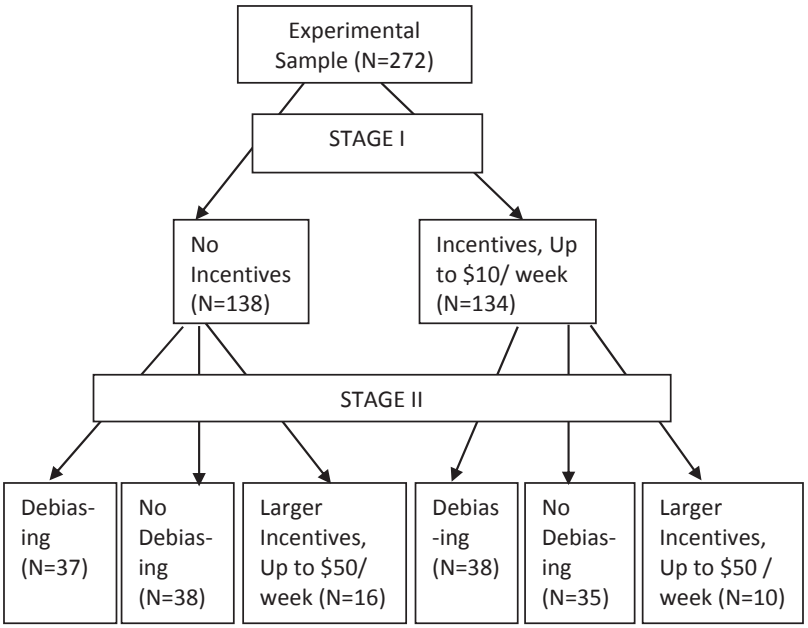
The first goal of the field experiment was to examine if using incentives for accurate guessing would have any effect on productivity beliefs, given that our main beliefs data from Firm A are non-incentivized. In addition, we sought to "test" our main counterfactual of debiasing (that is, of eliminating 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).[9]

[8] Including IQ or personality 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 in those tables.

[9] Compared to other papers estimating structural models using subjective beliefs (Appendix A.11), ours is the only one (to our knowledge) to collect some beliefs using incentives for accuracy. Our result that having quadratic scoring rule incentives does not affect beliefs is consistent with lab experiments by Friedman and Massaro (1998) and

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. (Given that our two goals for the experiment do not involve training contracts, this should not be a concern.) We attempted to contact all workers who had started at the company in the last several months. 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.

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.

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 group and 138 were in the No Incentive group. 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 between the numbers in the two groups because for logistical reasons, workers were randomized to the $10 incentive or control groups before they agreed to participate in the experiment.[10] 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

Sonnemans and Offerman (2001).

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

which experimental economists elicit agents' expectations in an incentive-compatible manner.[11]

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.[12] The 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.

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).

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

[11] 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).

[12] 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: $argmax_b \int A - B(x-b)^2 f(x)dx = argmin_b \int (x-b)^2 f(x)dx$. This leads to a first-order condition of $\int \frac{d}{dx}(x-b)^2 f(x)dx = 0$, which simplifies to $b = \int x f(x)dx$.

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 different experimental groups in attrition.[13]

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 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).[14] For this group, we randomized the order in which drivers would receive the larger incentive.[15]

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-\$15 per hour.[16]

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

[13] 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 group or not, and there was no significant correlation.

[14] 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.

[15] 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 (that is, the order) in which drivers received the larger incentive was randomly assigned. Running a regression of mileage prediction on a dummy for having the larger incentive, driver fixed effects, and week of interview controls, restricting the sample to workers who eventually get the larger incentive, we see no evidence that the larger incentive reduces mileage prediction (though standard errors are fairly large). 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.

[16] Suppose that workers could improve the distance between their weekly guesses and realized miles in the firm data by 253 miles (the level of average worker overconfidence in column 1 of Table B2). Then, subjects' experimental earnings per week calculated based on paying by miles would increase by around \$3 and \$12 in the lower and higher incentive treatments, respectively. Thus, we believe that the experimental design provided subjects with reasonable incentives to guess accurately given subjects' average value of time.

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.

We know of no work on the impact of stake size on the effectiveness of quadratic scoring rules.[17] In experimental economics as a whole, there is no general evidence that experimental results with smaller stakes are undone by using larger stakes.[18] In light of this, we would speculate that our conclusions would not be undermined had we used incentives beyond up to $50.

B.3 Debiasing

Psychologists have long been interested in whether overconfidence and other behavioral biases can be eliminated, focusing primarily on laboratory settings.[19] 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:

[17] To our knowledge, all existing work with quadratic scoring rules uses relatively low stakes, usually payments of up to $20 or less.

[18] 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.

[19] 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, the only paper we are aware of that explicitly attempts to debias overconfidence is 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).

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,[20] 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.[21] For purposes of driver payment, we calculated our best guess of driver earnings.

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 the week relative to 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.[22]

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, at 9 weeks of tenure, eliminating 50% of overconfidence increases quitting by 12 percentage

[20] 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.

[21] We have also done analysis of the impact of incentives and debiasing on earnings predictions. We found no 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.

[22] 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.

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)	-30.9	-54.5	-54.8
	(50.3)	(46.6)	(46.6)
Larger incentives (up to $50/wk)	-4.8	-45.0	-42.7
	(89.8)	(84.9)	(85.1)
Debiasing	-95.9	-112.9*	
	(70.6)	(65.4)	
Debiasing X 0wk post-treat			-206.9**
			(80.2)
Debiasing X 1wk post-treat			-91.2
			(89.4)
Debiasing X 2wk post-treat			-111.4
			(82.0)
Debiasing X 3wk post-treat			-82.4
			(95.2)
Debiasing X 4-6wks post-treat			-62.0
			(73.0)
Joint significance of the two incentive treatments (p-value)	0.824	0.494	0.492
Demographic Controls	No	Yes	Yes
Observations	1,096	1,071	1,071
Mean dep var	2316	2315	2315
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 not worked in a week. 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. * significant at 10%; ** significant at 5%; *** significant at 1%

points.[23] 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.

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). We control for driver experience in all regressions. In column 1, we control for month-year of hire. In columns 2-3, we also add week of tenure dummies and dummies for the number of days not worked in a week. * significant at 10%; ** significant at 5%; *** significant at 1%

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)	-9.8	-31.5	40.1	33.3
	(52.7)	(49.7)	(69.8)	(69.0)
Demographic Controls	No	Yes	No	Yes
Observations	567	556	463	463
Mean dep var	2321	2321	304.3	304.3
Subjects (clusters)	247	241	223	223

Notes: OLS regressions with standard errors clustered by driver in parentheses. An observation is a worker-week. All regressions include dummies for the number of days not worked in a week, worker tenure in weeks. The demographic controls are gender, age, trucking experience, and region of home residence. To limit the effect of outliers, we trim the lower and upper 5% on each dependent variable. * significant at 10%; ** significant at 5%; *** significant at 1%

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] We look at the impact of debiasing at 9 weeks in our counterfactual simulation since this time period corresponds most closely 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 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; dummies for the number of days not worked in a week; worker tenure in weeks; and demographic controls (gender, age, trucking experience, and region of home residence). To limit the effect of outliers, we trim the lower and upper 5% on each dependent variable. * significant at 10%; ** significant at 5%; *** significant at 1%

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?[24]

[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 earnings guess, with which one chosen randomly. Does this make sense? [If not, explain to them.

[24]Note that we made small changes to survey wording over the course of the experiment. 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 the question shift.

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?

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?

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 $W + \epsilon$ 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.[25]

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 where an individual characteristic variable is missing, leaving an estimation sample of 699 drivers.

Probability of Staying. Let $\Lambda(x) = \frac{exp(x)}{1+exp(x)}$. At time T, the probability of staying, given the state variables, is:

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

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 demeaned average of past productivity. That is, *the demeaned 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:

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

[25] 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.

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

$$E^b(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t) = E^b_{y_t} E^b_{\epsilon|y_t}(V_{t+1}(\mathbf{x}_{t+1})|\mathbf{x}_t) \tag{7}$$

$$= E^b_{y_t} E^b_{\epsilon}(max\{\overline{V}^S_{t+1}(\mathbf{x}_{t+1}) + \epsilon^S_{t+1}, \overline{V}^Q_{t+1} + \epsilon^Q_{t+1}\}|\mathbf{x}_t) \tag{8}$$

$$= \int \tau \log\left(\exp(\frac{\overline{V}^S_{t+1}(\mathbf{x}_{t+1})}{\tau}) + \exp(\frac{\overline{V}^Q_{t+1}}{\tau})\right) f^b(y_t|y_1, ..., y_{t-1})dy_t \tag{9}$$

$$= \sum_k \tau \log\left(\exp(\frac{\overline{V}^S_{t+1}(\mathbf{x}_{t+1})}{\tau}) + \exp(\frac{\overline{V}^Q_{t+1}}{\tau})\right) \Pr(y^k_t|y_1, ..., y_{t-1}). \tag{10}$$

(7) expresses that the value function depends on future miles and future idiosyncratic shocks. (8) uses the definition of V, that V^Q_s is independent of the state variable, and that the idiosyncratic shocks are independent of miles. (9) integrates out y_t, which are not observed when the driver makes his period t decision, but are observed in the future, as well as integrates out the idiosyncratic shocks. (10) follows because, in implementation, miles will be discretized into K possible values. The probability $\Pr(y^k_t|y_1, ..., y_{t-1})$ can be easily shown to depend only on $\overline{y}_{t-1}$, and is expressed below in (17). As long as average miles to date by time $t+1$, $\overline{y}_t$, is a sufficient statistic for x_{t+1}, then it follows that average miles by time t, $\overline{y}_{t-1}$, is a sufficient statistic for x_t. Related derivations can be found in Rust (1987) and 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.

$$L_i = \int L(d_{i1}, ..., d_{it}, y_{i1}, ..., y_{it}, b_{i1}, ..., b_{it}|\alpha, \eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b \tag{11}$$

$$= \int \{L(d_{i1}, ..., d_{it}|y_{i1}, ..., y_{it}, b_{i1}, ..., b_{it}, \alpha, \eta_b) * L(b_{i1}, ..., b_{it}|y_{i1}, ..., y_{it}, \alpha, \eta_b) \tag{12}$$
$$*L(y_{i1}, ..., y_{it}|\alpha, \eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b\}$$

$$= \left[\int L(d_{i1}, ..., d_{it}|y_{i1}, ..., y_{it}, \alpha, \eta_b) * L(b_{i1}, ..., b_{it}|y_{i1}, ..., y_{it}, \eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b\right] L(y_{i1}, ..., y_{it}) \tag{13}$$

$$= \left[\int \prod_{s=1}^{t} L(d_{is}|d_{i1}, ..., d_{is-1}, y_{i1}, ..., y_{it}, \alpha, \eta_b) * \prod_{s=1}^{t} L(b_{is}|b_{i1}, ..., b_{is-1}, y_{i1}, ..., y_{it}, \eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b\right] \tag{14}$$
$$*L(y_{i1}, ..., y_{it})$$

$$= \left[\int \prod_{s=1}^{t} L(d_{is}|y_{i1}, ..., y_{is-1}, \alpha, \eta_b) \prod_{s=1}^{t} L(b_{is}|y_{i1} ..., y_{is-1}, \eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b\right] \left(\prod_{s=1}^{t} L(y_{is}|y_{i1} ..., y_{is-1})\right) \tag{15}$$

$$\equiv \{\int L^1_i(\alpha, \eta_b) L^3_i(\eta_b) f(\alpha, \eta_b)\, d\alpha d\eta_b\} L^2_i \tag{16}$$

Equations (11) and (12) follow by the law of total probability. (13) holds because productivity is unaffected by the taste and overconfidence heterogeneity and because beliefs are unaffected by the taste heterogeneity. (14) follows because (a) future miles are not observed when a worker decides to quit and (b) quit decisions are independent of reported subjective beliefs conditional on the overconfidence unobserved heterogeneity. (15) follows because (a) since the ϵ shocks are iid, the decision to quit is conditionally independent of all prior decisions to quit (given the miles realizations and the unobserved heterogeneity) and (b) reported subjective beliefs are conditionally independent of past reported subjective beliefs conditional on productivity and the belief heterogeneity. In (16), we define the part of the likelihood due to the quitting decisions as $L^1_i(\alpha, \eta_b)$, the part due to the miles realizations as L^2_i, and the part due to subjective beliefs as $L^3_i(\eta_b)$.

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

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

with

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

and 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 $\gamma_s^b = \frac{s\sigma_0^2}{s\sigma_0^2 + \widetilde{\sigma_y}^2}$.[26]

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.

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

Transitions Between Miles. As mentioned in the main text, 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, $K = 14$). Transitions between miles states are given by:

$$\Pr(y_s^k|y_1, ..., y_{s-1}) = \Phi\left(\frac{y_s^k + .5 * kstep - E^b(y_s^k|y_1, ..., y_{s-1})}{\sqrt{\Omega_{s-1}}}\right) - \Phi\left(\frac{y_s^k - .5 * kstep - E^b(y_s^k|y_1, ..., y_{s-1})}{\sqrt{\Omega_{s-1}}}\right) \quad (17)$$

kstep is the distance between grid points. See Rust (1996) and Stange (2012) for similar formulas. 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.

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). We these in hand, we use backwards recursion to solve for the choice-specific value function V_t^S and V_t^Q for $t = 1, ..., T$.

12-Month and 18-Month Contracts in Structural Model. The quit penalties under the training contracts varied slightly by school. 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. For the 18-month contract, we assume a penalty of \$5,250 that is reduced linearly over 78 weeks to 0.

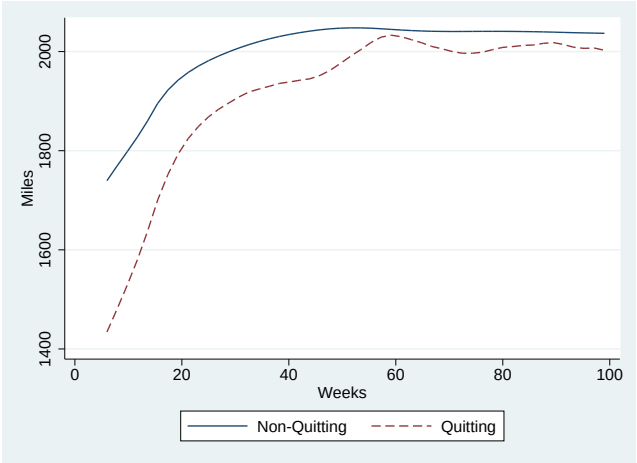
[26] This follows by applying the standard formula for the conditional density for a multivariate normal distribution: $X_1|(X_2 = x_2) \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$.

Zero Mile Weeks. The data contain a significant number of zero mile weeks for drivers. These weeks may occur for several reasons, including that the driver is not working or is working in a role that does not pay by miles. 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.

Compensation and Additional Bonuses. At Firm A, Drivers may receive small quarterly bonuses (based on customer/shipper satisfaction, good fuel economy, and other factors).[27] 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 all 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 pay per mile continues to increase after three years of tenure, for simplicity in our model, we assume that pay per mile for three years and after is the same as the rate before the three-year mark.

E Further Reduced-Form Results

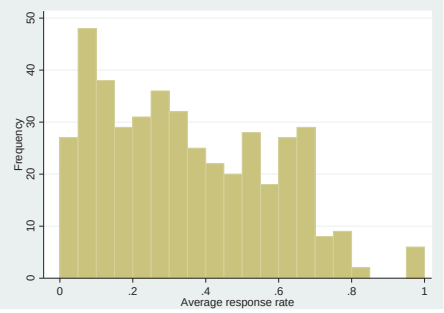
Figure E1: Conditional on Week of Tenure, Workers who Quit that Week Have Lower Average Productivity to Date Than Workers Who Stay



Notes: This figure analyzes the average miles to date in the full sample (excluding weeks where miles are not positive). In each week, it compares the average miles of drivers who quit with those of drivers who do not quit that week. For example, it can be used to compare the average miles per week (from weeks 1-19) of drivers who quit in the 20th week with the average miles per week (from weeks 1-19) of drivers who survive to the 20th week, but do not quit. Zero mile weeks are excluded. The figures are plotted using a local polynomial regression with an Epanechnikov kernel (bandwidth = 3 weeks).

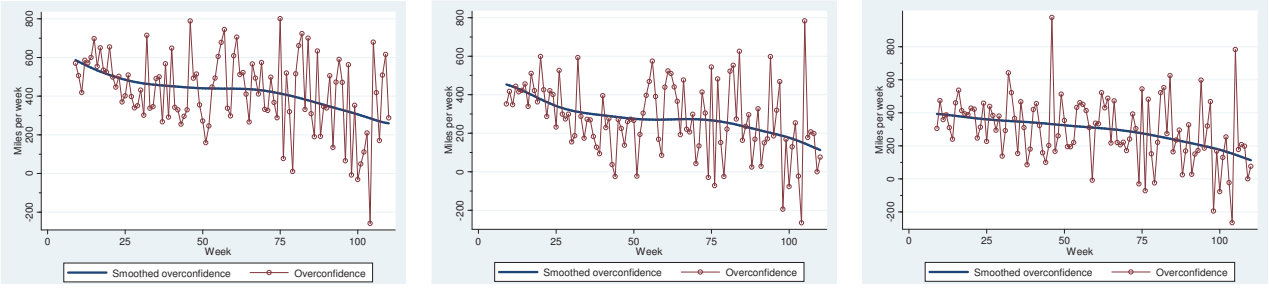
[27] At some points in the past, new inexperienced drivers only became eligible to receive a quarterly bonus after one year of tenure.

Figure E2: 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.

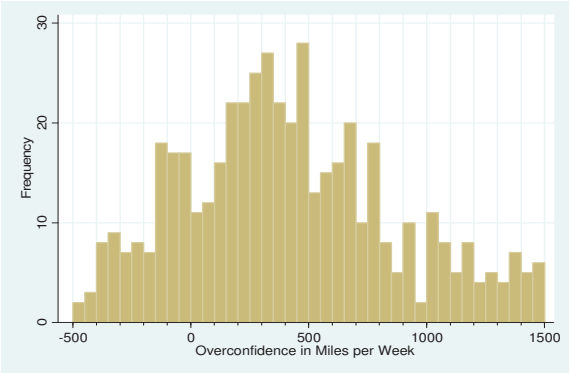
Figure E3: Tenure and Overconfidence (Productivity Beliefs *minus* Productivity)



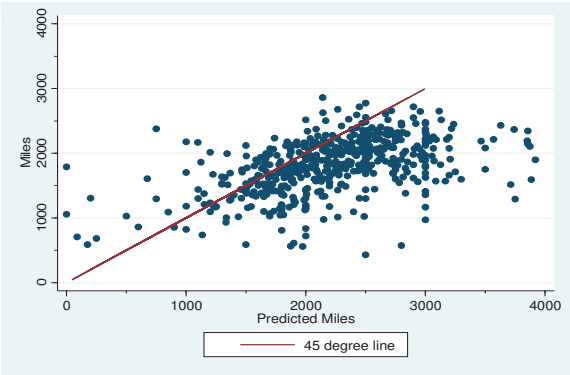
(a) Means (b) Medians (c) Medians, ≥75 Wks, Responder

Notes: This figure analyzes the evolution of average driver overconfidence as a function of driver tenure in the data subset. Overprediction, defined as productivity beliefs *minus* realized productivity, is collapsed (across all drivers) by week of tenure. The dots correspond to the collapsed means. The smoothed curve is plotted using a lowess regression with a bandwidth of 0.5. 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. Across these figures (and several others not reported with different samples), we see a slow pattern of gradual reduction in overprediction. Regressions of overprediction on tenure including control variables also confirm the pattern of slowly decreasing overprediction.

Figure E4: Distribution of Overconfidence Across Drivers



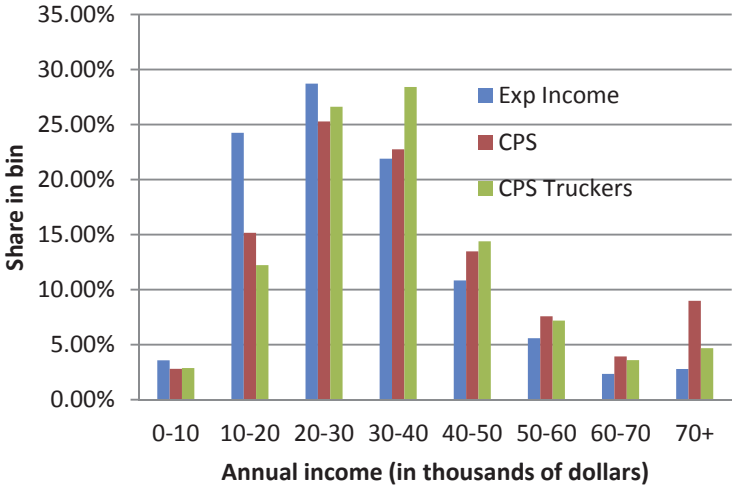
(a) Dist of Individual Overconfidence



(b) Predicted Miles and Actual Miles

Notes: This figure analyzes overconfidence among drivers in the data subset. It 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 beliefs and productivity. A driver's overconfidence level is calculated by averaging over all the weeks with productivity beliefs and actual productivity. In panel (b), each driver is represented by a dot located at their average productivity and beliefs (averaged over all weeks the driver is observed).

Figure E5: 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: $\$0 - \$10,000$, $\$10,000 - \$20,000$, $\$20,000 - \$30,000$, $\$30,000 - \$40,000$, $\$40,000 - \$50,000$, $\$50,000 - \$60,000$, $\$60,000 - \$70,000$, $\$70,000+$. The CPS comparison data are from the 2006 March CPS. "CPS" is the average income and earnings for 35-year old male workers with a high school degree who worked full-time last year. "CPS Truckers" is the average income and earnings for 30-40 year old male workers with a high school degree who work as truckdrivers (Occ=913). *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 Productivity? OLS Regressions at Firm B

	(1)	(2)	(3)	(4)	(5)	(6)
L. Pred miles	0.292***	0.290***	0.254***	0.133**	0.136**	0.049
	(0.052)	(0.053)	(0.065)	(0.052)	(0.053)	(0.051)
L. Avg miles to date				0.572***	0.584***	
				(0.077)	(0.075)	
$10 Incentive			-2.073			
			(2.740)			
$10 Incentive X L. Avg miles to date			0.091			
			(0.116)			
$50 Incentive			-4.516			
			(6.870)			
$50 Incentive X L. Avg miles to date			0.184			
			(0.307)			
Demographic Controls	No	Yes	Yes	No	Yes	No
Subject FE	No	No	No	No	No	Yes
Observations	803	803	803	695	695	803
R-squared	0.155	0.163	0.164	0.298	0.306	0.570

Notes: The dependent variable is miles driven per week (in hundreds). An observation is a driver-week. Standard errors clustered by driver in parentheses. The demographic controls are gender, age, trucking experience, and region of home residence. All regressions include worker tenure in weeks and dummies for the number of days not worked in a week. 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. We present the data here to show, as at Firm A, 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 productive of actual productivity when they are financially incentivized. * significant at 10%; ** significant at 5%; *** significant at 1%

Table E2: 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: An observation is a driver-week. The regressions are Cox proportional hazard models, where the dependent variable is quitting. Events where the driver is fired are treated as censored. Standard errors clustered by worker are in parentheses. Demographic controls are controls for gender, race, marital status, age, and education. Productivity is given in terms of hundreds of miles driven per week. Column 3 differs from column 2 in that it restricts to the sample of driver-weeks for which there is a corresponding belief expectation. All drivers are from the same training school and were hired in late 2005 or 2006. 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. * significant at 10%; ** significant at 5%; *** significant at 1%

Table E3: Do Productivity Beliefs Predict Quitting? Robustness Check with Lagged Values

	(1)	(2)	(3)	(4)	(5)	(6)
L. Predicted miles	-0.028*	-0.035*			-0.029	-0.032
	(0.017)	(0.019)			(0.019)	(0.020)
L. Avg miles to date			-0.053***	-0.039	0.007	-0.019
			(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	8,345	8,345	32,649	8,345	8,345	8,345

Notes: Column 4 differs from column 3 in that it restricts to the sample of driver-weeks for which there is a corresponding belief expectation. The notes are otherwise the same as in Table E2. While the coefficients in columns 5-6 just miss standard levels of statistical significance, they represent economically sizeable reductions in quitting, e.g., moving from the 50th (2,300) to the 90th percentile (3,000) in beliefs reduces quitting by 23% in column 6. * significant at 10%; ** significant at 5%; *** significant at 1%

Table E4: 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: Column 4 differs from column 3 in that it restricts to the sample of driver-weeks for which there is a corresponding belief expectation. The notes are otherwise the same as in Table E2. * significant at 10%; ** significant at 5%; *** significant at 1%

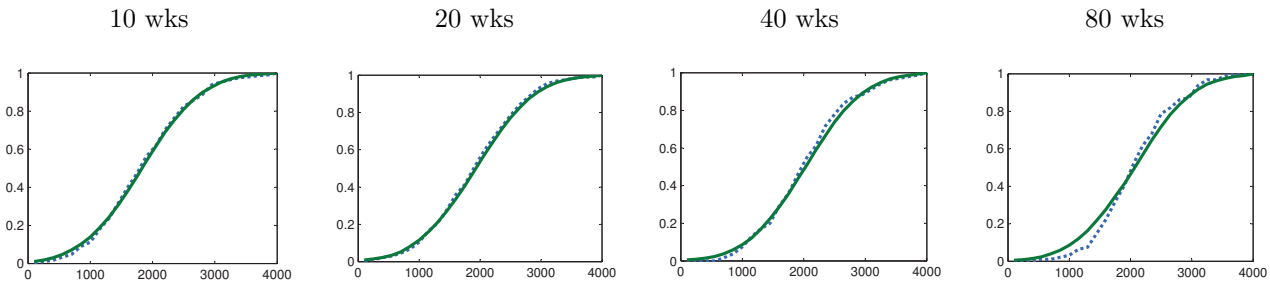
Table E5: Do Workers Update their Subjective Productivity Beliefs?

	(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				0.564***	
				(0.078)	
L. Miles				0.091***	
				(0.016)	
Demographic 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	0.332	0.614

Notes: 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 specifications include work type controls and week of tenure dummies. Demographic controls are controls for gender, race, marital status, age, and education. * significant at 10%; ** significant at 5%; *** significant at 1%

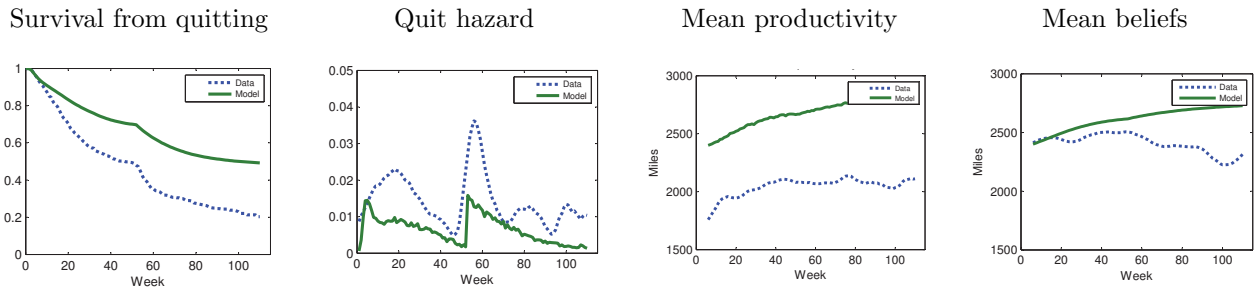
F Further Structural Results

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



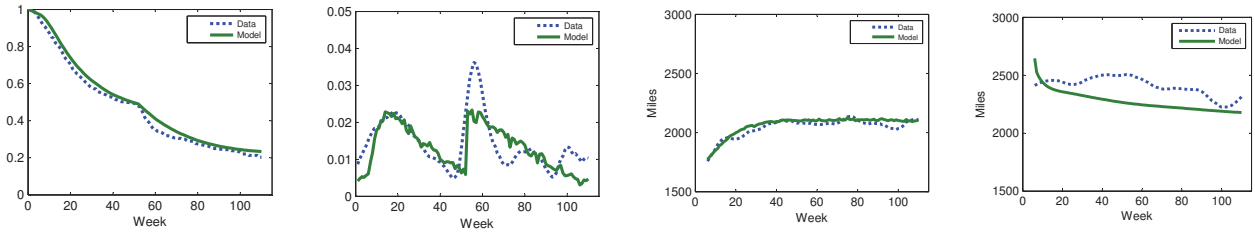
Notes: These figures plots the distribution of productivity at different tenure levels, both in the data and as simulated by the model. The data are from 699 drivers under the 12-month contract. The model is simulated using 40,000 simulated drivers. The figures show the model can well predict the changing distribution of worker productivity as tenure increases.

Figure F2: Model Fit: No Belief Bias



Notes: The notes are the same as for Figure 2 except that the model simulated corresponds to Column 1 in Table 5.

Figure F3: 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, $\widetilde{\sigma_y} = \sigma_y$.

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	2025	2024	2024	2025	2025	2025	2025
		(16.7)	(16.5)	(13)	(16.6)	(16.7)	(16.7)	(16.6)
σ_0	Std dev of prior productivity dist	286	284	286	286	287	286	285
		(10.8)	(10.7)	(8.4)	(10.8)	(10.8)	(10.8)	(10.7)
σ_y	Std dev of productivity shocks	706	706	708	707	707	706	706
		(3.6)	(3.6)	(2.8)	(3.6)	(3.6)	(3.6)	(3.6)
s_0	Value of skills gained in wks 1-5	8.6	3.8	8.7	9.3	8.7	12.0	7.9
		(4.4)	(2.8)	(3.5)	(5.6)	(4.4)	(4.4)	(3.8)
	Taste UH Parameters							
μ_1	Mass point 1 of taste UH	-260	-325	-260	-252	-255	-100	-221
		(14)	(14)	(11)	(21)	(14)	(14)	(8)
μ_2	Mass point 2 of taste UH	-135	-166	-134	-120	-136	25	-108
		(12)	(11)	(10)	(15)	(13)	(13)	(19)
μ_3	Mass point 3 of taste UH	135	119	140	164	109	295	160
		(41)	(40)	(33)	(45)	(36)	(41)	(39)
p_1	Probability type 1	0.34	0.31	0.34	0.31	0.35	0.34	0.56
		(0.07)	(0.05)	(0.05)	(0.08)	(0.07)	(0.07)	(0.06)
p_2	Probability type 2	0.43	0.46	0.44	0.48	0.43	0.43	0.24
		(0.06)	(0.05)	(0.05)	(0.08)	(0.07)	(0.06)	(0.05)
	Belief Parameters							
η_b	Belief bias	589	591	590	560	588	589	609
		(28)	(28)	(22)	(25)	(28)	(28)	(26)
$\widetilde{\sigma_y}$	Believed std dev of productivity shocks	1888	1847	1874	1606	1899	1888	1784
		(136)	(134)	(105)	(98)	(137)	(136)	(115)
σ_b	Std dev in beliefs	298	298	299	259	298	298	298
		(1.4)	(1.4)	(1.1)	(1.2)	(1.4)	(1.4)	(1.4)
	Scale Parameter							
τ	Scale param of idiosyncratic shock	2208	2776	2220	2781	2186	2208	1979
		(350)	(462)	(272)	(513)	(346)	(350)	(253)
	Log-likelihood	-90865	-90869	-90978	-89080	-90865	-90865	-90851
	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 in parentheses. 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; results are similar also with δ corresponding to an annual discount factor of 0.95. 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.

G Miscellaneous

G.1 Measuring Productivity

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 8 days. This translates to a federal legal limit of roughly 60 hours per calendar 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.[28] 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.

In addition, there is a small amount of measurement error in our productivity measure, miles per week. We describe the source of the measurement error, explain how we correct for it, and show it has little impact on our estimates and conclusions.[29]

Miles are imperfectly observed each week because miles are only recorded once a driver reaches his destination. 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.[30]

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 $t-1$

[28] 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.

[29] 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.

[30] 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 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," 2014. Mimeo, Duke.

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

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," *Proceedings of the National Academy of Science*, 2009, *106* (19), 7745–7750.

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.

Card, David, Alexandre Mas, Enrico Moretti, and Emmanuel Saez, "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction," *American Economic Review*, 2012, *102* (6), 2981–3003.

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

Choi, Syngjoo, Shachar Kariv, Wieland Muller, and Dan Silverman, "Who is More Rational?," *American Economic Review*, 2014, *Forthcoming.*

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

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," 2012. Mimeo.

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

— and Stephen Burks, "Training Contracts, Employee Turnover, and the Returns from Firm-sponsored General Training," 2016. Mimeo, University of Toronto.

Holt, Charles A. and Angela M. Smith, "An Update on Bayesian Updating," *Journal of Economic Behavior and Organization*, 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.

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?," *Journal 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 P. 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 E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, 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.

__ , "Numerical Dynamic Programming in Economics," in H. M. Amman, D. A. Kendrick, and J. Rust, eds., *Handbook of Computational Economics*, Vol. 1 of *Handbook of Computational Economics*, Elsevier, 1996, chapter 14, pp. 619–729.

Rustichini, Aldo, Colin DeYoung, Jon Anderson, and Stephen Burks, "Toward the Integration of Personality Theory and Decision Theory in the Explanation of Economic Behavior," Mimeo, University of Minnesota 2012.

Sanna, Lawrence J. et al., "When Debiasing Backfires: Accessible Content and Accessibility Experiences in Debiasing Hindsight," *Journal 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, Todd and Ralph Stinebrickner, "Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model," *Journal of Labor Economics*, 2014, *Forthcoming*.

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.