**Online supplementary appendix for:**

***Racial Disparities in Occupational Safety and Health***

In this appendix we provide additional information on the data and methods four our study that was not included in the full manuscript due to space constraints. This includes some additional detail about study design, a comparison of the samples included in the study, and full regression results.

**Study Design**

 We used two nationally representative large survey datasets published by the US Census Bureau to capture information on racial disparities in workplace injury risk: the 2006-2013 American Community Survey (ACS) and the 1996, 2001, 2004 and 2008 panels of the Survey of Income and Program Participation (SIPP). The ACS data were used to construct what we call “expected” workplace injury rates. These are equal to the rates of expected workplace injury that a person faces based on the types of jobs someone in that race-age-gender-education group typically faces. The SIPP data were used to estimate the prevalence disability due to workplace injury, which cannot be observed in the ACS. The expected workplace injury rates inform as to differences in exposure to workplace injury risk, while the prevalence of disability informs as to how the differential exposure (if any) contributes to differences in population health.

 All calculations for this study were done using STATA/MP Version 14. This study was deemed exempt from Human Subjects review because it used only publicly available, non-identifiable data.

Constructing expected workplace injury rates

The key advantage of the ACS for our study was that it is very large, allowing us to incorporate differences in employment tendencies for fairly detailed breakdowns on age, education, gender and race at the 4-digit occupation level.[[1]](#footnote-1)

 As noted in the text, to measure occupational risk, we identified individuals who were working age (18 to 64) and employed at least one week in the previous year using the ACS. We then identified the primary occupation – defined as the occupation that you worked for the most in the previous year – and linked workers on the basis of occupation to injury rate data from the Bureau of Labor Statistics (BLS). The BLS publishes annual data on work-related injuries from the annual Survey of Occupational Injuries and Illnesses (SOII). The BLS collects data from a sample of more than 200,000 firms annually and publishes aggregate statistics based on these reports (the SOII), including data on injury rates by detailed occupation. We focused on lost-workday injuries (as opposed to all injuries) because they are more likely to result in long-term health problems and disabilities. The ACS and BLS data were merged from 2006 to 2012, reflecting the years for which we had data from both sources. We merged injury data from both sources at the occupation level using 4-digit codes from the Standard Occupational Classification (SOC) system. SOC codes are created and maintained by the BLS for the purposes of classifying workers’ occupations for data collection and reporting purposes. As of 2010, there were 840 detailed occupations in the SOC system.

 We used these data to estimate the number of lost workday injuries per 1,000 workers at the race-age-gender-education level. Note that we refer to this as the expected workplace injury rate because it is calculated by taking the weighted average of the injury rates across all jobs, with the weights being the share of people in each group in each job. If one race-age-gender-education combination had a relatively high share of individuals working in high-risk jobs such as construction, that group would have had a higher expected workplace injury rate (all else equal).

Note that the BLS reports injury rates based on full-time equivalents (FTEs), based on an assumption of 2,000 hours per year. To adjust for possible differences in hours worked per year across racial groups, we adjusted the injury rate for each worker according to the percent of an FTE she worked (see Appendix for more detail on how this was done). So, suppose someone worked 1,000 hours in a job that had a reported injury rate of 10 injuries per 1,000 workers. In this case, we would assign an expected workplace injury rate of 5 per 1,000 to that individual. This approach has some limitations, because it assumes injury risk is directly proportional to hours worked and ignores the possibility that part-time workers may differ systematically (e.g., they may work more dangerous shifts or have less skill, and be injured with a likelihood that is more than proportional to hours worked). However, we feel this approach is more consistent with the BLS data and we note that the qualitative findings are the same if we do not adjust for hours worked. Note that this approach means that individuals in groups that work fewer hours or are less likely to have any job will have lower injury rates even if the jobs they do have are riskier.

Constructing disability prevalence estimates

The SIPP is a smaller survey, that isn’t as useful for providing breakdowns on the occupation level. However, the SIPP do contain similar data on demographics and labor market outcomes, and they also include detailed information on disability status. Importantly for the purposes of this study, the SIPP include (self-reported) detail on whether or not a disability was caused due to a work-related injury, which is not available in the ACS.[[2]](#footnote-2)

The SIPP collects information on respondents on a monthly basis for up to 4 years. So, for individuals in the 2004 panel, core data elements are available monthly from January 2004 to December 2007. Data are collected in four-month waves, and different waves include “topical modules” that ask supplemental questions on selected topics. The survey units are households, and all members of the household age 15+ are surveyed. The topical modules can be linked to the core data files at the person level.

The SIPP was collected annually from 1984 to 1993, but after a redesign in 1996 it was collected at varying intervals (1996, 2001, 2004 and 2008). The 1996 redesign changed the panel structure of the data, resulting in larger sample but with fewer panels, though much of the actual survey was unchanged. We used only data from 1996+ because the race and ethnicity variables were different in the earlier years of the survey.

A topical module asked in the second wave of each panel includes questions that cover “functional limitations and disability; health and disability; health status and utilization of health care services; long-term care; medical expenses and work disability and work disability history.” In particular, questions were asked about the presence and nature of any work limitations, the date at which any work limitation began, whether the respondent worked at the time the limitation began, whether the limitation was due to an accident or injury and whether the accident or injury occurred at work.

Specifically, the SIPP asks the following sequence of questions that we used to identify work-related disabilities:

LMTVER

I have recorded that [fill HISHER] health or condition limits the kind or amount of work [fill HESHE] can do. Is that correct? [Yes/No]

MNCAUS

MAIN CONDITION: [fill TEMP] ASK OR VERIFY: Was this condition caused by an accident or injury? [Yes/No]

MNLOC

ASK OR VERIFY: Where did the accident or injury take place? Was it...

(1) ...on the job?

(2) ...during service in the Armed Forces?

(3) ...in the home?

(4) ...or somewhere else?

Using these questions, we identified someone as having a work-related disability if they (1) answered yes to the first question (indicating they had a disability), (2) answered yes to the second question (indicating the disability was caused by an accident or injury), and (3) responded with 1 on the third question (indicating the disability was caused by an accident or injury at work). We use the term work-related disability even though we acknowledge that we miss some disabilities caused by work-related illnesses. While important, work-related illnesses cause a relatively small fraction of lost-workday cases (less than 10% typically). Despite this limitation, we use the term work-related disability as opposed to disability caused by a workplace injury for ease of exposition and because of word limitations.

 We restricted the data to those individuals who had ever worked, because those who never work are by definition not subject to work-related disabilities. With these data, we estimated the overall prevalence of a work-limiting disability, the prevalence of a disability that was caused by a workplace injury (which we refer to as work-related disability), and the percent of disability attributable to disability.

Demographic breakdowns

 For our racial breakdowns, we compared non-Hispanic whites to non-Hispanic blacks, Hispanics, Asians and a general category of “Other race” that captured groups that were too small in our samples to be broken out separately. Because average economic opportunities and job types may differ between native-born and foreign-born Hispanics, we considered these groups separately.

Note that in principle you may see similar differences across other ethnic groups according to whether they were native-born or foreign born, but for most other races the share of foreign-born individuals was too small (typically less than 10%, compared to roughly 50% for Hispanics). One exception was Asians, where the share was larger, but the average characteristics (e.g., education, earnings and job risk) were generally similar for foreign-born and native-born Asians, which was not the case for Hispanics.

In terms of education, we grouped individuals in terms of less than high school, high school diploma with no college, some college, or 4-year college degree or higher. We created broad age categories corresponding to 18 to 29 years, 30 to 39, 40 to 49 and 50 to 64. We dropped individuals less than 18 or over 64 to focus on the working-age population. In principle, we could have included the 65+ population for the disability prevalence sample, but we were concerned that there could be differences in self-reported disability for the retired population that would make it inconsistent with the working age population (e.g., individuals may not consider a health limitation to be work-limiting if they were already expecting to be retired).

Statistical Analysis

 We conducted five separate analyses in the course of this study, the results of each being described in Exhibits 1-4 in the text (analysis 2 was not included in the manuscript to save space and the results are reported in Appendix Table 2):

1. Unadjusted comparison of the expected workplace injury rates according to race, overall and separately by male/female. (ACS sample)
2. Adjusted comparison of the expected workplace injury rate by race while conditioning on age, education and race. (ACS sample)
3. Unadjusted prevalence of work-related disability by race for the younger (18 to 29) and older (50 to 64) age groups. (SIPP)
4. Unadjusted and adjusted odds ratios of having a work-related disability for each racial group compared to non-Hispanic whites. (SIPP)
5. Predicted probability of a work-related disability by race for two scenarios – one where expected job risk is the same for all groups (equal to the mean value for whites) and one in which different racial groups have their own mean value of expected job risk.

Analyses 1 and 3 were simple unadjusted mean comparisons across different subpopulations. These means were constructed using sampling weights from the ACS and SIPP that reflected the design of each survey.

For Analysis 2, we used linear regression to estimate the impact of race on injury rate. Note that we used linear regression because the outcome variable was the expected workplace injury rate at the group level, not a binary indicator of whether any given individual was injured. This can be interpreted as examining the predicted injury rate for different racial groups as if each race category had the same breakdown of gender, age and educational mix. Standard errors and *p*-values for this regression were computed using heteroscedasticity-consistent “robust” variance estimates.

 An alternative approach would have been to not aggregate to the group level before estimating the regression, and running an individual regression using each person’s job risk as the dependent variable. In principle, as long as we cluster the standard errors at the occupation level, this approach should provide valid inference. However, we feel that the group-level regression is more intuitive and better represents what is available in the data. We also note that the choice is somewhat immaterial, because the results are nearly identical with the individual-level regression.

We used a different approach for Analysis 4. Because we had individual data on disability prevalence, we used multivariable logistic regression to test for racial differences in the probability of a work-related disability, with and without conditioning on other confounding factors. The logic behind comparing the unadjusted and adjusted odds ratios was to study the extent to which differences in disability prevalence across racial groups was explained by observable characteristics of individuals. The covariates included age, gender, and education. We also merged the expected workplace injury risk from the SIPP data at the race-age-education-gender level. Because we adjust for hours worked, this incorporates both the expected risk when a group is working and the possibility that they are less likely to work. We use the expected job risk to test for how the types of jobs someone in a certain demographic is more likely to hold predicts the prevalence of work-related disability.

Note that we did not include contemporaneous job characteristics—e.g., earnings or occupation—for the SIPP respondents as confounders. The reason for this is that these could be affected by the presence of health limitations. These factors could also be related to race and ethnicity, so including them could over-control for the disparities we are trying to detect. In this model, where we were interested in comparing the odds ratio estimates, we simply computed odds ratios overall and separately for the youngest and oldest individuals, rather than estimate a full set of interaction terms. As with Analysis 2, standard errors and *p*-values for this regression were computed using heteroscedasticity-consistent “robust” variance estimates.

For analysis 5, we used the logistic regressions to estimate the predicted probability that an individual has a work-related disability under two scenarios. The first scenario was a hypothetical in which we computed the predicted probability of a work-related disability for each race under the hypothetical scenario where we held the expected workplace injury rate constant at the mean value for white workers. In the second scenario, we computed the predicted probability of a work-related disability with the expected workplace injury rate equal to the mean value for each race category. In both cases we focused on those age 50-64 because work-related disability prevalence is relatively low in younger populations. The difference between these two sets of probabilities indicates how disparities in the expected workplace injury rates based on job types are associated with disparities in the prevalence of work-related disability.

Though not directly reported in any exhibit in the text, the mean expected workplace injury rate among ages 50-64 for whites was 8.8 per 1,000 workers, compared to 12.3, 11.3, 13.5, 9.6 and 9.5 for black, native-born Hispanic, foreign-born Hispanic, Asian and other race workers, respectively.

**Sample Comparison**

 Appendix Table 1 reports the characteristics of individuals in the matched ACS sample and the SIPP. As discussed in the text,the linked ACS data used information from 11,632,466 respondents to estimate the expected workplace injury rates for 432 combinations of race-age category, gender and education). From the SIPP we had data on 198,308 respondents who were 18 to 64 and ever worked from the 1996, 2001, 2004 and 2008 panels. Overall, the demographic features of the two samples were quite similar (**Appendix Table 1**). We expect that there are two factors that explain those differences that are present, most notably a higher share of males (52.5% compared to 49.9%) and a lower share of non-Hispanic whites (66.6% compared to 71.4%, respectively) in the ACS compared to the SIPP. The higher share of males in the ACS sample was most likely due to the fact that we restricted the sample to current workers (as opposed to those who ever worked in the SIPP). The lower share of whites in the ACS data was most likely because those data are more recent, and the share of the white population has been declining over time.

**Regression results**

 **Appendix Table 2** reports the regression results for the adjusted analysis of workplace injury rates by race using the ACS data. **Appendix Table 3** reports the full multivariable logistic regression results that generated the odds ratios for Exhibit 5 in the text.

|  |
| --- |
| **Appendix Table 1. Summary of demographics, labor market outcomes and disability according to race and ethnicity** |
|  | ACS | SIPP |
| *Years* | 2006-2012 | 1996, 2001, 2004, 2008 |
| *Demographics* |  |  |
| Age 18 to 29, % | 26.5% | 22.5% |
| Age 30 to 39, % | 22.3% | 24.6% |
| Age 40 to 49, % | 24.1% | 25.8% |
| Age 50 to 64, % | 27.2% | 27.1% |
| Male, % | 52.5% | 49.9% |
| Native born, % | 83.9% | 86.9% |
| *Race* |  |  |
| White, Non-Hispanic | 66.6% | 71.4% |
| Black, Non-Hispanic | 11.1% | 11.2% |
| Hispanic, Native born | 7.1% | 6.2% |
| Hispanic, Foreign born | 7.9% | 5.8% |
| Asian | 4.9% | 3.1% |
| Other race | 2.3% | 2.3% |
| *Education* |  |  |
| No high school diploma, % | 10.2% | 11.1% |
| High school, % | 26.7% | 28.0% |
| Some college, % | 33.2% | 34.3% |
| 4 year college degree or higher, % | 29.9% | 26.6% |
|  |  |  |
| Observations | 11,632,466 | 198,308 |
|  |

|  |
| --- |
| **Appendix Table 2. Regression results for average injury rate in the ACS sample** |
| *Outcome: Expected injury rate per 1,000 workers* |
|  | Coefficient estimate (standard error) |
|  |  |
| Omitted: White, Non-Hispanic | - |
|  |  |
| Black, Non-Hispanic | 1.464\*\*\* |
|  | (0.376) |
| Hispanic, Native born | 0.316 |
|  | (0.344) |
| Hispanic, Foreign born | 2.269\*\*\* |
|  | (0.309) |
| Asian | -0.763\*\* |
|  | (0.333) |
| Other race | 0.214 |
|  | (0.323) |
|  |  |
| Omitted: Age 18 to 29 years | - |
|  |  |
| Age 30 to 39, % | 2.407\*\*\* |
|  | (0.279) |
| Age 40 to 49, % | 3.070\*\*\* |
|  | (0.277) |
| Age 50 to 64, % | 3.044\*\*\* |
|  | (0.291) |
|  |  |
| Male | 3.659\*\*\* |
|  | (0.181) |
|  |  |
| Omitted: No high school diploma | - |
|  |  |
| High school, % | -0.368 |
|  | (0.261) |
| Some college, % | -2.523\*\*\* |
|  | (0.239) |
| 4 year college degree or higher, % | -5.936\*\*\* |
|  | (0.297) |
|  |  |
| Observations | 192 |
| R-squared | 0.887 |
| Note: Outcome is the average number of workplace injuries per 1,000 workers (mean value in estimation sample of 10.5). Standard errors are computed using heteroscedasticity-consistent robust variance estimates. A \*, \*\* or \*\*\* indicates statistical significance at the 10%, 5% or 1% level, respectively. |

|  |
| --- |
| **Appendix Table 3.** **Logistic regression results of the prevalence of work-related disability as a function of race and other respondent characteristics, SIPP sample** |
|  | I | II | III | IV | V | VI |
| Outcome: Had work-related disability |
|  |  |  |  |  |  |  |
| Odds ratios for difference from White, Non-Hispanic(*p-*value) |
| Black, Non-Hispanic | 1.268\*\*\* | 0.820\*\*\* | 0.716 | 0.560\* | 1.806\*\*\* | 1.062 |
|  | (0.000) | (0.001) | (0.305) | (0.061) | (0.000) | (0.546) |
| Hispanic, native | 1.068 | 1.015 | 1.229 | 1.106 | 1.434\*\* | 0.987 |
|  | (0.394) | (0.848) | (0.471) | (0.708) | (0.011) | (0.921) |
| Hispanic, foreign | 1.022 | 0.582\*\*\* | 2.169\*\*\* | 1.172 | 1.698\*\*\* | 0.814 |
|  | (0.782) | (0.000) | (0.003) | (0.635) | (0.000) | (0.131) |
| Asian | 1.286\*\*\* | 1.452\*\*\* | 0.947 | 1.159 | 1.643\*\*\* | 1.638\*\*\* |
|  | (0.004) | (0.000) | (0.907) | (0.723) | (0.000) | (0.000) |
| Other race | 0.690\*\*\* | 0.774\* | 0.894 | 1.300 | 0.948 | 0.885 |
|  | (0.006) | (0.060) | (0.827) | (0.610) | (0.783) | (0.515) |
|  |  |  |  |  |  |  |
| Age |  | 1.050\*\*\* |  | 1.162\*\*\* |  | 1.020\*\*\* |
|  |  | (0.000) |  | (0.000) |  | (0.004) |
| Male |  | 0.801\*\*\* |  | 1.379 |  | 1.032 |
|  |  | (0.003) |  | (0.361) |  | (0.797) |
| *Education (4+ degree or higher omitted)* |  |  |  |  |
| No high school |  | 2.111\*\*\* |  | 3.617\*\* |  | 2.768\*\*\* |
|  |  | (0.000) |  | (0.017) |  | (0.000) |
| High school |  | 1.453\*\*\* |  | 3.470\*\* |  | 1.733\*\*\* |
|  |  | (0.001) |  | (0.022) |  | (0.004) |
| Some college |  | 2.180\*\*\* |  | 2.791\*\* |  | 2.648\*\*\* |
|  |  | (0.000) |  | (0.019) |  | (0.000) |
|  |  |  |  |  |  |  |
| *Survey year (2008 panel omitted)* |  |  |  |  |
| 1996 |  | 1.079 |  | 1.676\*\* |  | 0.898 |
|  |  | (0.115) |  | (0.014) |  | (0.131) |
| 2001 |  | 0.911\* |  | 0.793 |  | 0.787\*\*\* |
|  |  | (0.080) |  | (0.384) |  | (0.002) |
| 2004 |  | 1.062 |  | 1.017 |  | 0.982 |
|  |  | (0.197) |  | (0.941) |  | (0.779) |
|  |  |  |  |  |  |  |
| Expected injury rate |  | 1.171\*\*\* |  | 1.050 |  | 1.114\*\*\* |
|  |  | (0.000) |  | (0.463) |  | (0.000) |
|  |  |  |  |  |  |  |
| Observations | 198,308 | 198,308 | 42,043 | 42,043 | 55,615 | 55,615 |
| Note: Results from logistic regression, where outcome is a binary indicator for whether the respondent self-reported a disability caused by a workplace injury. The *p*-values in parentheses are computed using heteroscedasticity-consistent robust variance estimates. A \*, \*\* or \*\*\* indicates statistical significance at the 10%, 5% or 1% level, respectively. |

1. Detailed information on the ACS, including sampling procedures, data dictionaries, questionnaires and downloads, are available here:

<http://www.census.gov/programs-surveys/acs/data.html> (accessed on September 3, 2016). [↑](#footnote-ref-1)
2. Detailed information on the SIPP, including sampling procedures, data dictionaries, questionnaires and downloads, are available here:

<http://www.census.gov/programs-surveys/sipp/data.html> (accessed on September 3, 2016). [↑](#footnote-ref-2)