Experimental Age Discrimination Evidence and the Heckman Critique

By David Neumark, Ian Burn, and Patrick Button
American Economic Review (2016)

James J. Heckman

THE UNIVERSITY OF CHICAGO

Econ 350, Winter 2021

I. Evidence on Age Discrimination

- Experimental audit or correspondence (AC) studies can provide compelling evidence on discrimination in hiring decisions.

- Both types of studies use fictitious job applicants.

- Audit studies use in-person applicants leading to actual job offers.

- Correspondence studies create paper or electronic applicants, and capture "callbacks" for job interviews, avoiding experimenter effects and making feasible the collection of very large samples.

- Existing field experiments on age almost always find substantial age discrimination in hiring.

- For example, Bendick, Jackson, and Romero (1997) find that in 43 percent of pairs only younger applicants (age 32) received positive responses, versus 16.5 percent for older applicants (age 57)—a "net discrimination" estimate of 26.5 percent.

- Heckman (1998) argues, however, that differences in the variances of unobservables, which the study design cannot eliminate, can create biases in either direction.
- This problem could be important in studying age discrimination.
- In the model of human capital investment, earnings become more dispersed as workers age, as differences in unobserved investment accumulate, which could generate a larger variance of unobservables for older versus younger applicants.
- To assess such bias in the context of age discrimination, we analyze data from a new, largescale field experiment.
- The study design lets us use a method developed in Neumark (2012) to identify the effect of age discrimination when the variance of unobservables can differ between groups.

II. Addressing the Heckman Critique

- We explain the analysis of data from AC studies, the Heckman critique, and how to address it.

- Assume that productivity depends linearly and additively on two characteristics: a measure $X\ I$ included on the resumes and standardized at $X\ I*$ across applicants in the study; and $X\ II$, which is unobserved by firms.

- Let S denote older ("senior") applicants and Y denote younger applicants.

- Define γ as an additional linear, additive term that reflects taste discrimination (undervaluation of productivity) or statistical discrimination (an assumption that $E(XS\ II) \neq E(XY\ II))$ regarding older workers—both
- of which are illegal in the United States.

- With the data from an AC study, we estimate γ from a model for callbacks as a linear function of $X\ I$ and an indicator for age.

- Suppose a callback results if a worker's perceived productivity exceeds a threshold $c(> 0)$.

- Then the hiring rules for older and younger applicants are

$$(1)\ T(X^{I^*}, X_S^{II})|(S = 1) = 1 \text{ if } \beta_1 X^{I^*} + X_S^{II} + \gamma > C$$

$$(1')T\left(X^{I^*}, X_Y^{II}\right)|(S = 0) = 1 \text{ if } \beta_1 X^{I^*} + X_Y^{II} > C,$$

where XS II or XY II are the residuals.

- If X_S^{II} and X_Y^{II} are normally distributed, with zero means, standard deviations $\sigma_\mathrm{S}^\mathrm{II}$ and $\sigma_\mathrm{Y}^\mathrm{II}$, and distribution function Φ, the callback probabilities are

$$(2)\ S = 1\colon \Phi[(\beta_1 X^{I^*} + \gamma - C)/\sigma_S^{II}]$$

$$(2')\ S = 0\colon \Phi[(\beta_1 X^{I^*} - C)/\sigma^{II}\_Y].$$

- Without a restriction on σ_S^{II} and σ_Y^{II} , γ is unidentified—the basis of Heckman's critique.

- If X_S^{II} and X_Y^{II} are normally distributed, with zero means, standard deviations $\sigma_{\mathrm{S}}^{\mathrm{II}}$ and $\sigma_{\mathrm{Y}}^{\mathrm{II}}$, and distribution function Φ, the callback probabilities are

(2) $S = 1\colon \Phi[(\beta_1 X^{I^*} + \gamma - C)/\sigma_S^{II}]$

(2′) $S = 0\colon \Phi[(\beta_1 X^{I^*} - C)/\sigma^{II}\_Y]$.

- Without a restriction on σ_S^{II} and σ_Y^{II} , γ is unidentified—the basis of Heckman's critique.

- This example shows that the relative variances of the unobservables interact with the level of quality chosen for the resumes in a correspondence study.

- Thus, without knowing how resume quality compares to those that employers receive, we cannot sign the bias even if we know whether σ_S^{II} is greater or less than σ_Y^{II}.

III. The Field Experiment

- The standard procedures for correspondence studies include: creation of data on artificial job applicants; applying for jobs; collection of data on hiring-related outcomes; and statistical analysis.

- The statistical analysis without quality variation in resumes is straightforward, and the extension to consider the Heckman critique follows the previous section.

- As described in Neumark, Burn, and Button (2015), we grounded the creation of resumes as much as possible in empirical evidence on actual resumes posted by job seekers.

- We created job applicants aged 29–31, 49–51, and 64–66.

- Hiring of 64–66-year-olds is significant because policymakers are trying to induce working longer via Social Security reforms, and this is likely to require hiring in new jobs as older workers leave their main jobs for other jobs, for health or other reasons, before retiring.

- To explore the implications for the Heckman critique, we generated applicants of different skill levels for each job for which we apply.
- We chose quality- or skill-related items based on extensive reading of actual resumes.
- High-skill resumes can include a post-secondary degree (B.A. for sales and security guard applicants, and Associate of Arts for janitor applicants), while all low-skill resumes only list a high school diploma.
- High-skill resumes can also include computer skills of some kind (appropriate to the job), fluency in Spanish as a second language, and other occupation-specific skills, such as licensing and CPR for security jobs, and certification for janitor jobs.
- The skills section can include one of three volunteer activities (food bank, homeless shelter, or animal shelter).

- All low-skill resumes include two typos, and some high-skill resumes do not.
- Finally, some high-skill resumes include recent "employee of the month" awards.
- We randomly assign five of seven possible skill indicators to each high-skill resume.
- We assign all applicants to the same employer as either high skilled or low skilled, with 50 percent probability for each, so that random assignment of high-skill or low-skill resumes within a triplet does not dominate the effect of age.
- Other resume characteristics that are not supposed to affect hiring are randomized across resumes, as in other audit and correspondence studies.

IV. Results

- Table 1 reports callback rates and statistical tests of independence.
- In retail, the callback rate for older applicants was significantly lower— 14.7 percent versus 20.9 percent for young applicants ($p = 0.00$).
- For security jobs, callback rates were lower for older applicants—21.7 versus 24.3 percent—with a marginally significant difference ($p = 0.12$).
- There were far fewer ads for janitor jobs.
- The callback rate differential is similar to security, but the difference is not statistically significant.
- The evidence of age discrimination in hiring based on the raw data is not as strong as in past studies.
- However, these conclusions can be misleading because of the problem of differences in variances of the unobservables—our main focus to which we turn next.

Table 1: Callback Rates by Age

		Young (29–31)	Old (64–66)
Sales (N = 3,570)			
Callback (%)	No	79.11	85.30
	Yes	20.89	14.70
Test of independence (p-value)		0.00	
Security (N = 2,746)			
Callback (%)	No	75.72	78.26
	Yes	24.28	21.74
Test of independence (p-value)		0.12	
Janitors (N = 845)			
Callback (%)	No	67.92	70.38
	Yes	32.08	29.62
Test of independence (p-value)		0.48	

Notes: The p-values reported for the tests of independence are from Fisher's exact test (two-sided). For the janitor resumes, only older resumes with commensurate experience are used.

- Table 2 reports estimates of models that include a dummy variable for older applicants, control variables including the skill indicators, and interactions between the skills and the old indicator.

- The interactions are informative because, under the identifying assumption that the underlying coefficients for the two age groups are equal, differences between the probit coefficients by age are informative about differences in the variances of the unobservables.

- For example, if the unobserved variance is larger for older workers, then, if the main effect of the skill variable is positive, the estimated interaction should be negative and reduce the overall effect toward zero.

- For sales workers, the skill variables are relatively unsuccessful in predicting hiring.

- The only main effect with a t-statistic exceeding one is employee of the month, for which the estimated interaction is of the opposite sign and points to a diminished effect for older applicants, although there are also estimates pointing to a larger effect for older applicants (e.g., computer skills).

- Table 2 reports estimates of models that include a dummy variable for older applicants, control variables including the skill indicators, and interactions between the skills and the old indicator.

- The interactions are informative because, under the identifying assumption that the underlying coefficients for the two age groups are equal, differences between the probit coefficients by age are informative about differences in the variances of the unobservables.

- For example, if the unobserved variance is larger for older workers, then, if the main effect of the skill variable is positive, the estimated interaction should be negative and reduce the overall effect toward zero.

- For sales workers, the skill variables are relatively unsuccessful in predicting hiring.

- The only main effect with a t-statistic exceeding one is employee of the month, for which the estimated interaction is of the opposite sign and points to a diminished effect for older applicants, although there are also estimates pointing to a larger effect for older applicants (e.g., computer skills).

Table 2: Probit Estimates for Callbacks by Age, Old versus Young, Effects of Skills and Interactions of Old with Skills, Marginal Effects

	Sales	Security	Janitor
Old (64–66)	−0.062	−0.037	0.073
	(0.085)	(0.057)	(0.229)
Common skills			
Spanish	0.007	0.081*	−0.022
	(0.025)	(0.045)	(0.049)
Spanish × Old	−0.046	0.038	−0.083
	(0.032)	(0.060)	(0.117)
Grammar	−0.017	0.025	0.002
	(0.020)	(0.034)	(0.047)
Grammar × Old	0.041	−0.019	0.036
	(0.037)	(0.045)	(0.124)
College	0.008	0.023	0.129**
	(0.023)	(0.038)	(0.053)
College × Old	−0.007	0.003	−0.055
	(0.031)	(0.049)	(0.107)
Employee of the month	0.033	−0.071*	−0.061
	(0.028)	(0.036)	(0.045)
Employee of the month × Old	−0.017	0.024	0.171
	(0.034)	(0.053)	(0.118)

Notes: Marginal effects computed as the discrete change in the probability associated with the variables, evaluating other variables at their means. Standard errors are computed based on clustering at the resume level. Other controls include city, order of resume submission, and employed/ unemployed. All controls are interacted with "Old" so main effect of "Old" is not meaningful. See notes to Table 1. *** Significant at the 1 percent level. ** Significant at the percent level. * Significant at the 10 percent level.

Table 2: Probit Estimates for Callbacks by Age, Old versus Young, Effects of Skills and Interactions of Old with Skills, Marginal Effects, Cont'd

	Sales	Security	Janitor
Volunteer	−0.027	−0.019	−0.103**
	(0.024)	(0.039)	(0.048)
Volunteer × Old	0.053	−0.034	−0.042
	(0.040)	(0.051)	(0.102)
Occupation-specific skills	*1: computer, 2: customer service*	*1: CPR, 2: license*	*1: technical skills, 2: certificate*
Skill 1	0.001	−0.064*	0.135**
	(0.024)	(0.034)	(0.066)
Skill 1 × Old	0.034	0.111**	−0.102
	(0.039)	(0.060)	(0.091)
Skill 2	0.012	0.065*	−0.009
	(0.024)	(0.039)	(0.064)
Skill 2 × Old	0.008	−0.052	−0.053
	(0.036)	(0.044)	(0.109)
Observations	3,570	2,746	845

Notes: Marginal effects computed as the discrete change in the probability associated with the variables, evaluating other variables at their means. Standard errors are computed based on clustering at the resume level. Other controls include city, order of resume submission, and employed/ unemployed. All controls are interacted with "Old" so main effect of "Old" is not meaningful. See notes to Table 1. *** Significant at the 1 percent level. ** Significant at the percent level. * Significant at the 10 percent level.

- Table 3 turns to the heteroscedastic probit estimates that correct for bias from differences in the variances of unobservables. Panel

- A reports the marginal effects from the standard probit model for each specification and sample.

- These estimates show significant evidence of age discrimination only in sales jobs, although all of the point estimates are in this direction.

- The first row of panel B reports the overall effect from the heteroscedastic probit estimates, which are similar to the probit estimates.

- Next, we report the p-values from the overidentification test that the ratios of the skill coefficients between younger and older workers are equal across all of the skills.

- These p-values are uniformly high, indicating that we never reject the overidentifying restrictions.

Table 3: Heteroscedastic Probit Estimates for Callbacks by Age, Old versus Young (Corrects for Potential Biases from Difference in Variance of Unobservables)

	Sales	Security	Janitor
Panel A. Probit estimates			
Old (64–66, marginal)	−0.044***	−0.028	−0.032
	(0.012)	(0.017)	(0.037)
Panel B. Heteroscedastic probit estimates			
Old (marginal)	−0.049***	−0.022	−0.031
	(0.012)	(0.020)	(0.040)
Overidentification test: ratios of coefficients on skills for old relative to young are equal (*p*-value)	0.88	0.85	0.97
Standard deviation of unobservables, old/young	0.84	1.16	1.33
Test: ratio of standard deviations = 1 (*p*-value)	0.23	0.35	0.59
Old-level (marginal)	−0.005	−0.058*	−0.084
	(0.039)	(0.030)	(0.094)
Old-variance (marginal)	−0.043	0.036	0.053
	(0.040)	(0.035)	(0.086)
Observations	3,570	2,746	845

Notes: Marginal effects computed as the change in the probability associated with "Old," using the continuous approximation, evaluating other variables at their means; the continuous approximation yields an unambiguous decomposition of the heteroscedastic probit estimates. p-values are based on Wald tests. See notes to Tables 1 and 2. *** Significant at the 1 percent level. ** Significant at the 5 percent level. * Significant at the 10 percent level.

- We next report the ratio of the standard deviation of the unobservables for old relative to young applicants.
- For sales applicants, the estimated ratio of standard deviations is a bit below one (0.84)—in contrast to our conjecture, lower for older workers.
- The p-value for the test that the ratio equals one is above 0.1 (0.23), but still relatively low.
- The last two rows of the table decompose the heteroscedastic probit estimates.
- "Old-level" is the unbiased estimate of the effect of age.
- The estimated level effect is near zero (-0.005), and nearly all of the effect comes from the variance ("Old-variance")—interpreted as spurious evidence from the research design—although these estimates are imprecise. Note also that the lower variance for older male sales applicants would predict that the standard probit estimates would overstate discrimination if the resumes were on average low quality, which is what we find.

V. Conclusions

- Our evidence points to more ambiguous evidenc of age discrimination against older men than past research.
- For the three occupations we study—retail, security, and janitors—the point estimates always indicate age discrimination; but the standard evidence is only strongly significant for retail.
- Moreover, the analysis indicates that conclusions are sensitive to accounting for the Heckman critique, adding to the ambiguity.
- The strongest evidence of age discrimination—in retail sales—disappears completely once the estimate is corrected for the bias identified by this critique.
- For security and janitor jobs, in contrast, the evidence of discrimination strengthens, although it is significant—and only weakly—just for security jobs.