IN THE UNITED STATES DISTRICT COURT
FOR THE WESTERN DISTRICT OF KENTUCKY
LOUISVILLE DIVISION

Chelsey Nelson Photography LLC, and **Chelsey Nelson,**

Plaintiffs,

v.

Louisville/Jefferson County Metro Government; Louisville Metro Human Relations Commission–Enforcement; Louisville Metro Human Relations Commission–Advocacy; Verná Goatley, in her official capacity as Executive Director of the Louisville Metro Human Relations Commission–Enforcement; and **Marie Dever, Kevin Delahanty, Charles Lanier, Sr., Leslie Faust, William Sutter, Ibrahim Syed,** and **Leonard Thomas,** in their official capacities as members of the Louisville Metro Human Relations Commission–Enforcement,

Defendants.

Case No. 3:19-cv-00851-BJB-CHL

Plaintiffs' Reply in Support of Their Motion to Exclude Testimony of Netta Barak-Corren

TABLE OF CONTENTS

Table of Authorities

Cases

Rules

Statutes, Codes and Ordinances

Other Authorities

Introduction

Professor Netta Barak-Corren's conclusions are speculative, her methods are unreliable, and her findings are irrelevant. *See* Pls.' Mot. to Exclude Testimony of Netta Barak-Corren (MTE), ECF No. 90. Her testimony should be excluded.

More fundamentally, Barak-Corren's conclusions are troubling. According to her, courts should tolerate religious hostility and reject religious exemptions because the media *might* incorrectly describe court decisions, which in turn *might* cause creative professionals to exercise their constitutional rights (what Barak-Corren calls discrimination). Barak-Corren makes this conclusion with no evidence about what media creative professionals consume or how they react to which media and without using any reliable methodology. And then she applies her conclusion to this litigation with no evidence about Louisville, Louisville media, Louisville court decisions, or Louisville creative professionals—all the while overlooking the evidence *from her own study* that indicates that Louisville won't be affected by the so-called *Masterpiece* effect.

The unspoken assumption beneath Barak-Corren's testimony is that courts should consider how the media report on court decisions to avoid problematic results that may or may not occur because of those reports. But neither the media nor the masses dictate fundamental freedoms. The Bill of Rights "withdraw[s] certain subjects from the vicissitudes of political controversy, to place them beyond the reach of majorities and officials and to establish them as legal principles to be applied by the courts." *W. Va. State Bd. of Educ. v. Barnette*, 319 U.S. 624, 638 (1943). Barak-Corren's testimony should be excluded.

Argument

Barak-Corren's testimony should be excluded because it is (I) unreliable and (II) irrelevant.

I. Barak-Corren's testimony should be excluded as unreliable.

Barak-Corren's testimony should be excluded because she does not use "reliable principles and methods," has insufficient "facts or data," and does not "reliably appl[y] the principles and methods to" this case. Fed. R. Evid. 702(b)–(d).

Contrary to Louisville's claim, Chelsey does not challenge Barak-Corren's testimony because of imperfect or weak methodology. Defs.' Resp. to Pls.' Mot. to Exclude Testimony of Netta Barak-Corren (RMTE) 6–7, ECF No. 99. Her testimony is unreliable. That's Rule 702's purpose—to weed out unreliable testimony. *See Tamraz v. Lincoln Elec. Co.*, 620 F.3d 665, 675 (6th Cir. 2010) (experts much reach conclusions "via a sound methodology."); Fed. R. Evid. 702 advisory committee's note, 2000 amend. ("[A]ny step that renders the analysis unreliable ... renders the expert's testimony inadmissible.").

Though Louisville disagrees, this purpose applies to "a bench trial." RMTE 6. Rule 702 has no bench-trial exception. Louisville's cases don't say otherwise. *Cf. Deal v. Hamilton Cnty. Bd. of Educ.*, 392 F.3d 840, 851 (6th Cir. 2004) (opposers "provide[d] no legal arguments" for to exclude expert); *Douglas v. United States*, 2011 WL 2633612, at *6 (E.D. Ky. July 5, 2011) ("[E]xperts must pass *Daubert* scrutiny."). And courts nix unreliable experts at summary-judgment. *See, e.g.*, *Newell Rubbermaid, Inc. v. Raymond Corp.*, 676 F.3d 521, 526–29 (6th Cir. 2012).

This Court should likewise exclude Barak-Corren's unreliable testimony that (A) speculates about media; (B) reaches unreliable results with unreliable methods; (C) misapplies legal regimes; and (D) has no facts about Louisville.

A. Barak-Corren's testimony is unreliable because she depends on speculative assumptions about the media.

The *Masterpiece* Study depends on what creative professionals understood about *Masterpiece* "as mediated through and filtered by the media." RMTE Ex. 2 (Tr.) 189, ECF No. 99–2. But Barak-Corren's testimony speculates about

professionals' exposure to media reports about the *Masterpiece* decision. That speculation makes her testimony unreliable.

Louisville admits "it was not possible to examine what media sources" professionals saw after the *Masterpiece* decision. RMTE 8. Barak-Corren cannot measure post-*Masterpiece* effects without knowing if professionals knew about *Masterpiece* in the first place. MTE 5–6 (making same point). So her conclusions without that knowledge are unreliable.

Louisville cites two studies to cover up this problem. RMTE 8.[1] But neither fill the holes Louisville needs. In the first study, Tankard and Paluck knew about participants' media consumption because they "reported consuming mass media news about 3 to 4 times per week, with a 57% majority reporting daily consumption, which suggests that few participants were unaware of the ruling." 28 Psych. Sci. at 1339. Barak-Corren relies on this study to conclude that professionals *could* learn of the decision without media *if* the decision "spread[s] in society and shape[s] social views." Tr. 93:7–9. This is speculative. Unlike Tankard and Paluck, she has no evidence about whether professionals followed media or whether society generally knew about *Masterpiece*.

Likewise, Linos and Twist measured survey subjects' "exposure" to news coverage by asking them "whether they had heard about a series of seven prominent news stories around the time of the decisions." 45 J. Legal Stud. at 240. Then they divided the subjects between "the treated group"—those with some level of exposure—from the untreated group. *Id.* Throughout, they stressed

[1] Margaret E. Tankard & Elizabeth Levy Paluck, *The Effect of a Supreme Court Decision Regarding Gay Marriage on Social Norms and Personal Attitudes* (Tankard & Paluck), 28 Psych. Sci. 1334 (2017); Katerina Linos & Kimberly Twist, *The Supreme Court, the Media, and Public Opinion: Comparing Experimental and Observational Methods* (Linos & Twist), 45 J. Legal Stud. 223 (2016).

"distinguish[ing] people who heard and understood the decisions from those who did not." *Id.* at 239. Barak-Corren did none of this.

Besides no evidence of media exposure, Barak-Corren also cannot measure what specific media professionals saw (if any). MTE 7–8. Louisville re-frames the argument as "media bias" and claims any media bias is "irrelevant" because the *Masterpiece* Study doesn't "measur[e] the impact of any media bias but rather the effect of the decision itself." RMTE 8. Louisville's argument misses a few key steps.

Recall that Barak-Corren relies on the media to communicate about *Masterpiece*. MTE Ex. B (HCRCL) 24–27, 47–48 n.150, ECF No. 90–3. And professionals' understanding of *Masterpiece* is filtered through the media. Tr. 189. So what the media says about *Masterpiece* to the public is critical. And what the media says about *Masterpiece* depends on whether the media source is mainstream, progressive, or conservative. "[M]ainstream" and "progressive" outlets classified *Masterpiece* narrowly or critically while "conservative" media supposedly had "less reservations about its scope." HCRCL 25–27.

Barak-Corren's failure to match professionals with media sources makes her testimony unreliable. If professionals saw only mainstream or progressive media, they would have understood the decision narrowly or critically. If professionals saw only conservative media, then their *Masterpiece* knowledge would have depended on what type of conservative media they saw (because those reports varied). MTE 7–8.

This failure also undermines Barak-Corren's conclusion that *Masterpiece* led to "[c]hanges in social norm perceptions." HCRCL 48. This theory assumes "that an individual's behavior depends … on what actions she believes others will approve or disapprove." Richard H. McAdams, *An Attitudinal Theory of Expressive Law*, 79 Or. L. Rev. 339, 340 (2000). HCRCL 47 n. 149 (citing same). Professionals who saw only mainstream media—where the decision was explained as "narrow" and as "not resolv[ing] the big constitutional questions at issue"—or only progressive media—

who reported on the case as a "license to discriminate"—would have no reason to think that *Masterpiece* ushered in a new social norm. HCRCL 25–27. Even those who viewed only "conservative" media would not have reached that conclusion because "conservative" reports on *Masterpiece* varied. MTE 8.

And professionals who saw nuanced conservative coverage or a mix of mainstream, progressive, or conservative coverage saw "two-sided coverage,"—i.e. coverage with both "supportive and critical information"—which makes it unlikely they would change their opinion about providing services for same-sex weddings Linos & Twist, 45 J. Legal Stud. at 225–26. That's because two–sided coverage "reduce[s] the impact of the Court decision on opinion change." *Id.* Louisville's contrary argument about two-sided coverage misstates the study.[2] RMTE 8.

Louisville claims that these "nuances are irrelevant because the *Masterpiece* decision was broadly reported and understood as a victory for a baker." RMTE 9. That claim points up the problem. There's no evidence about what professionals "understood" about the decision through media because Barak-Corren never measured that. And the media that Barak-Corren cites—mainstream, progressive, and conservative—demonstrate that the decision was not "broadly reported" as a victory. Indeed, most often, media called the decision narrow or criticized it. HCRCL 25–27. MTE Ex. E (GY Report) ¶¶ 30–31, ECF No. 90–6.

Finally, Louisville says that "[t]he fact that Professor Barak-Corren still found a significant *Masterpiece* effect" even with narrow and critical media reports "demonstrates the significant impact of the ruling." RMTE 9. But that assumes,

---

[2] *See, e.g.*, Linos & Twist, 45 J. Legal Stud. at 226 ("Two-sided coverage, … reduce[d] the impact of the Court decision on opinion change."); *id.* at 230 ("[I]ndividuals who receive two-sided, competing frames are more likely to retain their original views."); *id.* at 232 ("[P]eople who receive two-sided information should respond less positively; indeed, the net effect might be zero or even negative, depending on the relative strength of the competing frames."); *id.* at 242 ("[O]ne-sided coverage produces larger effects than does two-sided coverage.").

without evidence, the conclusion—that publicity about *Masterpiece* caused professionals to change behaviors. Assumptions without facts are unreliable. MTE 9–10 (collecting Sixth Circuit cases).

Even beyond that, Louisville never addresses the many other problems with Barak-Corren's dependence on media. For example, her study cannot be replicated and Barak-Corren goes against the grain of the generally accepted practice of studying media, the Supreme Court, and public opinion. *Id.* at 8–9. These problems independently justify excluding Barak-Corren's testimony.

B. Barak-Corren's methodological mistakes lead to unreliable conclusions.

The *Masterpiece* Study's methodological errors make it unreliable. For starters, Barak-Corren cannot measure pre-*Masterpiece* discrimination because of attrition rates in responses between Waves 1 and 2. MTE 12. Louisville argues Barak-Corren solved this problem because she "tested her conclusion against different data sets": (1) a control group of professionals; (2) businesses that indicated willingness to serve both types of couples pre-*Masterpiece*; (3) businesses that indicated willingness to serve same-sex couples pre-*Masterpiece*; and (4) within business transitions before and after *Masterpiece*. RMTE 12. Louisville is incorrect.

The control group is a misnomer. Normally, control groups and experimental groups are the same "except that the experimental groups are subjected to treatments … believed to have an effect on the outcome of interest while the control group is not." Mary Earick Godby, *Control Group*, Britannica, https://bit.ly/3nf2ACN. Here, the "treatment" was supposedly exposure to the *Masterpiece* decision. But Barak-Corren does not examine the "control group's" knowledge of *Masterpiece*. So the control group isn't really controlling for anything. It's just a group of professionals contacted after *Masterpiece*.

The final three data sets suffer from the regression fallacy because they depend on the unusually high positive response rate to same-sex requests in Wave 1. HCRCL 36 (70.8% Wave 1 response rate). MTE 13–14. Professionals who positively responded to same-sex wedding requests pre-*Masterpiece* will regress to their average same-sex responsiveness post-*Masterpiece*. *Id.* This logic applies to professionals who responded positively to same-sex *and* opposite-sex couples pre-*Masterpiece*. Those professionals' responsiveness to same-sex requests will regress towards the mean. In both cases, Barak-Corren only subsets professionals who responded favorably to same-sex requests before *Masterpiece*. By definition, these response rates to same-sex inquiries after *Masterpiece* had nowhere to go but down.

Likewise, the regression fallacy applies to within business transitions. Barak-Corren measured the percent of professionals transitioning "from no/negative response pre-*Masterpiece* to positive response post-*Masterpiece*, and vice-versa, for same-sex and opposite-sex couples." MTE Ex. C (JLS) 29–31, ECF No. 90–4. Barak-Corren posits that "same-sex couples were twice as likely to experience a negative transition, such that a previously willing business would decline to provide service post-*Masterpiece*." *Id.* at 30. But again, Wave 1 professionals had atypically high positive responses to same-sex requests pre-*Masterpiece*. So any business transition from agreeing pre-*Masterpiece* to declining post-*Masterpiece* likely underwent regression. Statistics—not discrimination—accounts for the responsiveness change.[3]

[3] Louisville also denies the regression fallacy because Barak-Corren compared responses for opposite-sex and same-sex couples in Waves 3 and 4. RMTE 13. According to Louisville, "[a]bsent the *Masterpiece* effect, the response rates would have been the same for both couple types." *Id.* But there's no evidence for that—Louisville just assumes that all professionals respond identically to all inquiries. Barak-Corren doesn't even agree with that. MTE Ex. D (App.) 1–10, ECF No. 90–5 (explaining photographers are "pickier in general about their customers …").

Take a sports analogy. There's a theory that athletes decline between the first and second years of their careers—i.e., the "Sophomore Slump." Debra Wetcher-Hendricks, *Does the Sophomore Slump Really Exist?,* Theory in Action 7(3) (2014) (attached as Ex. S). A professor tested this theory by analyzing baseball careers of Rookie of the Year (ROY) award recipients. *Id.* at 65–68. She found that after outstanding rookie seasons, these players' performance dipped in their second year and then remained consistent for the rest of their careers. *Id.* at 67. This led to two conclusions. First, ROY players regress to their average performance after their first-year high-level performance. *Id.* at 68. Next, "[s]econd-year performance … is not unusually low" but "a Freshman Fluke, characterized by a comparatively good performance during the players' first years in the Major Leagues, exists." *Id.*

The Freshman Fluke. That's Barak-Corren's Wave 1. Like the unusually successful rookie seasons for ROY players, Wave 1 contained an unusually high positive-response rate to same-sex requests relative to opposite-sex requests. All of Barak-Corren's post-*Masterpiece* analysis (and the so-called *Masterpiece* effect) depends on and is measured against Wave 1. But Barak-Corren never considers the possibility that any changes in post-*Masterpiece* responsiveness occurred because of a statistical correction. To continue the sports analogy, "a true regression to the mean situation suggests that the change in performance [after the first year] reflects statistical, rather than an athletic phenomenon." MTE Ex. S at 68. Applied here, the responsiveness change could result from regression to the mean rather than discrimination. But Barak-Corren just assumes discrimination. MTE 14.

Barak-Corren's testimony strikes out in other ways too. For example, Barak-Corren significantly altered the content of the emails she sent to professionals in each wave. MTE 14–16. Louisville cites a study by to remedy this problem.[4] RMTE

[4] Marianne Bertrand & Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market*

MTE 14–15. But that study doesn't help. That study measured racial bias in the labor market by submitting nearly identical resumes to employers while assigning half of the resumes "White-sounding names" and the other half "African-American-sounding names." Bertrand & Mullainathan, 94 Am. Econ. Rev. at 992; *id.* at 1006 (conclusions based on "identical individuals" with different names). The study "randomly assigned" a race "to each resume" to "guarantee[] that any differences [in employer callbacks] … are caused solely by the race manipulation." *Id.* at 994. By contrast, Barak-Corren did not randomize her emails—i.e., she always sent one email from same-sex couples and a different email from opposite-sex couples. App. 1–10. In sum, Barak-Corren altered the content of the emails systematically (i.e., same-sex and opposite-sex requests) and did not randomize them, so the change in response rate could be because of the different content. GY Report ¶¶ 11–14.

Louisville doubles down by arguing these changes don't alter the *Masterpiece* Study because Barak-Corren measured professionals' responses "to identical inquiries from same- and opposite-sex couples after the *Masterpiece* decision." RMTE 15. But this measurement is meaningless because Barak-Corren cannot know "the extent of discrimination towards same-sex couples … before *Masterpiece*." Tr. 147:12–21, 156:7–23. Without knowledge of pre-*Masterpiece* refusals, Barak-Corren cannot claim that refusals increased post-*Masterpiece*.[5] And any equal distribution "across couple types" or any differences in "non-responses" just confirms the regression to the mean analysis—i.e., that post-*Masterpiece* professionals responded based on their average responsiveness. RMTE 14–15.

Discrimination (Bertrand & Mullainathan), 94 Am. Econ. Rev. 991 (2004) (attached as Exhibit R)).

[5] For these same reasons, Louisville is wrong to suggest that Barak-Corren's conclusions do "not depend on any comparison to pre-*Masterpiece* responses." RMTE 13. The study evaluates the effects of religious exemptions on professionals' willingness to provide services for same-sex weddings as judged by their pre-and-post-*Masterpiece* responses to same-sex inquiries. *See, e.g.*, HCRCL 1, 24.

Another problem with Barak-Corren's testimony is that she codes non-responses as negative responses. MTE 14–17. Louisville claims this doesn't matter because Barak-Corren evaluated "systemic differences" between same-sex and opposite-sex inquiries "not isolated instances of non-responses." RMTE 13. Even so, there were many variables that could have explained systematic differences in responses. MTE 14–17. Barak-Corren never addresses them.

C. Barak-Corren's testimony is unreliable because she misclassifies and misapplies legal regimes.

Next, Barak-Corren's testimony is unreliable because she tries to distinguish Louisville from Texas's +RFRA/+AD jurisdictions based on a legal error and mischaracterizes other jurisdiction's legal regimes.

In Texas, Barak-Corren found that +RFRA/+AD jurisdictions were immune from the *Masterpiece* Study. MTE 10– 11. She does not extend this conclusion to Louisville (even though it is also a +RFRA/+ AD jurisdiction) because she believes Kentucky's Religious Freedom Restoration Act differs from Texas's. *Id.* That was a mistake—the laws' text is basically the same. *Compare* K.R.S. § 446.350 *with* Tex. Civ. Prac. & Rem. Code Ann. § 110.003(a). While Texas's RFRA cannot be used as a defense "to a civil action … under a federal or state civil rights law," this exemption doesn't apply to local civil rights laws. Tex. Civ. Prac. & Rem. Code § 110.011(a). So both RFRAs provide defenses to local anti-discrimination laws. MTE 10–11.

Louisville tries to avoid this conclusion, first saying that Texas's "carve-out" appears "on the face of the language in Texas['s] … RFRA." RMTE 11. But Barak-Corren never explains how this limited carve-out for federal and state law distinguishes Texas's RFRA's from Kentucky's. Next, Louisville says that neither caselaw nor Austin's defense of its local antidiscrimination law show Texas's RFRA protects against local antidiscrimination laws. RMTE 10 n.1. But the text of Texas's RFRA speaks for itself. Tex. Civ. Prac. & Rem. Code § 110.011(a). Other authorities

agree. MTE 11. Louisville finally hedges her conclusions as "preliminary." RMTE 11. But Barak-Corren unequivocally testified that "the report would have looked very differently" if Chelsey lived in Austin. Tr. 172:21–24. That's decisive.

Louisville also never addresses how Barak-Corren's misclassified legal regimes. MTE 11 n.8. Barak-Corren felt using varied legal regimes was "necessary" to uncover "real-world variation." HCRCL 32. That Barak-Corren botched a pivotal part of her study shows her testimony rests on unreliable foundations.

D. Barak-Corren's testimony is unreliable because she has no facts about how her study would apply in Louisville.

Barak-Corren's testimony is also unreliable because she cannot compare Louisville to any of the studied states and she has no evidence about Louisville's creative professionals, how her study would apply to a district court opinion, or how her study would apply today. MTE 17–22.

Start with comparing Louisville to the studied states. Barak-Corren tried to link *Louisville* to the studied states by comparing *Kentucky*'s demographics with the four studied states. MTE Ex. A (NBC Report) ¶¶ 17–23, ECF No. 90–2. Aside from religiosity, Barak-Corren made no attempt to compare Louisville's demographics to the studied states. *Id.* So there's no evidence that Louisville shares any similarities with any state in the *Masterpiece* Study. And there's no basis for Barak-Corren to assume that Louisville parallels Kentucky. MTE 18–19 (noting different attitudes and political affiliations).

As for Louisville's religiosity, Barak-Corren's "evidence" was a Wikipedia page and three websites. NBC Report ¶ 20 nn. 2–3. Those are not reliable sources. MTE 19 (citing cases). And Louisville concedes Barak-Corren had no data on religious density in Louisville "as compared to Kentucky more generally." RMTE 16. While Barak-Corren may have had information about religious density about Kentucky (*id.* at 16), Louisville is still not Kentucky. Without religiosity evidence,

Barak-Corren's testimony is unreliable. Her conclusion depends on Louisville's religiosity—she "expect[s] to observe the *Masterpiece* effect in Louisville" because of "the high degree of religiosity in the area." NBC Report ¶ 23. Barak-Corren has no evidence about Louisville's religious density. So that conclusion has no support.

Next, Barak-Corren never audited professionals in Louisville. Louisville does not dispute this. So Barak-Corren cannot know if professionals in Louisville would react the same to the *Masterpiece* Study as a florist in Floyd, Iowa, a photographer in Plano, Texas, or a baker in Bakersville, North Carolina.

Barak-Corren also has no evidence about how the *Masterpiece* Study would apply to a favorable ruling from this Court. Barak-Corren has no information or study on media and district court decisions and never claims district court and Supreme Court opinions attract the same media attention. MTE 16-17. Barak-Corren has no facts about how anyone would react to a district court decision.

Even so, Louisville cites several local articles and a press release to argue that media "should certainly be expected to publicize the Court's ultimate decision." RMTE 18. This response has two problems. First, Barak-Corren never mentioned or considered these articles. Louisville cannot add evidence to Barak-Corren's testimony after-the-fact. Second, Barak-Corren had no methodology for choosing which media sources supposedly communicate to the public about judicial opinions. MTE 8–9 (making this point). So Louisville cannot even test its own theory. It is impossible to know if Barak-Corren would find the local articles to be relevant news sources, much less Louisville creatives. HCRCL 25–27.

Finally, Barak-Corren has no evidence about how the *Masterpiece* Study might apply to present-day Louisville. MTE 19–20. Louisville counters that the *Masterpiece* effect harms "same-sex couples even if it is short in duration." RMTE 19. That misses the point. The so-called *Masterpiece* effects are not just short-lived. The point is that there's no evidence that any effect would happen in Louisville

today were the study re-run.[6] GY Report ¶ 21. Is the present-day social environment conducive to or immune from the *Masterpiece* Study? No one knows.

Louisville next argues that the lack of complaints "could be caused by numerous factors." RMTE 18. Yes, present-day factors are relevant. And they're relevant to the wedding industry and professionals too. But Barak-Corren does not account for any of these factors. For example, she does not "know what COVID did to weddings" and agrees "maybe the [wedding] field has changed." Tr. 113:6–13. These contemporary nuances prove that Barak-Corren has no evidence about how the *Masterpiece* Study would apply today. That makes her conclusions unreliable.

Barak-Corren's total lack-of-evidence distinguishes her testimony from the other cases Louisville cites where plaintiffs challenged less-than "perfect methodology," critiqued research designs, or extrapolated from some data. RMTE 6–7, 15–16. Barak-Corren's testimony has no data about Louisville and is unreliable.

II. Barak-Corren's testimony should be excluded as irrelevant.

Barak-Corren's testimony should also be excluded because it is irrelevant to Chelsey's free-speech claim and analyzing her claims under strict scrutiny.

As for Chelsey's free-speech claim, Chelsey claims that the First Amendment protects her from creating photographs and writing blogs celebrating messages she opposes. Pls.' Br. in Supp. of Their Summ. J. Mot. 6–15, ECF No. 92–1. And there's no dispute that Chelsey's photographs and blogs are speech. *Id.* at 6–7.

The *Masterpiece* Study is irrelevant to this claim because it only measured the effect of media reports on religious exemptions, as Louisville admits. Tr. 188:7–189:25; RMTE 20. Barak-Corren does not claim that her study applies to free-

[6] Louisville claims that "the impact of judicial decisions is prolonged and significant." RMTE 18–19. The cited studies are irrelevant because they either dealt with legislation or proved any effect only lasted up to 18 months. RMTE 19. And Barak-Corren admitted no study addressed effects on professionals. MTE 20.

speech exemptions. NBC Report ¶ 12 (granting Chelsey "a religious exemption from the application of Louisville's" law could increase discrimination).

Louisville counters that Barak-Corren "*had no assumptions* about" her potential findings. RMTE 19. That may be. But it is irrelevant. The *conclusion*—that all negative or non-responses were discriminatory—*assumes* that professionals declined same-sex wedding requests post-*Masterpiece* due to discrimination rather than because of a legitimate, constitutionally protected, message-based objection to celebrating same-sex marriage. Those assumptions make her conclusions about religious exemptions irrelevant to Chelsey's free-speech claim. MTE 23.

As for strict scrutiny, Louisville claims that Chelsey's argument is "inconsisten[t]"—that she cannot say strict scrutiny demands evidence but argue that "empirical evidence" about exempting Chelsey is "irrelevant." RMTE 20. But that's the rub. Barak-Corren has no evidence—empirical or otherwise—about how the *Masterpiece* Study applies here. She never audited Louisville's professionals, studied Louisville's religiosity, tested the public's interactions with media or district court opinions, or claims that religious-exemptions are the same as message-based objections. *See* MTE 20–21, 24–25. Barak-Corren also disclaims being an expert in current wedding markets. Tr. 113:1–13. And only .00051% (1/1,977) of professionals explicitly declined to celebrate same-sex weddings for conscience reasons. MTE 25.

Louisville counters that Chelsey puts "an undue emphasis on testimony from HRC witnesses" about the lack of complaints because their testimony "was not based on any rigorous or comprehensive analysis." RMTE 18. But Louisville—not Chelsey—must show an actual problem exists. *Brown v. Ent. Merchants Ass'n*, 564 U.S. 786, 799–800 (2011) (state "bears risk of uncertainty," not speakers). Barak-Corren's testimony is irrelevant to that showing because she identifies no problem in Louisville and Louisville identifies no reason to doubt its own officials' conclusion that the injunction has caused no issue in Louisville.

Besides being irrelevant, consider some consequences of Barak-Corren's testimony. Pretend Barak-Corren performed the same study, but studied parade organizers instead. Call it the *Hurley* Study. Pretend further that the *Hurley* Study found parades more often excluded pro-LGBT messages from their parades after a Supreme Court decision than before. Would that justify compelling all parade organizers to include pro-LGBT messages? Of course not. *See Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos.*, 515 U.S. 557, 572–73 (1995) (parade had First Amendment right to reject units that would "alter the expressive content of their parade"). Parades have a constitutional right to decide for themselves the messages they want to promote. Barak-Corren, though, relabels constitutional freedoms as discrimination and Louisville equates constitutional "message-based objections" with "discriminatory animus." RMTE 20. Barak-Corren and Louisville are wrong.

Take another example. *Masterpiece* found that Colorado treated Jack Phillips with "a clear and impermissible hostility toward" his religious beliefs. *Masterpiece Cakeshop, Ltd. v. Colorado C.R. Comm'n*, 138 S. Ct. 1719, 1729 (2018). Under Barak-Corren's theory, the Supreme Court should have ruled against Jack Phillips—and tolerated Colorado's hostility—because of the speculative possibility that the media would misreport the case, the public would misunderstand the ruling, and professionals would discriminate more based on their incorrect understanding of a court opinion. GY Report ¶ 39. That cannot be right.

In the end, Barak-Corren's testimony is irrelevant to Chelsey's claims and to strict scrutiny. Her testimony should be excluded.

Conclusion

Professor Netta Barak-Corren's testimony, including her report, her written articles, and any additional testimony she may provide should be excluded.

Respectfully submitted this 25th day of October, 2021.

By: s/ Bryan D. Neihart

Ryan Bangert
TX Bar No. 24045446*
Jonathan A. Scruggs
AZ Bar No. 030505*
Katherine L. Anderson
AZ Bar No. 033104*
Bryan D. Neihart
AZ Bar No. 035937*
Alliance Defending Freedom
15100 N. 90th Street
Scottsdale, AZ 85260
Telephone: (480) 444–0020
rbangert@adflegal.org
jscruggs@adflegal.org
kanderson@adflegal.org
bneihart@adflegal.org

Hailey M. Vrdolyak
IL Bar No. 6333515*
Alliance Defending Freedom
440 First Street NW, Suite 600
Washington, DC 20001
Telephone: (202) 393–8690
hvrdolyak@adflegal.org

David A. Cortman
GA Bar No. 188810*
Alliance Defending Freedom
1000 Hurricane Shoals Road NE
Suite D–1100
Lawrenceville, GA 30043
Telephone: (770) 339–0774
dcortman@adflegal.org

Joshua D. Hershberger
KY Bar No. 94421
Hershberger Law Office
P.O. Box 233
Hanover, IN 47243
Telephone: (812) 274–0441
josh@hlo.legal

Attorneys for Plaintiffs

* Admission *Pro Hac Vice*

Attorneys for Plaintiffs

CERTIFICATE OF SERVICE

I hereby certify that on the 25th day of October, 2021, I electronically filed the foregoing document with the Clerk of Court using the ECF system which will send notification of such filing to all counsel of record who are registered users of the ECF system.

By: s/ Bryan D. Neihart

Bryan D. Neihart
AZ Bar No. 035937*
Alliance Defending Freedom
15100 N. 90th Street
Scottsdale, AZ 85260
Telephone: 480–444–0020
bneihart@ADFlegal.org

Attorney for Plaintiffs
* Admitted *Pro Hac Vice*

UNITED STATES DISTRICT COURT
WESTERN DISTRICT OF KENTUCKY
LOUISVILLE DIVISION

Chelsey Nelson Photography LLC and **Chelsey Nelson,** Plaintiffs, **v.** **Louisville/Jefferson County Metro Government; Louisville Metro Human Relations Commission-Enforcement; Louisville Metro Human Relations Commission-Advocacy; Verná Goatley,** in her official capacity as Executive Director of the Louisville Metro Human Relations Commission-Enforcement; and **Marie Dever, Kevin Delahanty, Charles Lanier, Sr., Leslie Faust, William Sutter, Ibrahim Syed,** and **Leonard Thomas,** in their official capacities as members of the Louisville Metro Human Relations Commission-Enforcement, Defendants.	**Case No. 3:19-cv-00851-BJB-CHL** **Bryan D. Neihart's Supplemental Declaration in Support of Plaintiffs' Motion to Exclude Testimony of Netta Barak-Corren**

I, Bryan D. Neihart, declare as follows:

1. I am over the age of eighteen and competent to testify, and I make this declaration based on my personal knowledge.

2. I am one of the attorneys representing Plaintiffs Chelsey Nelson Photography LLC and Chelsey Nelson in this litigation.

3. Defendants' Response to Plaintiffs' Motion to Exclude Testimony of Netta Barak-Corren (ECF No. 99) cites a study by Marianne Bertrand and Sendhil

Mullainathan entitled *Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination*, 94 Am. Econ. Rev. 991 (2004). A true and correct copy of that article as accessed at https://www.uh.edu/~adkugler/Bertrand&Mullainathan.pdf is attached as Exhibit R.

4. A true and correct copy of Debra Wetcher-Hendricks, *Does the Sophomore Slump Really Exist?* Theory in Action 7(3) (2014) is attached as Exhibit S.

Declaration Under Penalty of Perjury

I, Bryan D. Neihart, a citizen of the United States and a resident of the State of Arizona, hereby declare under penalty of perjury pursuant to 28 U.S.C. § 1746 that the foregoing is true and correct to the best of my knowledge.

Executed this 25th day of October, 2021, at Scottsdale, Arizona.

Bryan D. Neihart

EXHIBIT R

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

By MARIANNE BERTRAND AND SENDHIL MULLAINATHAN*

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. White names receive 50 percent more callbacks for interviews. Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U.S. labor market. (*JEL* J71, J64).

Every measure of economic success reveals significant racial inequality in the U.S. labor market. Compared to Whites, African-Americans are twice as likely to be unemployed and earn nearly 25 percent less when they are employed (Council of Economic Advisers, 1998). This inequality has sparked a debate as to whether employers treat members of different races differentially. When faced with observably similar African-American and White applicants, do they favor the White one? Some argue yes, citing either employer prejudice or employer perception that race signals lower productivity. Others argue that differential treatment by race is a relic of the past, eliminated by some combination of employer enlightenment, affirmative action programs and the profit-maximization motive. In fact, many in this latter camp even feel that stringent enforcement of affirmative action programs has produced an environment of reverse discrimination. They would argue that faced with identical candidates, employers might favor the African-American one.[1] Data limitations make it difficult to empirically test these views. Since researchers possess far less data than employers do, White and African-American workers that appear similar to researchers may look very different to employers. So any racial difference in labor market outcomes could just as easily be attributed to differences that are observable to employers but unobservable to researchers.

To circumvent this difficulty, we conduct a field experiment that builds on the correspondence testing methodology that has been primarily used in the past to study minority outcomes in the United Kingdom.[2] We send resumes in response to help-wanted ads in Chicago and Boston newspapers and measure callback for interview for each sent resume. We

* Bertrand: Graduate School of Business, University of Chicago, 1101 E. 58th Street, R0 229D, Chicago, IL 60637, NBER, and CEPR (e-mail: marianne.bertrand@gsb.uchicago.edu); Mullainathan: Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, E52-380a, Cambridge, MA 02142, and NBER (e-mail: mullain@mit.edu). David Abrams, Victoria Bede, Simone Berkowitz, Hong Chung, Almudena Fernandez, Mary Anne Guediguian, Christine Jaw, Richa Maheswari, Beverley Martis, Alison Tisza, Grant Whitehorn, and Christine Yee provided excellent research assistance. We are also grateful to numerous colleagues and seminar participants for very helpful comments.

[1] This camp often explains the poor performance of African-Americans in terms of supply factors. If African-Americans lack many basic skills entering the labor market, then they will perform worse, even with parity or favoritism in hiring.

[2] See Roger Jowell and Patricia Prescott-Clarke (1970), Jim Hubbuck and Simon Carter (1980), Colin Brown and Pat Gay (1985), and Peter A. Riach and Judith Rich (1991). One caveat is that some of these studies fail to fully match skills between minority and nonminority resumes. For example some impose differential education background by racial origin. Doris Weichselbaumer (2003, 2004) studies the impact of sex-stereotypes and sexual orientation. Richard E. Nisbett and Dov Cohen (1996) perform a related field experiment to study how employers' response to a criminal past varies between the North and the South in the United States.

experimentally manipulate perception of race via the name of the fictitious job applicant. We randomly assign very White-sounding names (such as Emily Walsh or Greg Baker) to half the resumes and very African-American-sounding names (such as Lakisha Washington or Jamal Jones) to the other half. Because we are also interested in how credentials affect the racial gap in callback, we experimentally vary the quality of the resumes used in response to a given ad. Higher-quality applicants have on average a little more labor market experience and fewer holes in their employment history; they are also more likely to have an e-mail address, have completed some certification degree, possess foreign language skills, or have been awarded some honors.[3] In practice, we typically send four resumes in response to each ad: two higher-quality and two lower-quality ones. We randomly assign to one of the higher- and one of the lower-quality resumes an African-American-sounding name. In total, we respond to over 1,300 employment ads in the sales, administrative support, clerical, and customer services job categories and send nearly 5,000 resumes. The ads we respond to cover a large spectrum of job quality, from cashier work at retail establishments and clerical work in a mail room, to office and sales management positions.

We find large racial differences in callback rates.[4] Applicants with White names need to send about 10 resumes to get one callback whereas applicants with African-American names need to send about 15 resumes. This 50-percent gap in callback is statistically significant. A White name yields as many more callbacks as an additional eight years of experience on a resume. Since applicants' names are randomly assigned, this gap can only be attributed to the name manipulation.

Race also affects the reward to having a better resume. Whites with higher-quality resumes receive nearly 30-percent more callbacks than Whites with lower-quality resumes. On the other hand, having a higher-quality resume has a smaller effect for African-Americans. In other words, the gap between Whites and African-Americans widens with resume quality. While one may have expected improved credentials to alleviate employers' fear that African-American applicants are deficient in some unobservable skills, this is not the case in our data.[5]

The experiment also reveals several other aspects of the differential treatment by race. First, since we randomly assign applicants' postal addresses to the resumes, we can study the effect of neighborhood of residence on the likelihood of callback. We find that living in a wealthier (or more educated or Whiter) neighborhood increases callback rates. But, interestingly, African-Americans are not helped more than Whites by living in a "better" neighborhood. Second, the racial gap we measure in different industries does not appear correlated to Census-based measures of the racial gap in wages. The same is true for the racial gap we measure in different occupations. In fact, we find that the racial gaps in callback are statistically indistinguishable across all the occupation and industry categories covered in the experiment. Federal contractors, who are thought to be more severely constrained by affirmative action laws, do not treat the African-American resumes more preferentially; neither do larger employers or employers who explicitly state that they are "Equal Opportunity Employers." In Chicago, we find a slightly smaller racial gap when employers are located in more African-American neighborhoods.

The rest of the paper is organized as follows. Section I compares this experiment to earlier work on racial discrimination, and most notably to the labor market audit studies. We describe the experimental design in Section II and present the results in Section III, subsection A. In Section IV, we discuss possible interpretations of our results, focusing especially on two issues. First, we examine whether the

[3] In creating the higher-quality resumes, we deliberately make small changes in credentials so as to minimize the risk of overqualification.

[4] For ease of exposition, we refer to the effects uncovered in this experiment as racial differences. Technically, however, these effects are about the racial soundingness of names. We briefly discuss below the potential confounds between name and race. A more extensive discussion is offered in Section IV, subsection B.

[5] These results contrast with the view, mostly based on nonexperimental evidence, that African-Americans receive higher returns to skills. For example, estimating earnings regressions on several decades of Census data, James J. Heckman et al. (2001) show that African-Americans experience higher returns to a high school degree than Whites do.

race-specific names we have chosen might also proxy for social class above and beyond the race of the applicant. Using birth certificate data on mother's education for the different first names used in our sample, we find little relationship between social background and the name-specific callback rates.[6] Second, we discuss how our results map back to the different models of discrimination proposed in the economics literature. In doing so, we focus on two important results: the lower returns to credentials for African-Americans and the relative homogeneity of the racial gap across occupations and industries. We conclude that existing models do a poor job of explaining the full set of findings. Section V concludes.

I. Previous Research

With conventional labor force and household surveys, it is difficult to study whether differential treatment occurs in the labor market.[7] Armed only with survey data, researchers usually measure differential treatment by comparing the labor market performance of Whites and African-Americans (or men and women) for which they observe similar sets of skills. But such comparisons can be quite misleading. Standard labor force surveys do not contain all the characteristics that employers observe when hiring, promoting, or setting wages. So one can never be sure that the minority and nonminority workers being compared are truly similar from the employers' perspective. As a consequence, any measured differences in outcomes could be attributed to these unobserved (to the researcher) factors.

This difficulty with conventional data has led some authors to instead rely on pseudo-experiments.[8] Claudia Goldin and Cecilia Rouse (2000), for example, examine the effect of blind auditioning on the hiring process of orchestras. By observing the treatment of female candidates before and after the introduction of blind auditions, they try to measure the amount of sex discrimination. When such pseudo-experiments can be found, the resulting study can be very informative; but finding such experiments has proven to be extremely challenging.

A different set of studies, known as audit studies, attempts to place comparable minority and White actors into actual social and economic settings and measure how each group fares in these settings.[9] Labor market audit studies send comparable minority (African-American or Hispanic) and White auditors in for interviews and measure whether one is more likely to get the job than the other.[10] While the results vary somewhat across studies, minority auditors tend to perform worse on average: they are less likely to get called back for a second interview and, conditional on getting called back, less likely to get hired.

These audit studies provide some of the cleanest nonlaboratory evidence of differential treatment by race. But they also have weaknesses, most of which have been highlighted in Heckman and Siegelman (1992) and Heckman (1998). First, these studies require that both members of the auditor pair are identical in all dimensions that might affect productivity in employers' eyes, except for race. To accomplish this, researchers typically match auditors on several characteristics (height, weight, age, dialect, dressing style, hairdo) and train them for several days to coordinate interviewing styles. Yet, critics note that this is unlikely to erase the numerous differences that exist between the auditors in a pair.

Another weakness of the audit studies is that they are not double-blind. Auditors know the purpose of the study. As Turner et al. (1991)

[6] We also argue that a social class interpretation would find it hard to explain some of our findings, such as why living in a better neighborhood does not increase callback rates more for African-American names than for White names.

[7] See Joseph G. Altonji and Rebecca M. Blank (1999) for a detailed review of the existing literature on racial discrimination in the labor market.

[8] William A. Darity, Jr. and Patrick L. Mason (1998) describe an interesting nonexperimental study. Prior to the Civil Rights Act of 1964, employment ads would explicitly state racial biases, providing a direct measure of differential treatment. Of course, as Arrow (1998) mentions, discrimination was at that time "a fact too evident for detection."

[9] Michael Fix and Marjery A. Turner (1998) provide a survey of many such audit studies.

[10] Earlier hiring audit studies include Jerry M. Newman (1978) and Shelby J. McIntyre et al. (1980). Three more recent studies are Harry Cross et al. (1990), Franklin James and Steve W. DelCastillo (1991), and Turner et al. (1991). Heckman and Peter Siegelman (1992), Heckman (1998), and Altonji and Blank (1999) summarize these studies. See also David Neumark (1996) for a labor market audit study on gender discrimination.

note: "The first day of training also included an introduction to employment discrimination, equal employment opportunity, and a review of project design and methodology." This may generate conscious or subconscious motives among auditors to generate data consistent or inconsistent with their beliefs about race issues in America. As psychologists know very well, these demand effects can be quite strong. It is very difficult to insure that auditors will not want to do "a good job." Since they know the goal of the experiment, they can alter their behavior in front of employers to express (indirectly) their own views. Even a small belief by auditors that employers treat minorities differently can result in measured differences in treatment. This effect is further magnified by the fact that auditors are not in fact seeking jobs and are therefore more free to let their beliefs affect the interview process.

Finally, audit studies are extremely expensive, making it difficult to generate large enough samples to understand nuances and possible mitigating factors. Also, these budgetary constraints worsen the problem of mismatched auditor pairs. Cost considerations force the use of a limited number of pairs of auditors, meaning that any one mismatched pair can easily drive the results. In fact, these studies generally tend to find significant differences in outcomes across pairs.

Our study circumvents these problems. First, because we only rely on resumes and not people, we can be sure to generate comparability across race. In fact, since race is randomly assigned to each resume, the same resume will sometimes be associated with an African-American name and sometimes with a White name. This guarantees that any differences we find are caused solely by the race manipulation. Second, the use of paper resumes insulates us from demand effects. While the research assistants know the purpose of the study, our protocol allows little room for conscious or subconscious deviations from the set procedures. Moreover, we can objectively measure whether the randomization occurred as expected. This kind of objective measurement is impossible in the case of the previous audit studies. Finally, because of relatively low marginal cost, we can send out a large number of resumes. Besides giving us more precise estimates, this larger sample size also allows us to examine the nature of the differential treatment from many more angles.

II. Experimental Design

A. *Creating a Bank of Resumes*

The first step of the experimental design is to generate templates for the resumes to be sent. The challenge is to produce a set of realistic and representative resumes without using resumes that belong to actual job seekers. To achieve this goal, we start with resumes of actual job searchers but alter them sufficiently to create distinct resumes. The alterations maintain the structure and realism of the initial resumes without compromising their owners.

We begin with resumes posted on two job search Web sites as the basis for our artificial resumes.[11] While the resumes posted on these Web sites may not be completely representative of the average job seeker, they provide a practical approximation.[12] We restrict ourselves to people seeking employment in our experimental cities (Boston and Chicago). We also restrict ourselves to four occupational categories: sales, administrative support, clerical services, and customer services. Finally, we further restrict ourselves to resumes posted more than six months prior to the start of the experiment. We purge the selected resumes of the person's name and contact information.

During this process, we classify the resumes within each detailed occupational category into two groups: high and low quality. In judging resume quality, we use criteria such as labor market experience, career profile, existence of gaps in employment, and skills listed. Such a classification is admittedly subjective but it is made independently of any race assignment on the resumes (which occurs later in the experimental design). To further reinforce the quality gap between the two sets of resumes, we add to each high-quality resume a subset of the following features: summer or while-at-school employment experience, volunteering experience, extra computer skills, certification degrees, foreign language skills, honors, or some military

[11] The sites are www.careerbuilder.com and www.americasjobbank.com.

[12] In practice, we found large variation in skill levels among people posting their resumes on these sites.

experience. This resume quality manipulation needs to be somewhat subtle to avoid making a higher-quality job applicant overqualified for a given job. We try to avoid this problem by making sure that the features listed above are not all added at once to a given resume. This leaves us with a high-quality and a low-quality pool of resumes.[13]

To minimize similarity to actual job seekers, we use resumes from Boston job seekers to form templates for the resumes to be sent out in Chicago and use resumes from Chicago job seekers to form templates for the resumes to be sent out in Boston. To implement this migration, we alter the names of the schools and previous employers on the resumes. More specifically, for each Boston resume, we use the Chicago resumes to replace a Boston school with a Chicago school.[14] We also use the Chicago resumes to replace a Boston employer with a Chicago employer in the same industry. We use a similar procedure to migrate Chicago resumes to Boston.[15] This produces distinct but realistic looking resumes, similar in their education and career profiles to this subpopulation of job searchers.[16]

B. *Identities of Fictitious Applicants*

The next step is to generate identities for the fictitious job applicants: names, telephone numbers, postal addresses, and (possibly) e-mail addresses. The choice of names is crucial to our experiment.[17] To decide on which names are uniquely African-American and which are uniquely White, we use name frequency data calculated from birth certificates of all babies born in Massachusetts between 1974 and 1979. We tabulate these data by race to determine which names are distinctively White and which are distinctively African-American. Distinctive names are those that have the highest ratio of frequency in one racial group to frequency in the other racial group.

As a check of distinctiveness, we conducted a survey in various public areas in Chicago. Each respondent was asked to assess features of a person with a particular name, one of which is race. For each name, 30 respondents were asked to identify the name as either "White," "African-American," "Other," or "Cannot Tell." In general, the names led respondents to readily attribute the expected race for the person but there were a few exceptions and these names were disregarded.[18]

The final list of first names used for this study is shown in Appendix Table A1. The table reports the relative likelihood of the names for the Whites and African-Americans in the Massachusetts birth certificates data as well as the recognition rate in the field survey.[19] As Appendix Table A1 indicates, the African-American first names used in the experiment are quite common in the population. This suggests that by using these names as an indicator of race, we are actually covering a rather large segment of the African-American population.[20]

Applicants in each race/sex/city/resume quality cell are allocated the same phone number. This guarantees that we can precisely track employer callbacks in each of these cells. The phone lines we use are virtual ones with only a voice mailbox attached to them. A similar outgoing message is recorded on each of the voice mailboxes but each message is recorded by someone of the appropriate race and gender.

[13] In Section III, subsection B, and Table 3, we provide a detailed summary of resume characteristics by quality level.

[14] We try as much as possible to match high schools and colleges on quality and demographic characteristics.

[15] Note that for applicants with schooling or work experience outside of the Boston or Chicago areas, we leave the school or employer name unchanged.

[16] We also generate a set of different fonts, layouts, and cover letters to further differentiate the resumes. These are applied at the time the resumes are sent out.

[17] We chose name over other potential manipulations of race, such as affiliation with a minority group, because we felt such affiliations may especially convey more than race.

[18] For example, Maurice and Jerome are distinctively African-American names in a frequency sense yet are not perceived as such by many people.

[19] So many of names show a likelihood ratio of ∞ because there is censoring of the data at five births. If there are fewer than five babies in any race/name cell, it is censored (and we do not know whether a cell has zero or was censored). This is primarily a problem for the computation of how many African-American babies have "White" names.

[20] We also tried to use more White-sounding last names for White applicants and more African-American-sounding last names for African-American applicants. The last names used for White applicants are: Baker, Kelly, McCarthy, Murphy, Murray, O'Brien, Ryan, Sullivan, and Walsh. The last names used for African-American applicants are: Jackson, Jones, Robinson, Washington, and Williams.

Since we allocate the same phone number for applicants with different names, we cannot use a person name in the outgoing message.

While we do not expect positive feedback from an employer to take place via postal mail, resumes still need postal addresses. We therefore construct fictitious addresses based on real streets in Boston and Chicago using the White Pages. We select up to three addresses in each 5-digit zip code in Boston and Chicago. Within cities, we randomly assign addresses across all resumes. We also create eight e-mail addresses, four for Chicago and four for Boston.[21] These e-mail addresses are neutral with respect to both race and sex. Not all applicants are given an e-mail address. The e-mail addresses are used almost exclusively for the higher-quality resumes. This procedure leaves us with a bank of names, phone numbers, addresses, and e-mail addresses that we can assign to the template resumes when responding to the employment ads.

C. *Responding to Ads*

The experiment was carried out between July 2001 and January 2002 in Boston and between July 2001 and May 2002 in Chicago.[22] Over that period, we surveyed all employment ads in the Sunday editions of *The Boston Globe* and *The Chicago Tribune* in the sales, administrative support, and clerical and customer services sections. We eliminate any ad where applicants were asked to call or appear in person. In fact, most of the ads we surveyed in these job categories ask for applicants to fax in or (more rarely) mail in their resume. We log the name (when available) and contact information for each employer, along with any information on the position advertised and specific requirements (such as education, experience, or computer skills). We also record whether or not the ad explicitly states that the employer is an equal opportunity employer.

For each ad, we use the bank of resumes to sample four resumes (two high-quality and two low-quality) that fit the job description and requirements as closely as possible.[23] In some cases, we slightly alter the resumes to improve the quality of the match, such as by adding the knowledge of a specific software program.

One of the high- and one of the low-quality resumes selected are then drawn at random to receive African-American names, the other high- and low-quality resumes receive White names.[24] We use male and female names for sales jobs, whereas we use nearly exclusively female names for administrative and clerical jobs to increase callback rates.[25] Based on sex, race, city, and resume quality, we assign a resume the appropriate phone number. We also select at random a postal address. Finally, e-mail addresses are added to most of the high-quality resumes.[26] The final resumes are formatted, with fonts, layout, and cover letter style chosen at random. The resumes are then faxed (or in a few cases mailed) to the employer. All in all, we respond to more than 1,300 employment ads over the entire sample period and send close to 5,000 resumes.

D. *Measuring Responses*

We measure whether a given resume elicits a callback or e-mail back for an interview. For each phone or e-mail response, we use the content of the message left by the employer (name of the applicant, company name, telephone number for contact) to match the response to the corresponding resume-ad pair.[27] Any attempt by employers to contact applicants via postal mail cannot be measured in our experiment since the addresses are fictitious. Several human resource managers confirmed to us that

[21] The e-mail addresses are registered on Yahoo.com, Angelfire.com, or Hotmail.com.

[22] This period spans tighter and slacker labor markets. In our data, this is apparent as callback rates (and number of new ads) dropped after September 11, 2001. Interestingly, however, the racial gap we measure is the same across these two periods.

[23] In some instances, our resume bank does not have four resumes that are appropriate matches for a given ad. In such instances, we send only two resumes.

[24] Though the same names are repeatedly used in our experiment, we guarantee that no given ad receives multiple resumes with the same name.

[25] Male names were used for a few administrative jobs in the first month of the experiment.

[26] In the first month of the experiment, a few high-quality resumes were sent without e-mail addresses and a few low-quality resumes were given e-mail addresses. See Table 3 for details.

[27] Very few employers used e-mail to contact an applicant back.

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (*p*-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the *p*-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

employers rarely, if ever, contact applicants via postal mail to set up interviews.

E. *Weaknesses of the Experiment*

We have already highlighted the strengths of this experiment relative to previous audit studies. We now discuss its weaknesses. First, our outcome measure is crude, even relative to the previous audit studies. Ultimately, one cares about whether an applicant gets the job and about the wage offered conditional on getting the job. Our procedure, however, simply measures callbacks for interviews. To the extent that the search process has even moderate frictions, one would expect that reduced interview rates would translate into reduced job offers. However, we are not able to translate our results into gaps in hiring rates or gaps in earnings.

Another weakness is that the resumes do not directly report race but instead suggest race through personal names. This leads to various sources of concern. First, while the names are chosen to make race salient, some employers may simply not notice the names or not recognize their racial content. On a related note, because we are not assigning race but only race-specific names, our results are not representative of the average African-American (who may not have such a racially distinct name).[28] We return to this issue in Section IV, subsection B.

Finally, and this is an issue pervasive in both our study and the pair-matching audit studies, newspaper ads represent only one channel for job search. As is well known from previous work, social networks are another common means through which people find jobs and one that clearly cannot be studied here. This omission could qualitatively affect our results if African-Americans use social networks more or if employers who rely more on networks differentiate less by race.[29]

III. Results

A. *Is There a Racial Gap in Callback?*

Table 1 tabulates average callback rates by racial soundingness of names. Included in brackets under each rate is the number of resumes sent in that cell. Row 1 presents our results for the full data set. Resumes with White

[28] As Appendix Table A1 indicates, the African-American names we use are, however, quite common among African-Americans, making this less of a concern.

[29] In fact, there is some evidence that African-Americans may rely less on social networks for their job search (Harry J. Holzer, 1987).

names have a 9.65 percent chance of receiving a callback. Equivalent resumes with African-American names have a 6.45 percent chance of being called back. This represents a difference in callback rates of 3.20 percentage points, or 50 percent, that can solely be attributed to the name manipulation. Column 4 shows that this difference is statistically significant.[30] Put in other words, these results imply that a White applicant should expect on average one callback for every 10 ads she or he applies to; on the other hand, an African-American applicant would need to apply to about 15 different ads to achieve the same result.[31]

How large are these effects? While the cost of sending additional resumes might not be large per se, this 50-percent gap could be quite substantial when compared to the rate of arrival of new job openings. In our own study, the biggest constraining factor in sending more resumes was the limited number of new job openings each week. Another way to benchmark the measured return to a White name is to compare it to the returns to other resume characteristics. For example, in Table 5, we will show that, at the average number of years of experience in our sample, an extra year of experience increases the likelihood of a callback by a 0.4 percentage point. Based on this point estimate, the return to a White name is equivalent to about eight additional years of experience.

Rows 2 and 3 break down the full sample of sent resumes into the Boston and Chicago markets. About 20 percent more resumes were sent in Chicago than in Boston. The average callback rate (across races) is lower in Chicago than in Boston. This might reflect differences in labor market conditions across the two cities over the experimental period or maybe differences in the ability of the MIT and Chicago teams of research assistants in selecting resumes that were good matches for a given help-wanted ad. The percentage difference in callback rates is, however, strikingly similar across both cities. White applicants are 49 percent more likely than African-American applicants to receive a callback in Chicago and 50 percent more likely in Boston. These racial differences are statistically significant in both cities.

Finally, rows 4 to 7 break down the full sample into female and male applicants. Row 4 displays the average results for all female names while rows 5 and 6 break the female sample into administrative (row 5) and sales jobs (row 6); row 7 displays the average results for all male names. As noted earlier, female names were used in both sales and administrative job openings whereas male names were used close to exclusively for sales openings.[32] Looking across occupations, we find a significant racial gap in callbacks for both males (52 percent) and females (49 percent). Comparing males to females in sales occupations, we find a larger racial gap among males (52 percent versus 22 percent). Interestingly, females in sales jobs appear to receive more callbacks than males; however, this (reverse) gender gap is statistically insignificant and economically much smaller than any of the racial gaps discussed above.

Rather than studying the distribution of callbacks at the applicant level, one can also tabulate the distribution of callbacks at the employment-ad level. In Table 2, we compute the fraction of employers that treat White and African-American applicants equally, the fraction of employers that favor White applicants and the fraction of employers that favor African-American applicants. Because we send up to four resumes in response to each sampled ad, the three categories above can each take three different forms. Equal treatment occurs when either no applicant gets called back, one White and one African-American get called back or two Whites and two African-Americans get called back. Whites are favored when either only one White gets called back, two Whites and no African-American get called back or two Whites and one African-American get called back. African-Americans are favored in all other cases.

As Table 2 indicates, equal treatment occurs for about 88 percent of the help-wanted ads. As expected, the major source of equal treatment comes from the high fraction of ads for which

[30] These statistical tests assume independence of callbacks. We have, however, verified that the results stay significant when we assume that the callbacks are correlated either at the employer or first-name level.

[31] This obviously assumes that African-American applicants cannot assess a priori which firms are more likely to treat them more or less favorably.

[32] Only about 6 percent of all male resumes were sent in response to an administrative job opening.

TABLE 2—DISTRIBUTION OF CALLBACKS BY EMPLOYMENT AD

Equal Treatment:	No Callback	1W + 1B	2W + 2B
88.13 percent	83.37	3.48	1.28
[1,166]	[1,103]	[46]	[17]
Whites Favored (WF):	1W + 0B	2W + 0B	2W + 1B
8.39 percent	5.59	1.44	1.36
[111]	[74]	[19]	[18]
African-Americans Favored (BF):	1B + 0W	2B + 0W	2B + 1W
3.48 percent	2.49	0.45	0.53
[46]	[33]	[6]	[7]
Ho: *WF* = *BF*			
p = 0.0000			

Notes: This table documents the distribution of callbacks at the employment-ad level. "No Callback" is the percent of ads for which none of the fictitious applicants received a callback. "1W + 1B" is the percent of ads for which exactly one White and one African-American applicant received a callback. "2W + 2B" is the percent of ads for which exactly two White applicants and two African-American applicants received a callback. "Equal Treatment" is defined as the sum of "No Callback," "1W + 1B," and "2W + 2B." "1W + 0B" is the percent of ads for which exactly one White applicant and no African-American applicant received a call back. "2W + 0B" is the percent of ads for which excatly two White applicants and no African-American applicant received a callback. "2W + 1B" is the percent of ads for which exactly two White applicants and one African-American applicant received a callback. "Whites Favored" is defined as the sum of "1W + 0B," "2W + 0B," and "2W + 1B." "1B + 0W" is the percent of ads for which exactly one African-American applicant and no White applicant received a callback. "2B + 0W" is the percent of ads for which exactly two African-American applicants and no White applicant received a callback. "2B + 1W" is the percent of ads for which exactly two African-American applicants and one White applicant received a callback. "African-Americans Favored" is defined as the sum of "1B + 0W," "2B + 0W," and "2B + 1W." In brackets in each cell is the number of employment ads in that cell. "*Ho*: *WF* = *WB*" reports the p-value for a test of symmetry between the proportion of employers that favor White names and the proportion of employers that favor African-American names.

no callbacks are recorded (83 percent of the ads). Whites are favored by nearly 8.4 percent of the employers, with a majority of these employers contacting exactly one White applicant. African-Americans, on the other hand, are favored by only about 3.5 percent of employers. We formally test whether there is symmetry in the favoring of Whites over African-Americans and African-Americans over Whites. We find that the difference between the fraction of employers favoring Whites and the fraction of employers favoring African-Americans is statistically very significant (p = 0.0000).

B. *Do African-Americans Receive Different Returns to Resume Quality?*

Our results so far demonstrate a substantial gap in callback based on applicants' names. Next, we would like to learn more about the factors that may influence this gap. More specifically, we ask how employers respond to improvements in African-American applicants' credentials. To answer this question, we examine how the racial gap in callback varies by resume quality.

As we explained in Section II, for most of the employment ads we respond to, we send four different resumes: two higher-quality and two lower-quality ones. Table 3 gives a better sense of which factors enter into this subjective classification. Table 3 displays means and standard deviations of the most relevant resume characteristics for the full sample (column 1), as well as broken down by race (columns 2 and 3) and resume quality (columns 4 and 5). Since applicants' names are randomized, there is no difference in resume characteristics by race. Columns 4 and 5 document the objective differences between resumes subjectively classified as high and low quality. Higher-quality applicants have on average close to an extra year of labor market experience, fewer employment holes (where an employment hole is defined as a period of at least six months without a reported job), are more likely to have worked while at school, and to report some military experience. Also, higher-quality applicants are more likely to have an e-mail address, to have received some honors, and to list some computer skills and other special skills (such as a certification degree or foreign language skills) on their resume. Note that the higher- and lower-quality resumes do not differ on average with regard to

TABLE 3—RESUME CHARACTERISTICS: SUMMARY STATISTICS

Sample:	All resumes	White names	African-American	Higher quality	Lower quality
Characteristic:					
College degree	0.72	0.72	0.72	0.72	0.71
(Y = 1)	(0.45)	(0.45)	(0.45)	(0.45)	(0.45)
Years of experience	7.84	7.86	7.83	8.29	7.39
	(5.04)	(5.07)	(5.01)	(5.29)	(4.75)
Volunteering experience?	0.41	0.41	0.41	0.79	0.03
(Y = 1)	(0.49)	(0.49)	(0.49)	(0.41)	(0.16)
Military experience?	0.10	0.09	0.10	0.19	0.00
(Y = 1)	(0.30)	(0.29)	(0.30)	(0.39)	(0.06)
E-mail address?	0.48	0.48	0.48	0.92	0.03
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.27)	(0.17)
Employment holes?	0.45	0.45	0.45	0.34	0.56
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.47)	(0.50)
Work in school?	0.56	0.56	0.56	0.72	0.40
(Y = 1)	(0.50)	(0.50)	(0.50)	(0.45)	(0.49)
Honors?	0.05	0.05	0.05	0.07	0.03
(Y = 1)	(0.22)	(0.23)	(0.22)	(0.25)	(0.18)
Computer skills?	0.82	0.81	0.83	0.91	0.73
(Y = 1)	(0.38)	(0.39)	(0.37)	(0.29)	(0.44)
Special skills?	0.33	0.33	0.33	0.36	0.30
(Y = 1)	(0.47)	(0.47)	(0.47)	(0.48)	(0.46)
Fraction high school dropouts in applicant's zip code	0.19	0.19	0.19	0.19	0.18
	(0.08)	(0.08)	(0.08)	(0.08)	(0.08)
Fraction college or more in applicant's zip code	0.21	0.21	0.21	0.21	0.22
	(0.17)	(0.17)	(0.17)	(0.17)	(0.17)
Fraction Whites in applicant's zip code	0.54	0.55	0.54	0.53	0.55
	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
Fraction African-Americans in applicant's zip code	0.31	0.31	0.31	0.32	0.31
	(0.33)	(0.33)	(0.33)	(0.33)	(0.33)
Log(median per capital income) in applicant's zip code	9.55	9.55	9.55	9.54	9.56
	(0.56)	(0.56)	(0.55)	(0.54)	(0.57)
Sample size	4,870	2,435	2,435	2,446	2,424

Notes: The table reports means and standard deviations for the resume characteristics as listed on the left. Column 1 refers to all resumes sent; column 2 refers to resumes with White names; column 3 refers to resumes with African-American names; column 4 refers to higher-quality resumes; column 5 refers to lower-quality resumes. See text for details.

applicants' education level. This reflects the fact that all sent resumes, whether high or low quality, are chosen to be good matches for a given job opening. About 70 percent of the sent resumes report a college degree.[33]

The last five rows of Table 3 show summary characteristics of the applicants' zip code address. Using 1990 Census data, we compute the fraction of high school dropouts, fraction of college educated or more, fraction of Whites, fraction of African-Americans and log(median per capital income) for each zip code used in the experiment. Since addresses are randomized within cities, these neighborhood quality measures are uncorrelated with race or resume quality.

The differences in callback rates between high- and low-quality resumes are presented in Panel A of Table 4. The first thing to note is that the resume quality manipulation works: higher-quality resumes receive more callbacks. As row 1 indicates, we record a callback rate of close to 11 percent for White applicants with a higher-quality resume, compared to 8.5 percent for White applicants with lower-quality resumes. This is a statistically significant difference of 2.29 percentage points, or 27 percent (p = 0.0557). Most strikingly, African-Americans experience much less of an increase in callback

[33] This varies from about 50 percent for the clerical and administrative support positions to more than 80 percent for the executive, managerial, and sales representatives positions.

TABLE 4—AVERAGE CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES AND RESUME QUALITY

	Panel A: Subjective Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p-value)
White names	8.50	10.79	1.27	2.29
	[1,212]	[1,223]		(0.0557)
African-American names	6.19	6.70	1.08	0.51
	[1,212]	[1,223]		(0.6084)
	Panel B: Predicted Measure of Quality (Percent Callback)			
	Low	High	Ratio	Difference (p- value)
White names	7.18	13.60	1.89	6.42
	[822]	[816]		(0.0000)
African-American names	5.37	8.60	1.60	3.23
	[819]	[814]		(0.0104)

Notes: Panel A reports the mean callback percents for applicant with a White name (row 1) and African-American name (row 2) depending on whether the resume was subjectively qualified as a lower quality or higher quality. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback rates are equal across quality groups within each racial group. For Panel B, we use a third of the sample to estimate a probit regression of the callback dummy on the set of resume characteristics as displayed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). We then use the estimated coefficients on the set of resume characteristics to estimate a predicted callback for the remaining resumes (two-thirds of the sample). We call "high-quality" resumes the resumes that rank above the median predicted callback and "low-quality" resumes the resumes that rank below the median predicted callback. In brackets is the number of resumes sent for each race/quality group. The last column reports the p-value of a test of proportion testing the null hypothesis that the callback percents are equal across quality groups within each racial group.

rate for similar improvements in their credentials. African-Americans with higher-quality resumes receive a callback 6.7 percent of the time, compared to 6.2 percent for African-Americans with lower quality resumes. This is only a 0.51-percentage-point, or 8-percent, difference and this difference is not statistically significant ($p = 0.6084$).

Instead of relying on the subjective quality classification, Panel B directly uses resume characteristics to classify the resumes. More specifically, we use a random subsample of one-third of the resumes to estimate a probit regression of the callback dummy on the resume characteristics listed in Table 3. We further control for a sex dummy, a city dummy, six occupation dummies, and a vector of job requirements as listed in the employment ads.[34] We then use the estimated coefficients on the resume characteristics to rank the remaining two-thirds of the resumes by predicted callback. In Panel B, we classify as "high" those resumes that have above-median-predicted callback; similarly, we classify as "low" those resumes that have below-median-predicted callback. As one can see from Panel B, qualitatively similar results emerge from this analysis. While African-Americans do appear to significantly benefit from higher-quality resumes under this alternative classification, they benefit less than Whites. The ratio of callback rates for high- versus low-quality resumes is 1.60 for African Americans, compared to 1.89 for Whites.

In Table 5, we directly report the results of race-specific probit regressions of the callback dummy on resume characteristics. We, however, start in column 1 with results for the full sample of sent resumes. As one can see, many of the resume characteristics have the expected effect on the likelihood of a callback. The addition of an e-mail address, honors, and special skills all have a positive and significant effect on the likelihood of a callback.[35] Also, more experienced applicants are more likely to get called back: at the average number of years of experience in our sample (eight years), each

[34] See Section III, subsection D, for more details on these occupation categories and job requirements.

[35] Note that the e-mail address dummy, because it is close to perfectly correlated with the subjective resume-quality variable, may in part capture some other unmeasured resume characteristics that may have led us to categorize a given resume as higher quality.

TABLE 5—EFFECT OF RESUME CHARACTERISTICS ON LIKELIHOOD OF CALLBACK

Dependent Variable: Callback Dummy Sample:	All resumes	White names	African-American names
Years of experience (*10)	0.07	0.13	0.02
	(0.03)	(0.04)	(0.03)
Years of experience^2 (*100)	−0.02	−0.04	−0.00
	(0.01)	(0.01)	(0.01)
Volunteering? (Y = 1)	−0.01	−0.01	0.01
	(0.01)	(0.01)	(0.01)
Military experience? (Y = 1)	−0.00	0.02	−0.01
	(0.01)	(0.03)	(0.02)
E-mail? (Y = 1)	0.02	0.03	−0.00
	(0.01)	(0.01)	(0.01)
Employment holes? (Y = 1)	0.02	0.03	0.01
	(0.01)	(0.02)	(0.01)
Work in school? (Y = 1)	0.01	0.02	−0.00
	(0.01)	(0.01)	(0.01)
Honors? (Y = 1)	0.05	0.06	0.03
	(0.02)	(0.03)	(0.02)
Computer skills? (Y = 1)	−0.02	−0.04	−0.00
	(0.01)	(0.02)	(0.01)
Special skills? (Y = 1)	0.05	0.06	0.04
	(0.01)	(0.02)	(0.01)
Ho: Resume characteristics effects are all zero (*p*-value)	54.50 (0.0000)	57.59 (0.0000)	23.85 (0.0080)
Standard deviation of predicted callback	0.047	0.062	0.037
Sample size	4,870	2,435	2,435

Notes: Each column gives the results of a probit regression where the dependent variable is the callback dummy. Reported in the table are estimated marginal changes in probability for the continuous variables and estimated discrete changes for the dummy variables. Also included in each regression are a city dummy, a sex dummy, six occupation dummies, and a vector of dummy variables for job requirements as listed in the employment ad (see Section III, subsection D, for details). Sample in column 1 is the entire set of sent resumes; sample in column 2 is the set of resumes with White names; sample in column 3 is the set of resumes with African-American names. Standard errors are corrected for clustering of the observations at the employment-ad level. Reported in the second to last row are the *p*-values for a χ^2 testing that the effects on the resume characteristics are all zero. Reported in the second to last row is the standard deviation of the predicted callback rate.

extra year of experience increases the likelihood of a callback by about a 0.4 percentage point. The most counterintuitive effects come from computer skills, which appear to negatively predict callback, and employment holes, which appear to positively predict callback.

The same qualitative patterns hold in column 2 where we focus on White applicants. More importantly, the estimated returns to an e-mail address, additional work experience, honors, and special skills appear economically stronger for that racial group. For example, at the average number of years of experience in our sample, each extra year of experience increases the likelihood of a callback by about a 0.7 percentage point.

As might have been expected from the two previous columns, we find that the estimated returns on these resume characteristics are all economically and statistically weaker for African-American applicants (column 3). In fact, all the estimated effects for African-Americans are statistically insignificant, except for the return to special skills. Resume characteristics thus appear less predictive of callback rates for African-Americans than they are for Whites. To illustrate this more saliently, we predict callback rates using either regression estimates in column 2 or regression estimates in column 3. The standard deviation of the predicted callback from column 2 is 0.062, whereas it is only 0.037 from column 3. In summary, employers simply seem to pay less attention or discount more the characteristics listed on the

TABLE 6—EFFECT OF APPLICANT'S ADDRESS ON LIKELIHOOD OF CALLBACK

Dependent Variable: Callback Dummy						
Zip code characteristic:	Fraction Whites		Fraction college or more		Log(per capital income)	
Zip code characteristic	0.020	0.020	0.054	0.053	0.018	0.014
	(0.012)	(0.016)	(0.022)	(0.031)	(0.007)	(0.010)
Zip code characteristic* African-American name	—	−0.000	—	−0.002	—	0.008
		(0.024)		(0.048)		(0.015)
African-American name	—	−0.031	—	−0.031	—	−0.112
		(0.015)		(0.013)		(0.152)

Notes: Each column gives the results of a probit regression where the dependent variable is the callback dummy. Reported in the table is the estimated marginal change in probability. Also included in columns 1, 3, and 5 is a city dummy; also included in columns 2, 4, and 6 is a city dummy and a city dummy interacted with a race dummy. Standard errors are corrected for clustering of the observations at the employment-ad level.

resumes with African-American-sounding names. Taken at face value, these results suggest that African-Americans may face relatively lower individual incentives to invest in higher skills.[36]

C. *Applicants' Address*

An incidental feature of our experimental design is the random assignment of addresses to the resumes. This allows us to examine whether and how an applicant's residential address, all else equal, affects the likelihood of a callback. In addition, and most importantly for our purpose, we can also ask whether African-American applicants are helped relatively more by residing in more affluent neighborhoods.

We perform this analysis in Table 6. We start (columns 1, 3, and 5) by discussing the effect of neighborhood of residence across all applicants. Each of these columns reports the results of a probit regression of the callback dummy on a specific zip code characteristic and a city dummy. Standard errors are corrected for clustering of the observations at the employment-ad level. We find a positive and significant effect of neighborhood quality on the likelihood of a callback. Applicants living in Whiter (column 1), more educated (column 3), or higher-income (column 5) neighborhoods have a higher probability of receiving a callback. For example, a 10-percentage-point increase in the fraction of college-educated in zip code of residence increases the likelihood of a callback by a 0.54 percentage point (column 3).

In columns 2, 4, and 6, we further interact the zip code characteristic with a dummy variable for whether the applicant is African-American or not. Each of the probit regressions in these columns also includes an African-American dummy, a city dummy, and an interaction of the city dummy with the African-American dummy. There is no evidence that African-Americans benefit any more than Whites from living in a Whiter, more educated zip code. The estimated interactions between fraction White and fraction college educated with the African-American dummy are economically very small and statistically insignificant. We do find an economically more meaningful effect of zip code median income level on the racial gap in callback; this effect, however, is statistically insignificant.

In summary, while neighborhood quality affects callbacks, African-Americans do not benefit more than Whites from living in better neighborhoods. If ghettos and bad neighborhoods are particularly stigmatizing for African-Americans, one might have expected African-Americans to be helped more by having a "better" address. Our results do not support this hypothesis.

D. *Job and Employer Characteristics*

Table 7 studies how various job requirements (as listed in the employment ads) and employer characteristics correlate with the racial gap in callback. Each row of Table 7 focuses on a specific job or employer characteristic, with

[36] This of course assumes that the changes in job and wage offers associated with higher skills are the same across races, or at least not systematically larger for African-Americans.

TABLE 7—EFFECT OF JOB REQUIREMENT AND EMPLOYER CHARACTERISTICS ON RACIAL DIFFERENCES IN CALLBACKS

Job requirement:	Sample mean (standard deviation)	Marginal effect on callbacks for African-American names
Any requirement? (Y = 1)	0.79	0.023
	(0.41)	(0.015)
Experience? (Y = 1)	0.44	0.011
	(0.49)	(0.013)
Computer skills? (Y = 1)	0.44	0.000
	(0.50)	(0.013)
Communication skills? (Y = 1)	0.12	−0.000
	(0.33)	(0.015)
Organization skills? (Y = 1)	0.07	0.028
	(0.26)	(0.029)
Education? (Y = 1)	0.11	−0.031
	(0.31)	(0.017)
Total number of requirements	1.18	0.002
	(0.93)	(0.006)
Employer characteristic:	**Sample mean (standard deviation)**	**Marginal effect on callbacks for African-American names**
Equal opportunity employer? (Y = 1)	0.29	−0.013
	(0.45)	(0.012)
Federal contractor? (Y = 1)	0.11	−0.035
(N = 3,102)	(0.32)	(0.016)
Log(employment)	5.74	−0.001
(N = 1,690)	(1.74)	(0.005)
Ownership status:		
(N = 2,878)		
Privately held	0.74	0.011
		(0.019)
Publicly traded	0.15	−0.025
		(0.015)
Not-for-profit	0.11	0.025
		(0.042)
Fraction African-Americans in employer's zip code	0.08	0.117
(N = 1,918)	(0.15)	(0.062)

Notes: Sample is all sent resumes (N = 4,870) unless otherwise specified in column 1. Column 2 reports means and standard deviations (in parentheses) for the job requirement or employer characteristic. For ads listing an experience requirement, 50.1 percent listed "some," 24.0 percent listed "two years or less," and 25.9 percent listed "three years or more." For ads listing an education requirement, 8.8 percent listed a high school degree, 48.5 percent listed some college, and 42.7 percent listed at least a four-year college degree. Column 3 reports the marginal effect of the job requirement or employer characteristic listed in that row on differential treatment. Specifically, each cell in column 3 corresponds to a different probit regression of the callback dummy on an African-American name dummy, a dummy for the requirement or characteristic listed in that row and the interaction of the requirement or characteristic dummy with the African-American name dummy. Reported in each cell is the estimated change in probability for the interaction term. Standard errors are corrected for clustering of the observations at the employment-ad level.

summary statistics in column 2. Column 3 shows the results of various probit regressions. Each entry in this column is the marginal effect of the specific characteristic listed in that row on the racial gap in callback. More specifically, each entry is from a separate probit regression of a callback dummy on an African-American dummy, the characteristic listed in that row and the interaction of that characteristic with the African-American dummy. The reported coefficient is that on the interaction term.

We start with job requirements. About 80 percent of the ads state some form of requirement. About 44 percent of the ads require some minimum experience, of which roughly 50 percent simply ask for "some experience," 24 percent less than two years, and 26 percent at least three years of experience. About 44 percent of

ads mention some computer knowledge requirement, which can range from Excel or Word to more esoteric software programs. Good communication skills are explicitly required in about 12 percent of the ads. Organization skills are mentioned 7 percent of the time. Finally, only about 11 percent of the ads list an explicit education requirement. Of these, 8.8 percent require a high school degree, 48.5 percent some college (such as an associate degree), and the rest at least a four-year college degree.[37]

Despite this variability, we find little systematic relationship between any of the requirements and the racial gap in callback. The point estimates in column 3 show no consistent economic pattern and are all statistically weak. Measures of job quality, such as experience or computer skills requirements, do not predict the extent of the racial gap. Communication or other interpersonal skill requirements have no effect on the racial gap either.[38]

We also study employer characteristics. Collecting such information is a more difficult task since it is not readily available from the employment ads we respond to. The only piece of employer information we can directly collect from the employment ad is whether or not the employer explicitly states being an "Equal Opportunity Employer." In several cases, the name of the employer is not even mentioned in the ad and the only piece of information we can rely on is the fax number which applications must be submitted to. We therefore have to turn to supplemental data sources. For employment ads that do not list a specific employer, we first use the fax number to try to identify the company name via Web reverse-lookup services. Based on company names, we use three different data sources (*Onesource Business Browser, Thomas Register,* and *Dun and Bradstreet Million Dollar Directory, 2001*) to track company information such as total employment, industry, and ownership status. Using this same set of data sources, we also try to identify the specific zip code of the company (or company branch) that resumes are to be sent to. Finally, we use the Federal Procurement and Data Center Web site to find a list of companies that have federal contracts.[39] The racial difference in callback rates for the subsamples where employer characteristics could be determined is very similar in magnitude to that in the full sample.

Employer characteristics differ significantly across ads. Twenty-nine percent of all employers explicitly state that they are "Equal Opportunity Employers." Eleven percent are federal contractors and, therefore, might face greater scrutiny under affirmative action laws. The average company size is around 2,000 employees but there is a lot of variation across firms. Finally, 74 percent of the firms are privately held, 15 percent are publicly traded, and 11 percent are not-for-profit organizations.

Neither "Equal Opportunity Employers" nor federal contractors appear to treat African-Americans more favorably. In fact, each of these employer characteristics is associated with a larger racial gap in callback (and this effect is marginally significant for federal contractors). Differential treatment does not vary with employer size.[40] Point estimates indicate less differential treatment in the not-for-profit sector; however, this effect is very noisily estimated.[41]

In an unpublished Appendix (available from the authors upon request), we also study how the racial gap in callback varies by occupation and industry. Based on the employment ad listings, we classify the job openings into six occupation categories: executives and managers; administrative supervisors; sales representatives; sales workers; secretaries and legal assistants; clerical workers. We also, when possible,

[37] Other requirements sometimes mentioned include typing skills for secretaries (with specific words-per-minute minimum thresholds), and, more rarely, foreign language skills.

[38] Other ways of estimating these effects produce a similar nonresult. Among other things, we considered including a city dummy or estimating the effects separately by city; we also estimated one single probit regression including all requirements at once.

[39] This Web site (www.fpdc.gov) is accurate up to and including March 21, 2000.

[40] Similar results hold when we measure employer size using a total sales measure rather than an employment measure.

[41] Our measurement of the racial gap by firm or employer type may not be a good indicator of the fraction of African-Americans actually employed in these firms. For example, "Equal Opportunity Employers" may receive a higher fraction of African-American resumes. Their actual hiring may therefore look different from that of non "Equal Opportunity Employers" when one considers the full set of resumes they receive.

classify employers into six industry categories: manufacturing; transportation and communication; wholesale and retail trade; finance, insurance, and real estate; business and personal services; health, educational, and social services. We then compute occupation and industry-specific racial gaps in callback and relate these gaps to 1990 Census-based measures of occupation and industry earnings, as well as Census-based measures of the White/African-American wage gap in these occupations and industries.

We find a positive White/African-American gap in callbacks in all occupation and industry categories (except for transportation and communication). While average earnings vary a lot across the occupations covered in the experiment, we find no systematic relationship between occupation earnings and the racial gap in callback. Similarly, the industry-specific gaps in callback do not relate well to a measure of inter-industry wage differentials. In fact, while the racial gap in callback rates varies somewhat across occupations and industries, we cannot reject the null hypothesis that the gap is the same across all these categories.

The last row of Table 7 focuses on the marginal effect of employer location on the racial gap in callback.[42] We use as a measure of employer location the zip code of the company (or company branch) resumes were to be sent to. More specifically, we ask whether differential treatment by race varies with the fraction of African-Americans in the employer's zip code. We find a marginally significant positive effect of employer location on African-American callbacks but this effect is extremely small. In regressions not reported here (but available from the authors upon request), we reestimate this effect separately by city. While the point estimates are positive for both cities, the effect is only statistically significant for Chicago.

IV. Interpretation

Three main sets of questions arise when interpreting the results above. First, does a higher callback rate for White applicants imply that employers are discriminating against African-Americans? Second, does our design only isolate the effect of race or is the name manipulation conveying some other factors than race? Third, how do our results relate to different models of racial discrimination?

A. *Interpreting Callback Rates*

Our results indicate that for two identical individuals engaging in an identical job search, the one with an African-American name would receive fewer interviews. Does differential treatment within our experiment imply that employers are discriminating against African-Americans (whether it is rational, prejudice-based, or other form of discrimination)? In other words, could the lower callback rate we record for African-American resumes *within our experiment* be consistent with a racially neutral review of the *entire pool* of resumes the surveyed employers receive?

In a racially neutral review process, employers would rank order resumes based on their quality and call back all applicants that are above a certain threshold. Because names are randomized, the White and African-American resumes we send should rank similarly on average. So, irrespective of the skill and racial composition of the applicant pool, a race-blind selection rule would generate equal treatment of Whites and African-Americans. So our results must imply that employers use race as a factor when reviewing resumes, which matches the legal definition of discrimination.

But even rules where employers are not trying to interview as few African-American applicants as possible may generate observed differential treatment in our experiment. One such hiring rule would be employers trying to interview a target level of African-American candidates. For example, perhaps the average firm in our experiment aims to produce an interview pool that matches the population base rate. This rule could produce the observed differential treatment if the average firm receives a higher proportion of African-American resumes than the population base rate because African-Americans disproportionately apply to the jobs and industries in our sample.[43]

[42] For previous work on the effect of employer location on labor market discrimination, see, for example, Steven Raphael et al. (2000).

[43] Another variant of this argument is that the (up to) two African-American resumes we sent are enough to signifi-

Some of our other findings may be consistent with such a rule. For example, the fact that "Equal Opportunity Employers" or federal contractors do not appear to discriminate any less may reflect the fact that such employers receive more applications from African-Americans. On the other hand, other key findings run counter to this rule. As we discuss above, we find no systematic difference in the racial gap in callback across occupational or industry categories, despite the large variation in the fraction of African-Americans looking for work in those categories. African-Americans are underrepresented in managerial occupations, for example. If employers matched base rates in the population, the few African-Americans who apply to these jobs should receive a higher callback rate than Whites. Yet, we find that the racial gap in managerial occupations is the same as in all the other job categories. This rule also runs counter to our findings on returns to skill. Suppose firms are struggling to find White applicants but overwhelmed with African-American ones. Then they should be less sensitive to the quality of White applicants (as they are trying to fill in their hiring quota for Whites) and much more sensitive to the quality of Black applicants (when they have so many to pick from). Thus, it is unlikely that the differential treatment we observe is generated by hiring rules such as these.

B. *Potential Confounds*

While the names we have used in this experiment strongly signal racial origin, they may also signal some other personal trait. More specifically, one might be concerned that employers are inferring social background from the personal name. When employers read a name like "Tyrone" or "Latoya," they may assume that the person comes from a disadvantaged background.[44] In the extreme form of this social background interpretation, employers do not care at all about race but are discriminating only against the social background conveyed by the names we have chosen.[45]

While plausible, we feel that some of our earlier results are hard to reconcile with this interpretation. For example, in Table 6, we found that while employers value "better" addresses, African-Americans are not helped more than Whites by living in Whiter or more educated neighborhoods. If the African-American names we have chosen mainly signal negative social background, one might have expected the estimated name gap to be lower for better addresses. Also, if the names mainly signal social background, one might have expected the name gap to be higher for jobs that rely more on soft skills or require more interpersonal interactions. We found no such evidence in Table 7.

We, however, directly address this alternative interpretation by examining the average social background of babies born with the names used in the experiment. We were able to obtain birth certificate data on mother's education (less than high school, high school or more) for babies born in Massachusetts between 1970 and

cantly distort the racial composition of the entire applicant pool. This is unlikely for two reasons. First, anecdotal evidence and the empirically low callback rates we record suggest that firms typically receive many hundreds of resumes in response to each ad they post. Hence, the (up to) four resumes we send out are unlikely to influence the racial composition of the pool. Second, the similar racial gap in callback we observe across the two cities goes counter to this interpretation since the racial composition base rates differ quite a lot across these two cities. Another variant of this argument is that, for some reason, the average firm in our sample receives a lot of high-quality resumes from African-American applicants and much fewer high-quality resumes from White applicants. Hypothetically, this might occur if high-quality African-Americans are much more likely to use help-wanted ads rather than other job search channels. If employers perform within-race comparisons and again want to target a certain racial mix in their interviewing and hiring, our African-American resumes may naturally receive lower callbacks as they are competing with many more high-quality applicants. This specific argument would be especially relevant in a case where the average sampled employer is "known" to be good to African-Americans. But our selection procedure for the employment ads did not allow for such screening: we simply responded to as many ads as possible in the targeted occupational categories.

[44] Roland Fryer and Steven Levitt (2003) provide a recent analysis of social background and naming conventions amongst African-Americans.

[45] African-Americans as a whole come from more disadvantaged backgrounds than Whites. For this social class effect to be something of independent interest, one must assert that African-Americans with the African-American names we have selected are from a lower social background than the average African-American and/or that Whites with the White names we have selected are from a higher social background than the average White. We come back to this point below.

TABLE 8—CALLBACK RATE AND MOTHER'S EDUCATION BY FIRST NAME

White female			African-American female		
Name	Percent callback	Mother education	Name	Percent callback	Mother education
Emily	7.9	96.6	Aisha	2.2	77.2
Anne	8.3	93.1	Keisha	3.8	68.8
Jill	8.4	92.3	Tamika	5.5	61.5
Allison	9.5	95.7	Lakisha	5.5	55.6
Laurie	9.7	93.4	Tanisha	5.8	64.0
Sarah	9.8	97.9	Latoya	8.4	55.5
Meredith	10.2	81.8	Kenya	8.7	70.2
Carrie	13.1	80.7	Latonya	9.1	31.3
Kristen	13.1	93.4	Ebony	9.6	65.6
Average		91.7	Average		61.0
Overall		83.9	Overall		70.2
Correlation	−0.318	(p = 0.404)	Correlation	−0.383	(p = 0.309)

White male			African-American male		
Name	Percent callback	Mother education	Name	Percent callback	Mother education
Todd	5.9	87.7	Rasheed	3.0	77.3
Neil	6.6	85.7	Tremayne	4.3	—
Geoffrey	6.8	96.0	Kareem	4.7	67.4
Brett	6.8	93.9	Darnell	4.8	66.1
Brendan	7.7	96.7	Tyrone	5.3	64.0
Greg	7.8	88.3	Hakim	5.5	73.7
Matthew	9.0	93.1	Jamal	6.6	73.9
Jay	13.4	85.4	Leroy	9.4	53.3
Brad	15.9	90.5	Jermaine	9.6	57.5
Average		91.7	Average		66.7
Overall		83.5	Overall		68.9
Correlation	−0.0251	(p = 0.949)	Correlation	−0.595	(p = 0.120)

Notes: This table reports, for each first name used in the experiment, callback rate and average mother education. Mother education for a given first name is defined as the percent of babies born with that name in Massachusetts between 1970 and 1986 whose mother had at least completed a high school degree (see text for details). Within each sex/race group, first names are ranked by increasing callback rate. "Average" reports, within each race-gender group, the average mother education for all the babies born with one of the names used in the experiment. "Overall" reports, within each race-gender group, average mother education for all babies born in Massachusetts between 1970 and 1986 in that race-gender group. "Correlation" reports the Spearman rank order correlation between callback rate and mother education *within* each race-gender group as well as the p-value for the test of independence.

1986.[46] For each first name in our experiment, we compute the fraction of babies with that name and, in that gender-race cell, whose mothers have at least completed a high school degree.

In Table 8, we display the average callback rate for each first name along with this proxy for social background. Within each race-gender group, the names are ranked by increasing callback rate. Interestingly, there is significant

[46] This longer time span (compared to that used to assess name frequencies) was imposed on us for confidentiality reasons. When fewer than 10 births with education data available are recorded in a particular education-name cell, the exact number of births in that cell is not reported and we impute five births. Our results are not sensitive to this imputation. One African-American female name (Latonya) and two male names (Rasheed and Hakim) were imputed in this way. One African-American male name (Tremayne) had too few births with available education data and was therefore dropped from this analysis. Our results are qualitatively similar when we use a larger data set of California births for the years 1989 to 2000 (kindly provided to us by Steven Levitt).

variation in callback rates by name. Of course, chance alone could produce such variation because of the rather small number of observations in each cell (about 200 for the female names and 70 for the male names).[47]

The row labeled "Average" reports the average fraction of mothers that have at least completed high school for the set of names listed in that gender-race group. The row labeled "Overall" reports the average fraction of mothers that have at least completed high school for the full sample of births in that gender-race group. For example, 83.9 percent of White female babies born between 1970 and 1986 have mothers with at least a high school degree; 91.7 percent of the White female babies with one of the names used in the experiment have mothers with at least a high school degree.

Consistent with a social background interpretation, the African-American names we have chosen fall below the African-American average. For African-American male names, however, the gap between the experimental names and the population average is negligible. For White names, both the male and female names are above the population average.

But, more interestingly to us, there is substantial between-name heterogeneity in social background. African-American babies named Kenya or Jamal are affiliated with much higher mothers' education than African-American babies named Latonya or Leroy. Conversely, White babies named Carrie or Neil have lower social background than those named Emily or Geoffrey. This allows for a direct test of the social background hypothesis within our sample: are names associated with a worse social background discriminated against more? In the last row in each gender-race group, we report the rank-order correlation between callback rates and mother's education. The social background hypothesis predicts a positive correlation. Yet, for all four categories, we find the exact opposite. The p-values indicate that we cannot reject independence at standard significance levels except in the case of African-American males where we can almost reject it at the 10-percent level ($p = 0.120$). In summary, this test suggests little evidence that social background drives the measured race gap.

Names might also influence our results through familiarity. One could argue that the African-American names used in the experiment simply appear odd to human resource managers and that any odd name is discriminated against. But as noted earlier, the names we have selected are not particularly uncommon among African-Americans (see Appendix Table A1). We have also performed a similar exercise to that of Table 8 and measured the rank-order correlation between name-specific callback rates and name frequency within each gender-race group. We found no systematic positive correlation.

There is one final potential confound to our results. Perhaps what appears as a bias against African-Americans is actually the result of *reverse discrimination.* If qualified African-Americans are thought to be in high demand, then employers with average quality jobs might feel that an equally talented African-American would never accept an offer from them and thereby never call her or him in for an interview. Such an argument might also explain why African-Americans do not receive as strong a return as Whites to better resumes, since higher qualification only strengthens this argument. But this interpretation would suggest that among the better jobs, we ought to see evidence of reverse discrimination, or at least a smaller racial gap. However, as we discussed in Section III, subsection D, we do not find any such evidence. The racial gap does not vary across jobs with different skill requirements, nor does it vary across occupation categories. Even among the better jobs in our sample, we find that employers significantly favor applicants with White names.[48]

[47] We formally tested whether this variation was significant by estimating a probit regression of the callback dummy on all the personal first names, allowing for clustering of the observations at the employment-ad level. For all but African-American females, we cannot reject the null hypothesis that all the first name effects in the same race-gender group are the same. Of course, a lack of a rejection does not mean there is no underlying pattern in the between-name variation in callbacks that might have been detectable with larger sample sizes.

[48] One might argue that employers who reverse-discriminate hire through less formal channels than help-wanted ads. But this would imply that African-Americans are less likely to find jobs through formal channels. The evidence on exit out of unemployment does not paint a clear picture in this direction (Holzer, 1987).

C. *Relation to Existing Theories*

What do these results imply for existing models of discrimination? Economic theories of discrimination can be classified into two main categories: taste-based and statistical discrimination models.[49] Both sets of models can obviously "explain" our average racial gap in callbacks. But can these models explain our other findings? More specifically, we discuss the relevance of these models with a focus on two of the facts that have been uncovered in this paper: (i) the lower returns to credentials for African-Americans; (ii) the relative uniformity of the race gap across occupations, job requirements and, to a lesser extent, employer characteristics and industries.

Taste-based models (Gary S. Becker, 1961) differ in whose prejudiced "tastes" they emphasize: customers, coworkers, or employers. Customer and co-worker discrimination models seem at odds with the lack of significant variation of the racial gap by occupation and industry categories, as the amount of customer contact and the fraction of White employees vary quite a lot across these categories. We do not find a larger racial gap among jobs that explicitly require "communication skills" and jobs for which we expect either customer or coworker contacts to be higher (retail sales for example).

Because we do not know what drives employer tastes, employer discrimination models could be consistent with the lack of occupation and industry variation. Employer discrimination also matches the finding that employers located in more African-American neighborhoods appear to discriminate somewhat less. However, employer discrimination models would struggle to explain why African-Americans get relatively lower returns to their credentials. Indeed, the cost of indulging the discrimination taste should increase as the minority applicants' credentials increase.[50]

Statistical discrimination models are the prominent alternative to the taste-based models in the economics literature. In one class of statistical discrimination models, employers use (observable) race to proxy for *unobservable* skills (e.g., Edmund S. Phelps, 1972; Kenneth J. Arrow, 1973). This class of models struggle to explain the credentials effect as well. Indeed, the added credentials should lead to a larger update for African-Americans and hence greater returns to skills for that group.

A second class of statistical discrimination models "emphasize the precision of the information that employers have about individual productivity" (Altonji and Blank, 1999). Specifically, in these models, employers believe that the same observable signal is more precise for Whites than for African-Americans (Dennis J. Aigner and Glenn G. Cain, 1977; Shelly J. Lundberg and Richard Startz, 1983; Bradford Cornell and Ivo Welch, 1996). Under such models, African-Americans receive lower returns to observable skills because employers place less weight on these skills. However, how reasonable is this interpretation for our experiment? First, it is important to note that we are using the same set of resume characteristics for both racial groups. So the lower precision of information for African-Americans cannot be that, for example, an employer does not know what a high school degree from a very African-American neighborhood means (as in Aigner and Cain, 1977). Second, many of the credentials on the resumes are in fact externally and easily verifiable, such as a certification for a specific software.

An alternative version of these models would rely on bias in the observable signal rather than differential variance or noise of these signals by race. Perhaps the skills of African-Americans are discounted because affirmative action makes it easier for African-Americans to get these skills. While this is plausible for credentials such as an employee-of-the-month honor, it is unclear why this would apply to more verifiable and harder skills. It is equally unclear why work experience would be less rewarded since our study suggests that getting a job is more, not less, difficult for African-Americans.

The uniformity of the racial gap across occupations is also troubling for a statistical discrimination interpretation. Numerous factors that should affect the level of statistical discrimination, such as the importance of unobservable skills, the observability of qualifications, the precision of observable skills and the ease of

[49] Darity and Mason (1998) provide a more thorough review of a variety of economic theories of discrimination.

[50] One could, however, assume that employer tastes differ not just by race but also by race and skill, so that employers have greater prejudice against minority workers with better credentials. But the opposite preferences, employers having a particular distaste for low-skilled African-Americans, also seem reasonable.

performance measurement, may vary quite a lot across occupations.

This discussion suggests that perhaps other models may do a better job at explaining our findings. One simple alternative model is lexicographic search by employers. Employers receive so many resumes that they may use quick heuristics in reading these resumes. One such heuristic could be to simply read no further when they see an African-American name. Thus they may never see the skills of African-American candidates and this could explain why these skills are not rewarded. This might also to some extent explain the uniformity of the race gap since the screening process (i.e., looking through a large set of resumes) may be quite similar across the variety of jobs we study.[51]

[51] Another explanation could be based on employer stereotyping or categorizing. If employers have coarser stereotypes for African-Americans, many of our results would follow. See Melinda Jones (2002) for the relevant psychology and Mullainathan (2003) for a formalization of the categorization concept.

V. Conclusion

This paper suggests that African-Americans face differential treatment when searching for jobs and this may still be a factor in why they do poorly in the labor market. Job applicants with African-American names get far fewer callbacks for each resume they send out. Equally importantly, applicants with African-American names find it hard to overcome this hurdle in callbacks by improving their observable skills or credentials.

Taken at face value, our results on differential returns to skill have possibly important policy implications. They suggest that training programs alone may not be enough to alleviate the racial gap in labor market outcomes. For training to work, some general-equilibrium force outside the context of our experiment would have to be at play. In fact, if African-Americans recognize how employers reward their skills, they may rationally be less willing than Whites to even participate in these programs.

TABLE A1—FIRST NAMES USED IN EXPERIMENT

White female Name	L(W)/L(B)	Perception White	African-American female Name	L(B)/L(W)	Perception Black
Allison	∞	0.926	Aisha	209	0.97
Anne	∞	0.962	Ebony	∞	0.9
Carrie	∞	0.923	Keisha	116	0.93
Emily	∞	0.925	Kenya	∞	0.967
Jill	∞	0.889	Lakisha	∞	0.967
Laurie	∞	0.963	Latonya	∞	1
Kristen	∞	0.963	Latoya	∞	1
Meredith	∞	0.926	Tamika	284	1
Sarah	∞	0.852	Tanisha	∞	1
	Fraction of all births:			Fraction of all births:	
	3.8 percent			7.1 percent	

White male Name	L(W)/L(B)	Perception White	African-American male Name	L(B)/L(W)	Perception Black
Brad	∞	1	Darnell	∞	0.967
Brendan	∞	0.667	Hakim		0.933
Geoffrey	∞	0.731	Jamal	257	0.967
Greg	∞	1	Jermaine	90.5	1
Brett	∞	0.923	Kareem	∞	0.967
Jay	∞	0.926	Leroy	44.5	0.933
Matthew	∞	0.888	Rasheed	∞	0.931
Neil	∞	0.654	Tremayne	∞	0.897
Todd	∞	0.926	Tyrone	62.5	0.900
	Fraction of all births:			Fraction of all births:	
	1.7 percent			3.1 percent	

Notes: This table tabulates the different first names used in the experiment and their identifiability. The first column reports the likelihood that a baby born with that name (in Massachusetts between 1974 and 1979) is White (or African-American) relative to the likelihood that it is African-American (White). The second column reports the probability that the name was picked as White (or African-American) in an independent field survey of people. The last row for each group of names shows the proportion of all births in that race group that these names account for.

REFERENCES

Aigner, Dennis J. and Cain. Glenn G. "Statistical Theories of Discrimination in Labor Markets." *Industrial and Labor Relations Review,* January 1977, *30*(1), pp. 175–87.

Altonji, Joseph G. and Blank, Rebecca M. "Race and Gender in the Labor Markey," in Orley Ashenfelter and David Card, eds., *Handbook of labor economics,* Vol. 30. Amsterdam: North-Holland, 1999, pp. 3143–259.

Arrow, Kenneth, J. "The Theory of Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in labor markets.* Princeton, NJ: Princeton University Press, 1973, pp. 3–33.

_____. "What Has Economics to Say about Racial Discrimination?" *Journal of Economic Perspectives*, Spring 1998, *12*(2), pp. 91–100.

Becker, Gary S. *The economics of discrimination,* 2nd Ed. Chicago: University of Chicago Press, 1961.

Brown, Colin and Gay, Pat. *Racial discrimination 17 years after the act.* London: Policy Studies Institute, 1985.

Cornell, Bradford and Welch, Ivo. "Culture, Information, and Screening Discrimination." *Journal of Political Economy,* June 1996, *104*(3), pp. 542–71.

Council of Economic Advisers. *Changing America: Indicators of social and economic well-being by race and Hispanic origin.* September 1998, http://w3.access.gpo.gov/eop/ca/pdfs/ca.pdf.

Cross, Harry; Kenney, Genevieve; Mell, Jane and Zimmerman, Wendy. *Employer hiring practices: Differential treatment of Hispanic and Anglo job applicants.* Washington, DC: Urban Institute Press, 1990.

Darity, William A., Jr. and Mason, Patrick L. "Evidence on Discrimination in Employment: Codes of Color, Codes of Gender." *Journal of Economic Perspectives,* Spring 1998, *12*(2), pp. 63–90.

Fix, Michael and Turner, Margery A., eds. *A national report card on discrimination in America: The role of testing.* Washington, DC: Urban Institute Press, 1998.

Fryer, Roland and Levitt, Steven. "The Causes and Consequences of Distinctively Black Names." Mimeo, University of Chicago, 2003.

Goldin, Claudia and Rouse, Cecilia. "Orchestrating Impartiality: The Impact of Blind Auditions on Female Musicians." *American Economic Review,* September 2000, *90*(4), pp. 715–41.

Heckman, James J. "Detecting Discrimination." *Journal of Economic Perspectives,* Spring 1998, *12*(2), pp. 101–16.

Heckman, James J.; Lochner, Lance J., and Todd, Petra E. "Fifty Years of Mincer Earnings Regressions." Mimeo, University of Chicago, 2001.

Heckman, James J. and Siegelman, Peter. "The Urban Institute Audit Studies: Their Methods and Findings," in Michael Fix and Raymond J. Struyk, eds., *Clear and convincing evidence: Measurement of discrimination in America.* Lanham, MD: Urban Institute Press, 1992, pp. 187–258.

Holzer, Harry J. "Informal Job Search and Black Youth Unemployment." *American Economic Review,* June 1987, *77*(3), pp. 446–52.

Hubbuck, Jim and Carter, Simon. *Half a chance? A report on job discrimination against young blacks in Nottingham.* London: Commission for Racial Equality, 1980.

James, Franklin and DelCastillo, Steve W. "Measuring Job Discrimination by Private Employers Against Young Black and Hispanic Seeking Entry Level Work in the Denver Metropolitan Area." Mimeo, University of Colorado-Denver, 1991.

Jones, Melinda. *Social psychology of prejudice.* Saddle River, NJ: Pearson Education, 2002.

Jowell, Roger and Prescott-Clark, Patricia. "Racial Discrimination and White-Collar Workers in Britain." *Race,* November 1970, *11*(4), pp. 397–417.

Lundberg, Shelly J. and Starz, Richard. "Private Discrimination and Social Intervention in Competitive Labor Market." *American Economic Review,* June 1983, *73*(3), pp. 340–47.

McIntyre, Shelby J.; Moberg, Dennis J. and Posner, Barry Z. "Discrimination in Recruitment: An Empirical Analysis: Comment." *Industrial and Labor Relations Review,* July 1980, *33*(4), pp. 543–47.

Mullainathan, Sendhil. "Thinking Through Categories." Mimeo, Massachusetts Institute of Technology, 2003.

Neumark, David. "Sex Discrimination in Restaurant Hiring: An Audit Study." *Quarterly Journal of Economics,* August 1996, *111*(3), pp. 915–42.

Newman, Jerry M. "Discrimination in Recruitment: An Empirical Analysis." *Industrial and Labor Relations Review,* October 1978, *32*(1), pp. 15–23.

Nisbett, Richard E. and Cohen, Dov. *The culture of honor: The psychology of violence in the South.* Boulder, CO: Westview Press, 1996.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *American Economic Review,* September 1972, *62*(4), pp. 659–61.

Raphael, Steven; Stoll, Michael A. and Holzer, Harry J. "Are Suburban Firms More Likely to Discriminate against African Americans?" *Journal of Urban Economics,* November 2000, *48*(3), pp. 485–508.

Riach, Peter A. and Rich, Judity. "Testing for Racial Discrimination in the Labour Market." *Cambridge Journal of Economics,* September 1991, *15*(3), pp. 239–56.

Turner, Margery A.; Fix, Michael and Struyk, Raymond J. *Opportunities denied, opportunities diminished: Racial discrimination in hiring.* Washington, DC: Urban Institute Press, 1991.

Weichselbaumer, Doris. "Sexual Orientation Discrimination in Hiring." *Labour Economics,* December 2003, *10*(6), pp. 629–42.

_____. "Is it Sex or Personality? The Impact of Sex-Stereotypes on Discrimination in Applicant Selection." *Eastern Economic Journal,* Spring 2004, *30*(2), pp. 159–86.

EXHIBIT S

Theory in Action, Vol. 7, No. 3, July (© 2014)
DOI: 10.3798/tia.1937-0237.14018

Does the Sophomore Slump Really Exist?

Debra Wetcher-Hendricks[1]

Various statisticians and sports analysts have addressed the role of regression to the mean in major league sports performances. This paper intensifies general findings on the topic, specifically, that of Schall and Smith (2000), to focus upon regression with respect to the widely-acknowledged sophomore slump. Results of the analysis yield two main findings. First, players' comparatively poor performances during their second seasons of play reflect uncharacteristically high first-season statistics, not low second-season statistics. Second, the degree of change in performance from the first to the second seasons clearly follows Campbell and Stanley's (1963) regression to the mean model. *[Article copies available for a fee from The Transformative Studies Institute. E-mail address: journal@transformativestudies.org Website: http://www.transformativestudies.org* ©2014 *by The Transformative Studies Institute. All rights reserved.]*

KEYWORDS: Regression to the Mean; Regression Artifacts; Sophomore Slump; Freshman Fluke.

INTRODUCTION

Sports players, analysts, and fans alike have long acknowledged the decline in performance that commonly occurs between the first and second years of players' careers. This phenomenon, nicknamed the

[1] **Debra Wetcher-Hendricks**, Ph.D., received her B.A. in journalism from Glassboro State College and her M.A. in Social Relations as well as her Ph.D. in Applied Social Research from Lehigh University. Currently, she is an Associate Professor of Sociology at Moravian College, in Bethlehem, PA. Her primary academic interests include social research and statistical methods, interpersonal communication, and mass communication. In addition to her published book, *Analyzing Quantitative Data: An Introduction for Social Researchers,* her articles describing analyses of statistical models and providing statistical analyses of sociological phenomena have appeared in peer-refereed journals and been presented at academic conferences. Various editions of *Who's Who Among American Women*, *Who's Who in America*, and *Who's Who in the World* have recognized her since 2008. Address correspondence to: Debra Wetcher-Hendricks, Moravian College, 1200 Main Street - Sociology Department, Bethlehem, PA 18018; e-mail: medwh02@moravian.edu.

1937-0229 ©2014 Transformative Studies Institute

sophomore slump, most often refers to former first-year standouts plagued with comparatively poor second-year performances.[1]

Analyses tend give the most attention to the sophomore slump when it affects those who received the Rookie of the Year (ROTY) award. Bennett describes the ROTY award as a Most Valuable Player Award, limited to "rookie eligible" players.[2] First officially given in 1947, the award honors a single rookie player (1947 and 1948) or a rookie player from the American League and a rookie player from the National League (1949 until present). Ex-post facto ROTY award winners between the years of 1872 and 1948, until each league began identifying its own ROTY, have also been named.[3]

The contention about the relationship between the ROTY award and the sophomore slump suggests that those who have won the award may be average players who just happened to have good seasons during their first years in the major leagues. Sabermetricians have, to varying degrees, addressed this possibility before. In a discussion of his recent research regarding ROTY award recipients, Bennett implies a bit of arbitrariness to qualifying for the award, as it focuses upon a player's single-year performance with no baseline (no pun intended) by which to judge his overall performance.[4] He specifically presents the sophomore slump as an aftermath of having received the award. Other studies on the matter focus upon the heightened attention to the sophomore slump for ROTY award recipients.[5] Taylor and Cuave also discuss the sophomore slump in relation to players with exceptional rookie-year performance.[6] Although they never specifically identify ROTY winners as the only players affected, they note that 80 percent of those who received the ROTY award between 2000 and 2003 performed more poorly during their second years in the major leagues than they did during the first. Taylor and Cuave also examined average productivity for all players receiving votes for the ROTY between 1994 and 2003, finding a 12 percent decrease between their freshman and sophomore years.

Finding examples of particular Rookie of the Year award winners who have declined in performance during their second years in the major leagues provides little challenge. Consider the following.[7]

- After a standout first season in the Major Leagues and receiving the 2002 American League ROTY award Eric Hinske's second season was plagued by injuries and a 20 percent decline in Win Shares (WS).

- The WS of 1976 American League ROTY, Mark Fidrych, also dropped 20 points by the end of his second years playing Major League Baseball.
- Angel Berroa, the 2003 American League ROTY, saw a 13-point drop in his On-Base Percentage Plus Slugging (OBP+) and a 6 point drop in his WS numbers between his first and second years in the Major Leagues. Berroa's performance declined so much that his team relegated him to the Minor Leagues for a portion of his second season.
- Statistics for Berroa's National League counterpart, Dontrelle Willis, also indicate a decline in performance just after he received the ROTY award. Specifically, his sophomore-year Earned Run Average (ERA) exceeded his freshman-year ERA by .72 and his WS for his sophomore year fell 4 points shy of that for his freshman year in the Major Leagues.

These examples only begin this list of ROTY players with worse performances during their second than during their first years in the Major Leagues. Others have had similar, although not necessarily so dramatic, experiences. Walters and Gleeman, for example, mention many of these players in their articles.[8] The repeated mention of having received the ROTY award and experiencing a sophomore slump in conjunction with one another clearly suggests a connection between the two. The analysis presented in this paper follows others' leads in discussing the sophomore slump in terms of former ROTY award winners.

The clear changes in performance have two possible explanations. Players may simply have uncharacteristically bad seasons during their second years of play, favoring the idea of a sophomore slump. In this case, players' performances should improve after their second years of play, making their statistics for the remainder of their careers resemble the statistics from the rookie year. But, it is also possible that players perform uncharacteristically well during their first years, indicating more of a freshman fluke than a sophomore slump. This idea suggests that players' statistics fall after their rookie years and do not improve thereafter.[9] Both situations produce the same initial result: a worsening of performance between the first and second years.

Each of these possibilities may indicate some overall pattern in players' performances as a result of sports-related trends. It may, for instance, take a year for established players to learn the strengths and weaknesses of a new player, thus making the new player's first season

his most outstanding. More likely, however, the decline in performance results from a statistical principle called regression to the mean, which dictates that extreme values within a large data set tend to serve as the exception, rather than the rule. According to this standard, very high or very low numbers based upon data gathered at a particular point in time generally do not repeat themselves in future trials. So, in terms of sports, incredibly good performance during the first year or incredibly poor performance during the second year does not last. The player's performance usually returns to normal.

Others acknowledge this idea. Specifically, Albert describes the possibility that data based upon a single season may indicate a very low or a very high performance even when these values are not evident in the player's career statistics.[10] Therefore, he argues, season performance cannot appropriately represent overall natural ability. Walters additionally clarifies the role of statistical artifacts, stating that "All players have their ups and downs, but we tend to give awards to players experiencing temporary ups. We shouldn't be surprised when they come back down to earth, a tendency statisticians call 'regression to the mean.'"[11]

Sir Francis Galton introduced the idea of regression to the mean in 1886.[12] He suggested that regression to the mean threatens the internal validity of a study. Even if values from two sets of data differ significantly, one cannot necessarily attribute the difference to a change in conditions under which the data was gathered. Galton proved that, when a particular data set has unusually high or unusually low scores, subsequent data sets tend to contain scores closer to the overall mean. In other words, extreme scores eventually return to "normal" on their own.

Since Galton's discussion of regression to the mean, many statisticians have examined the phenomenon. Schall and Smith presented perhaps the most straightforward contention that the possibility that regression to the mean could explain changes in baseball players' performances throughout their careers.[13] In fact, they presented a model to predict players' future performances based upon the previous changes in their performances from season to season. Their ability to do so, however, assumes the availability of statistics from previous seasons. When dramatic changes in performance occur early in a player's career, such as for those experiencing a sophomore slump, it remains unclear whether the first or second year of play qualifies as the extreme and which characterizes player's typical performance. This paper, therefore particularizes Schall and Smith's analysis of regression to the mean to its

role within the first two seasons of play for those plagued by the sophomore slump.

Interestingly, a regression model proposed by Campbell and Stanley more than 45 years before Schall and Smith's study can help in this analysis.[14] Campbell and Stanley's model provides an estimate for the amount that an extreme score can be expected to regress toward the mean in the subsequent data-gathering period or trial. The correlation between the two data sets plays a role in this relationship. Given an extreme value on an initial trial, a subject's score on the next trial falls closer to the mean of all data in a proportion equal to the correlation coefficient. So,

$$Y_{reg} = Y_1 - r_{12}(|Y_1 - \mu|) \qquad \text{eq. 1}$$

for extremely high initial trials and

$$Y_{reg} = Y_1 + r_{12}(|Y_1 - \mu|) \qquad \text{eq. 2}$$

for extremely low initial trials. Symbolically, for both of these equations,

Y_{reg} = regressed score
Y_1 = score on trial 1
r_{12} = correlation between scores on trial 1 and trial 2
μ = theoretical mean.

The correlation between first and second year statistics indicates the proportion of the difference between a player's extreme performance and the average player's performance that should "disappear" during the following year. An actual change in performance would result in a difference between first-year and second-year statistics that exceeds the difference explained by regression to the mean.

GOALS

Based upon the concept of regression to the mean in the context of athletics, this analysis seeks to fulfill three goals. It first investigates the difference between performance during the first and second years of play to assure that a significant decline actually does exist. This assessment simply involves a statistical test of the null hypothesis that no difference exists between Rookie of the Year players' statistics during the first and second years of play (H_o:μ_1 μ_2). Rejection of this hypothesis confirms that players' first-year and second-year performances significantly differ.

Assuming rejection of this null hypothesis and a difference in the predicted direction, (i.e. lower second-year than first-year statistics rather than the opposite) the second step of this analysis attempts to determine which year, the freshman or the sophomore year, qualifies as the extreme year. If the statistics from the first year are significantly higher than those of later years in players' careers ($\mu_1 > \mu_{after\ year\ 2}$), then the change in performance signifies a freshman fluke. Significantly lower statistics from the second year of play than from all following years ($\mu_2 > \mu_{after\ year\ 2}$), however, indicate a sophomore slump. The statistical test to evaluate this issue utilizes the null hypothesis that statistics from the first year, the second year, and all other years of play do not differ significantly H_o: ($\mu_1 = \mu_2\ \ \mu_{after\ year\ 2}$). Results of this analysis should indicate which, if either, of the years qualifies as the extreme one.

The third step of this analysis focuses on the extreme year. After determining the difference between Rookie of the Year players' performances during the extreme year of play and all players' average performances, an inquiry into the possibility of regression to the mean can take place. Should a freshman fluke exist, according to Campbell and Stanley's claim, second-year statistics should still lie above the statistics for the average player, but should be the correlation coefficient's proportion smaller than those associated with the first year.[15] Conversely, should the sophomore year emerge as a low extreme, Rookie of the Year players' statistics from the third year until the end of a player's career should be closer to the first-year than the second-year values.

Until now, most analyses of Major League Baseball players have characterized offensive performance. Such measures have traditionally focused upon the "single goal of scoring runs."[16] Some further argue that, beyond the fact that offense only constitutes half of the game, the most commonly used offensive performance statistic, the batting average doesn't even provide a valid representation of offensive performance. According to Bennett and Fleuck, the batting average simply addresses the percentage of times that a player achieves a hit in all of his times at bat. The assumption of a direct relationship between the number of hits and the number of runs scored provides an indication of a player's success in scoring runs.[17] But, as any baseball fan understands, scoring a run involves speed and thoughtful base running as well as a successful hit. To complement the batting average, many consider players' runs or runs batted in (RBIs). However, state Bennett and Flueck, these values also fail to provide valid indications of particular player's offensive performances in that they may reflect other, extraneous factors as well. Batting order, for instance, may affect these numbers[18]. An order that

places an individual immediately after good hitters benefits that individual because previous batters on base prove the opportunity for him to acquire RBIs. Similarly, a batter who reaches base must often rely on following batters to provide the opportunity for him to score a run.

Baseball analysts have recognized these shortcomings and have, in some cases instituted new, more suitable, statistics such as win shares on-base percentage, and expected run production[19]. But, these measures still focus only on offensive performance, failing to provide an all-inclusive indication of a player's performance. [20] Taylor and Cuave have made an attempt to include pitcher's field statistics in assessments of player performance. [21] Their study, however, provides somewhat inconclusive results, indicating a true sophomore slump in some contexts and regression to the mean in others.

METHOD

Palmer and Gillette's *The Baseball Encyclopedia* provided the data for this analysis. The subject group includes the 238 players who had received ROTY awards between 1872 and 2004. Assessing these players' overall performances requires a statistic that accounts for players' accomplishments in both offensive and defensive situations. The relatively new statistic of Player Overall Wins (POW) provides a more comprehensive summary of players' overall performances than the common Value Over Replacement Player and On-Base Plus Slugging numbers (which address only offensive performance) and, even, Win Shares (which does not directly address base-stealing successes) do. *The Baseball Encyclopedia* presents the POW as a means to comprehensively assess players, comparing the sum or a player's "batting wins, fielding wins, base-stealing wins, and pitching wins," to that of an average player. [22] More simply, the statistic represents the sum or a player's Batter-Fielder Wins (BFW) and Pitcher Wins (PW). Generally, then, players with negative POW values have put their teams at disadvantages and those with positive POW values have helped their teams to win games.[23]

According to the concept of the Sophomore Slump, players should have lower POW values during their second years of play than during their first years or, on average, during the remainder of their careers. Two paired t-tests, one comparing mean POW values for players' first and second years of play and another comparing mean POW values for players' second years and for the remainders of their careers address this claim. A second-season mean POW that lies above the value for the first

season and the value for the remainder of the career or a lack of significant difference between the three values suggests that a sophomore slump does not exist. In this case, there is no need to consider the possibility of regression to the mean as a possible explanation for the phenomenon. Assuming, however, that the second-year POW value falls significantly below the other two values, it becomes appropriate to determine whether players' change in performance follows a specific pattern, prompting an investigation into the possibility that regression to the mean can explain players' decline in performance.

Such an investigation utilizes Campbell and Stanley's[24] suggestion that extreme scores generally regress toward the mean with respect to the correlation between data sets. The correlation between first year an second year POW values for ROTY award winners is .331. (Year 1 and Year 2 POW values appear in Appendix A.) So, according to Campbell and Stanley, second-year value should fall 33.1 percent closer to the overall mean POW value than first-year scores do. As suggested on *The Baseball Encyclopedia*'s official website, the theoretical mean POW score for all Major League Players falls at 0.[25] This mean value makes subsequent calculations rather simple. Based upon Equation 1,

$$Y_{reg} = Y_1 - .331(Y_1 - 0)$$

and, then, with simplification,

$$Y_{reg} = Y_1 - .331Y_1$$
$$Y_{reg} = Y_1(1 - .331)$$
$$Y_{reg} = .669(Y_1).$$

So, first-year POW scores that regress toward the mean of 0 should be 66.9 percent of their original value. Regressed first-year statistics, as well as actual first-year and second-year statistics for the players used in this analysis, appear in the Appendix.

If the difference in first-year and second-year performance results from regression to the mean, a t-test should indicate no significant difference between the regressed first-year mean score, 1.12, and the second-year mean POW score. The lack of a significant difference would suggest that some factor other than regression to the mean accounts for the pattern.

RESULTS

A simple comparison of the first and second-year POW values for all those receiving the Rookie of the Year Award and those receiving the Ex Post Facto Rookie of the Year Award verifies that players' levels of performance actually do decrease between their first and second years of play. [26] Table 1 contains mean POWs for these players' rookie and sophomore years as well as for the remainder of their careers. Adding each player's POW values for all years following the second and dividing this sum by the number of years in the Major League less two produces his remainder POW value.

	$\bar{X}_{POW}$	s_{POW}
rookie year	1.67	1.63
sophomore year	1.15	2.35
remainder of career	0.6785	2.39

Table 1 – Pow Summary Statistics
The mean POW score for Rookie of the Year award winners during the year that they won the award and for the following year, and the combined years after the second appears in the column labeled $\bar{X}_{POW}$. Standard deviations for these values appears in the column labeled s_{POW}.

A t-test indicates a significant difference between the two mean POW values (t=3.340, p=.001). These results indicate that the decline in performance between the first and second years of play is too great to have occurred by chance.

The question of whether the first or second year of play qualifies as the uncharacteristic one, then, arises. If the regression to the mean premise holds true, then statistical analyses should support three research hypotheses.

1. Players' rookie-year POW values should differ significantly from their POW values for the remainder of their careers.
2. No significant difference should exist between players' second-year POW values and their POW values for the remainder of their careers.
3. No significant difference should exist between players' second-year POW values and their regressed rookie-year POW values as predicted using Campbell and Stanley's formula.

The first two of these hypotheses suggest comparisons involving players' remainder POW values. Results of t-tests performed at the .05 significance level support both hypotheses. A significant difference exists between rookie-year POW values, as used in the previous analysis, and remainder POW values (t=5.497, p=.000). Also, second-year POW values, as used in the previous analysis, and remainder POW values do not differ significantly (t=2.399, p=.017).

These results indicate that players who perform at the Rookie of the Year caliber during their first years in the Major Leagues perform considerably poorer during their second years and do not return to their original high-performance levels. Second-year performance, then, is not unusually low, as suggested by the concept of the Sophomore Slump. Rather, a Freshman Fluke, characterized by a comparatively good performance during the players' first years in the Major Leagues, exists.

Assessment of the third hypothesis considers the relationship between first and second-year POW values in comparison to the relationship proposed by Campbell and Stanley's claims regarding regression to the mean.[27] Given the correlation of .33 between rookie-year and second-year POW values, the mean regressed rookie-year POW value (found by adding all players' regressed rookie-year values and dividing by the total sample size of 238) becomes 1.12. A t-test comparing this value to the actual second year mean POW value of 1.13 indicates no significant difference between the values (t=.257, p=.789), verifying the proposed relationship. The pattern and magnitude of differences between POW scores, therefore, qualify the change in player performance between the first and second years of play as nothing more than an issue of regression artifacts.

DISCUSSION

Based upon these results, claims of the relatively well-accepted occurrence of the Sophomore Slump lack validity. Such references have two outstanding flaws. First, they incorrectly imply that the second year of play, not the first, emerges as the uncharacteristic one. Second, the presence of a true regression to the mean situation suggests that the change in performance reflects a statistical, rather than an athletic phenomenon. This understanding reflects the fact that extreme scores for any repeated measure, whether referring to athletes' performance statistics, college students' GPAs, or stock gains, includes extreme values that, eventually regress toward the mean.

The appropriateness of honoring athletes based upon their performance during one particular year, then, must receive attention. At no other time during a player's career do baseball analysts criticize follow-up performance as they do by referring to the sophomore slump immediately after a standout first season. Players clearly have standout years in the middle and late years of their careers, as evidenced by the presentation of League Most Valuable Player (MVP) awards. MVP recipients may also fall victim to regression of the mean. It may be that an analysis of statistics for MVP players during the years they won their awards and the following year would also likely reflect regression to the mean.

A follow-up study could investigate this possibility. Uncovering a trend similar to the change described for ROTY winners would suggest that the outstanding players of *any* year, not just the rookie year, inevitably perform more poorly during the subsequent season. Further, it would imply that a player's talent or effort, alone, cannot explain changes in performance. As the old cliché states, "Things are not always as they seem."

REFERENCES

Albert, Jim. "Exploring Baseball Hitting Data: What About Those Breakdown Statistics?" *Journal of the American Statistical Association* 89, no 427 (1994): 1066-1074. http://dx.doi.org/10.1080/01621459.1994.10476844

Bennett, Tommy. "Rookie of the Year and the Mythical Sophomore Slump," Beyond the Boxscore, November 16, 2009. http://www.beyondtheboxscore.com/2009/11/16/1159607/rookie-of-the-year-and-the. Accessed 1/19/2010.

Bennett, Jay .M. & Fleuck, John A. "An Evaluation of Major League Baseball Offensive Performance Models," *The American Statistician* 37, no 1 (1983): 76-82.

Campbell, Donald T. and Stanley, Julian C. *Experimental and Quasi-Experimental Designs for Research.* Boston: Houghton Mifflin Company, 1963.

Gleeman, Aaron. "The Sophomore Slump" *The Hardball Times*. 2005. www.hardballtimes.com/main/article/the_sophomore_slump.

Palmer Pete & Gillette, Gary. *The Baseball Encyclopedia.* New York:Barnes and Noble, 2004.

Schall, Edward and Smith, Gary. "Do Baseball Players Regress Toward the Mean?" *The American Statistician* 54, no 4 (2000): 231-235.

Taylor, Jim & Cuave, Kenneth. "The Sophomore Slump Among Professional Baseball Players: Real or Imagined?" *International Journal of Sport Psychology* 25 (2004): 230-238.

24-7 Baseball. 2004. http://www.247baseball.com/story.php?storyid=7.

Walters, Steve. "The Sophomore Sink." *Sporting News* 229, no 16 (2005): 71.

APPENDIX 1

ROOKIE OF THE YEAR P.O.W. VALUES

	YEAR 1	YEAR 2	REGRESSED
Candy Cummings	1.1	2.1	0.7359
Paul Hines	0.4	0.4	0.2676
Tommy Bond	0.6	3.1	0.4014
George Bradley	0.3	6.2	0.2007
Pete Browning	4.6	2.2	3.0774
Arlie Latham	1.1	4.1	0.7359
Dave Orr	2.8	2.6	1.8732
Norm Baker	0.0	0.0	0.0
Matt Kilroy	0.1	7.2	0.0669
Mike Griffin	0.4	1.2	0.2676
Mickey Hughes	2.1	-1.6	1.4049
Jesse Duryea	5.7	1.8	3.8133
Phil Knell	1.1	2.2	0.7359
Benny Kauff	3.7	4.7	2.4753
Charley Jones	1.6	2.6	1.0704
Terry Larkin	1.6	1.9	1.0704
Abner Dalrymple	1.5	-1.4	1.0035
John O'Rourke	2.1	1.8	1.4049
Rover Connor	1.3	0.3	0.8697
Jim Whitney	2.1	2.8	1.4049
Mike Muldoon	0.6	-1.9	0.4014
Charlie Buffington	-1.4	-1.4	-0.9366
John Clarkson	1.4	7.3	0.9366
Ed Daily	0.0	-0.5	0.0
Mark Baldwin	2.2	0.4	1.4718
Ben Sanders	-0.1	-0.1	-0.0669
Billy Rhines	6.0	0.4	4.014
Bill Dahlen	1.4	3.6	0.9366
Nig Cuppy	3.7	0.6	2.4753
Heinie Reitz	0.8	1.8	0.5352
Win Mercer	3.8	0.3	2.5422
Bill Hoffer	4.6	3.9	3.0774
Gene DeMontreville	3.3	1.9	2.2077
Chick Stahl	0.8	0.2	0.5352
Elmer Flick	3.7	2.5	2.4753

Debra Wetcher-Hendricks

Jimmy Williams	4.7	0.2	3.1443
Jimmy Barrett	1.2	1.2	0.8028
Homer Smoot	1.0	-0.5	0.669
Harry Lumlye	2.1	1.4	1.4049
George Stone	2.8	5.9	1.8732
Jack Pfiester	2.1	0.1	1.4049
Nap Rucker	0.7	1.2	0.4683
George McQuillan	3.8	0.7	2.5422
Harry Gaspar	1.0	1.2	0.669
King Cole	3.1	0.8	2.0739
Grover Alexander	3.7	3.0	2.4753
Larry Cheney	2.2	2.6	1.4718
Jim Viox	-0.9	-1.4	-0.6021
Jeff Pfeffer	3.2	2.9	2.1408
Tom Long	0.8	-0.8	0.5352
Rogers Hornsby	3.7	7.8	2.4753
Leon Cadore	1.7	0.2	1.1373
Charlie Hollocher	1.0	2.2	0.669
Ocsar Tuero	-0.6	-0.2	-0.4014
Pat Duncan	-0.1	-0.3	-0.0669
Ray Grimes	0.8	3.2	0.5352
Hack Miller	0.6	0.1	0.4014
George Grantham	1.2	1.8	0.8028
Kiki Cuyler	2.7	3.6	1.8063
Jimmy Welsh	0.1	-0.3	0.0669
Paul Waner	2.6	3.9	1.7394
Lloyd Waner	-0.3	-0.2	-0.2007
Del Bissonette	2.4	-1.3	1.6056
Johnny Frederick	1.7	1.4	1.1373
Wally Berger	2.0	3.6	1.338
Paul Derringer	1.1	-0.8	0.7359
Dizzy Dean	2.5	1.7	1.6725
Frank Demaree	-1.4	0.0	-0.9366
Curt Davis	5.7	2.0	3.8133
Cy Blanton	4.4	1.0	2.9436
Johnny Mize	2.5	2.9	1.6725
Cliff Melton	3.4	-0.5	2.2746
Johnny Rizzo	0.9	-0.6	0.6021
Bob Bowman	-0.3	-0.5	-0.2007
Babe Young	0.2	0.0	0.1338
Elmer Riddle	3.5	-0.3	2.3415

Johnny Beasley	3.5	-0.9	2.3415
Lou Klein	1.4	0.1	0.9366
Bll Voiselle	2.0	-1.7	1.338
Ken Burkhart	2.5	0.6	1.6725
Del Ennis	2.9	-1.0	1.9401
Socks Seybold	1.4	1.6	0.9366
Addie Joss	2.0	2.5	1.338
Fred Glade	0.7	-0.3	0.4683
George Stone	2.8	5.9	1.8732
Claude Rossman	-0.3	-1.5	-0.2007
Glenn Liebhardt	1.4	0.7	0.9366
Ed Summers	1.1	0.6	0.7359
Frank Baker	2.2	2.3	1.4718
Russ Ford	4.2	3.0	2.8098
Vean Gregg	4.6	2.8	3.0774
Del Pratt	3.1	1.9	2.0739
Reb Russell	4.7	-0.2	3.1443
George Burns	0.2	-1.2	0.1338
Babe Ruth	0.0	0.0	0.0
Jim Bagby	1.0	3.9	0.669
Joe Harris	2.5	2.0	1.6725
Scott Perry	3.5	-0.2	2.3415
Dickie Kerr	1.5	0.8	1.0035
Bob Meusel	0.4	1.3	0.2676
Joe Sweell	3.0	2.6	2.007
Herman Pillette	2.9	0.0	1.9401
Homer Summa	-1.9	-2.3	-1.2711
Ike Boone	0.36	0.0	0.24084
Earle Combs	0.8	-1.2	0.5352
Tony Lazzeri	-0.8	3.5	-0.5352
Wilcy Moore	4.5	-0.9	3.0105
Ed Morris	1.5	-0.1	1.0035
Dale Alexander	2.4	-0.1	1.6056
Smead Jolley	-0.1	0.0	-0.0669
Joe Vsomik	0.1	1.6	0.0669
Johnny Allen	0.9	-0.08	0.6021
Bob Johnston	1.9	3.0	1.2711
Hal Trosky	2.5	-1.7	1.6725
Jake Powell	-0.1	-0.7	-0.0669
Joe DiMaggio	2.2	5.7	1.4718
Rudy York	1.3	3.5	0.8697

Debra Wetcher-Hendricks

Ken Keltner	-0.7	2.8	-0.4683
Ted Williams	4.1	4.0	2.7429
Walt Judnich	0.9	1.1	0.6021
Phil Rizzuto	2.4	4.6	1.6056
Billy Johnson	2.0	0.05	1.338
Joe Berry	3.2	1.8	2.1408
Dave Ferriss	3.0	1.8	2.007
Hoot Evers	-1.0	1.1	-0.669
Spec Shea	1.2	1.2	0.8028
Gene Bearden	5.5	-1.7	3.6795
Roy Sievers	1.4	-1.4	0.9366
Walt Dropo	1.3	-1.6	0.8697
Gil McDougald	2.1	1.9	1.4049
Harry Byrd	1.4	-3.4	0.9366
Harvey Kuenn	-1.0	1.5	-0.669
Bob Grim	0.5	-0.7	0.3345
Herb Score	2.3	4.9	1.5387
Luis Aparicio	-1.1	-1.3	-0.7359
Tony Kubek	-0.5	0.9	-0.3345
Albie Pearson	-0.09	-1.0	-0.06021
Bob Allison	-0.03	1.2	-0.02007
Ron Hansen	3.1	1.1	2.0739
Don Schwall	1.9	-2.3	1.2711
Tom Tresh	1.1	2.0	0.7359
Gary Peters	4.8	3.9	3.2112
Tony Oliva	3.2	2.7	2.1408
Curt Blefary	2.1	1.6	1.4049
Tommie Agee	1.4	-0.8	0.9366
Rod Carew	1.6	0.5	1.0704
Stan Bahnsen	2.4	-1.0	1.6056
Lou Piniella	0.5	-0.3	0.3345
Thurman Munson	3.9	1.8	2.6091
Chris Chambliss	-1.3	-1.4	-0.8697
Carlton Fisk	5.1	1.9	3.4119
Al Bumbry	1.5	-1.2	1.0035
Mike Hargrove	2.5	1.8	1.6725
Fred Lynn	4.1	20.0	2.7429
Mark Fidrych	4.4	1.4	2.9436
Eddie Murray	1.0	2.8	0.669
Lou Whitaker	2.9	3.1	1.9401
John Castino	0.2	2.0	0.1338

Alfredo Griffin	1.3	-1.3	0.8697
Joe Charboneau	1.2	-0.8	0.8028
Dave Righetti	2.2	0.2	1.4718
Cal Ripken, Jr.	1.6	7.1	1.0704
Ron Kittle	-0.8	-1.0	-0.5352
Alvin Davis	2.9	1.1	1.9401
Ozzie Guillen	-0.1	-0.1	-0.0669
Jose Canseco	0.1	1.0	0.0669
Mark McGwire	3.5	0.6	2.3415
Walt Weiss	0.9	-1.5	0.6021
Gregg Olson	2.8	2.6	1.8732
Sandy Alomar, Jr.	0.7	-0.5	0.4683
Chuck Knoblauch	0.4	1.0	0.2676
Pat Listach	1.8	-1.4	1.2042
Tim Salmon	3.1	1.4	2.0739
Bob Hamelin	1.4	-1.7	0.9366
Marty Cordova	2.1	1.1	1.4049
Derek Jeter	1.3	1.2	0.8697
Nomar Garciaparra	3.4	3.0	2.2746
Ben Grieve	0.4	0.6	0.2676
Carlos Beltran	0.0	-1.5	0.0
Kazuhiro Sasaki	1.9	1.3	1.2711
Ichiro Suzuki	3.2	2.1	2.1408
Eric Hinske	1.3	-0.3	0.8697
Jackie Robinson	0.3	2.0	0.2007
Alvin Dark	0.04	-0.8	0.02676
Don Newcombe	2.9	1.9	1.9401
Sam Jethroe	0.06	2.0	0.04014
Willie Mays	1.2	0.4	0.8028
Joe Black	3.3	-1.0	2.2077
Jim Gilliam	1.9	-0.3	1.2711
Wally Moon	-1.4	-0.5	-0.9366
Bill Virdon	-1.5	-0.7	-1.0035
Frank Robinson	1.6	3.5	1.0704
Jack Sanford	1.8	-1.0	1.2042
Orlando Cepeda	0.9	1.6	0.6021
Willie McCovey	2.0	0.5	1.338
Frank Howard	-0.7	-0.1	-0.4683
Billy Williams	-0.4	0.9	-0.2676
Ken Hubbs	-1.6	1.3	-1.0704
Pete Rose	-0.6	-1.7	-0.4014

Debra Wetcher-Hendricks

Dick Allen	5.9	4.0	3.9471
Jim Lefebvre	1.8	3.1	1.2042
Tommy Helms	-2.0	-1.6	-1.338
Tom Seaver	2.4	3.0	1.6056
Johnny Bench	2.5	3.2	1.6725
Ted Sizemore	0.8	0.3	0.5352
Carl Morton	2.0	-3.4	1.338
Earl Williams	0.4	-0.6	0.2676
Jon Matlack	2.7	1.8	1.8063
Gary Matthews	0.9	0.6	0.6021
Bake McBride	1.2	0.7	0.8028
John Montefusco	2.1	2.32	1.4049
Butch Metzger	0.8	0.2	0.5352
Pat Zachry	1.4	-1.1	0.9366
Andre Dawson	1.0	0.7	0.669
Bob Horner	1.8	0.8	1.2042
Rick Sutcliffe	0.0	1.1	0.0
Steve Howe	1.2	0.8	0.8028
Fernando Valenzuela	2.6	2.4	1.7394
Steve Sax	1.3	-1.5	0.8697
Darryl Strawberry	1.4	1.7	0.9366
Dwight Gooden	2.9	7.5	1.9401
Vince Coleman	0.0	-2.1	0.0
Todd Worrell	3.6	3.1	2.4084
Benito Santiago	-0.4	1.4	-0.2676
Chris Sabo	2.0	-0.5	1.338
Jerome Walton	-0.3	-1.3	-0.2007
David Justice	1.5	1.7	1.0035
Jeff Bagwell	1.9	2.6	1.2711
Eric Karros	-0.3	0.0	-0.2007
Mike Piazza	5.3	1.9	3.5457
Raul Mondesi	0.9	2.3	0.6021
Hideo Nomo	2.1	1.98	1.4049
Todd Hollandsworth	0.06	-0.6	0.04014
Scott Rolen	4.0	5.1	2.676
Scott Williamson	4.2	1.8	2.8098
Rafael Furcal	1.8	-0.1	1.2042
Albert Pujols	5.4	3.6	3.6126
Jason Jennings	1.0	-0.04	0.669

ENDNOTES

[1]James Taylor and Kenneth Cuave, "The Sophomore Slump among Professional Baseball Players: Real or Imagined?," International Journal of Sport Psychology 25 (2004): 230 and Aaron Gleeman, "The Sophomore Slump," *The Hardball Times* (2004), 1. www.hardballtimes.com/main/article/sophomore_slumps.
[2] Bennett, Tommy. "Rookie of the Year and the Mythical Sophomore Slump" *Beyond the Box Score* (November 16, 2009), 1
[3] Pete Palmer and Gary Gilette. *The Baseball Encyclopedia.* (New York: Barnes and Noble, 2004.) 1667.
[4] Bennett, 1.
[5] Gleeman, 1. and Steve Walters. "The Sophomore Sink." *Sporting News* 229, no. 16 (2005),1.
[6] Taylor and Cuave, 230.
[7] Gleeman, 1-4.
[8] Walters,1, and Gleeman, 3.
[9] Taylor and Cuave, 232.
[10] Jim Albert. "Exploring Baseball Hitting Data: What About Those Breakdown Statistics?" *Journal of the American Statistical Association* 89, no 427 (1994): 1067.
[11] Walters, 71.
[12] Galton, Sir Francis. "Regression toward Mediocrity in Hereditary Stature." *Journal of the Anthropological Institute* 15 (1886): 246-253
[13] Edward Schall and Gary Smith. "Do Baseball Players Regress Toward the Mean?" *The American Statistician* 54, no 4 (2000): 233.
[14]Donald T. Campbell and Julian C. Stanley. *Experimental and Quasi-Experimental Designs for Research.* (Boston: Houghton Mifflin Company, (1963), 11.
[15] Campbell and Stanley. 11.
[16] Jay M. Bennett and John A. Fleuck. "An Evaluation of Major League Baseball Offensive Performance Models." *The American Statistician* 37, no 1 (1983): 76.
[17] Bennett and Fleuck, 76.
[18] Bennett and Fleuck, 76.
[19] Bennett and Fleuck, 77.
[20] Gleeman, 1, and Walters, 71.
[21] Taylor and Cuave, 233.
[22] Palmer and Gilette, 1696.
[23] Pete Palmer and Gary Gilette, *24-7 Baseball*, 2004, 1. http://www.247baseball.com/story.php?storyid=7.

[24] Campbell and Stanley.
[25] 24-7 Baseball, http://www.247baseball.com/. 1.
[26] Palmer and Gilette,1668.
[27] Campbell and Stanley, 11.