Vanderbilt Law Review

Volume 72
Issue 5 *Issue 5 - October 2019*

Article 4

10-2019

You Get What You Pay For: An Empirical Examination of the Use of MTurk in Legal Scholarship

Adriana Z. Robertson

Albert H. Yoon

ESSAY

You Get What You Pay For: An Empirical Examination of the Use of MTurk in Legal Scholarship

*Adriana Z. Robertson**
*Albert H. Yoon***

In recent years, legal scholars have come to rely on Amazon's Mechanical Turk ("MTurk") platform to recruit participants for surveys and experiments. Despite MTurk's popularity, there is no generally accepted methodology for its use in legal scholarship, and many questions remain about the validity of data gathered from this source. In particular, little is known about how the compensation structure affects the performance of respondents recruited using MTurk.

This Essay fills both of these gaps. We develop an experiment and test the effect of various compensation structures on performance along two dimensions: effort and attention. We find that both the level and the structure of the compensation scheme have substantial effects on the performance of MTurk workers, and that these effects differ across question types. We then propose a series of best practices for scholars to follow in conducting research

* Assistant Professor, University of Toronto Faculty of Law & Rotman School of Management, adriana.robertson@utoronto.ca.

** Professor and Chair in Law and Economics, University of Toronto Faculty of Law, albert.yoon@utoronto.ca. This research was made possible by generous funding from the James M. Tory Fund. We are grateful for helpful comments from Vikram Amar, Kenworthey Bilz, Vincent Chiao, Mitu Gulati, Paul Heald, Robert Lawless, Anthony Niblett, Jennifer Robbennolt, and Arnold Weinrib, as well as workshop participants at the Canadian Economics Association annual conference and the University of Illinois College of Law. All remaining errors are our own.

using MTurk. Adoption of these guidelines will improve both the transparency and the robustness of research conducted using this platform.

INTRODUCTION

Legal scholars have long been criticized for their propensity to navel-gaze.[1] The degree to which legal scholarship is useful or even relevant to anyone outside a narrow slice of legal academia has been questioned by both academics[2] and judges, including Chief Justice Roberts of the U.S. Supreme Court.[3] Perhaps partly in response to these concerns, there has been a move in recent years to inject some "real-world" grounding into legal scholarship. One common example of this is the rise of the use of Amazon's Mechanical Turk ("MTurk") platform as a means of tethering academic ideas to the lives, beliefs, and reactions of ordinary individuals.

Over the last ten years, law reviews and other scholarly legal publications have published dozens of articles that rely on data gathered using MTurk.[4] Some of these applications are purely survey-based—where the primary objective is to collect information about the thoughts, perceptions, and beliefs of ordinary individuals—while others

---

1. For perhaps the first instance of this, see Fred Rodell, *Goodbye to Law Reviews*, 23 VA. L. REV. 38, 43 (1937):

> Law review writers seem to rank among our most adept navel-gazers. When they are not busy adding to and patching up their lists of cases and their farflung lines of logic, so that some smart practising lawyer can come along and grab the cases and the logic without so much as a by-your-leave, they are sure to be found squabbling earnestly among themselves over the meaning or content of some obscure principle that nine judges out of ten would not even recognize if it hopped up and slugged them in the face.

For a more recent article surveying the literature and proposing "an incremental contribution," see generally Michael Klinger, *Escape from the Navel-Gazing Academy: A Modest Proposal for Student-Edited Legal Scholarship*, 5 U.C. IRVINE L. REV. 179 (2015).

2. *See, e.g.*, Rodell, *supra* note 1, at 44; LawProfBlawg, *Why Do Law Professors Write Law Review Articles?*, ABOVE LAW (May 9, 2017, 2:00 PM), https://abovethelaw.com/2017/05/why-do-law-professors-write-law-review-articles/ [https://perma.cc/7XMR-C68J] (asking why law professors write law review articles); *see also* Stephen Bainbridge, *"Why Do Law Professors Write Law Review Articles?" Is the Wrong Question*, PROFESSORBAINBRIDGE.COM (May 11, 2017, 2:29 PM), https://www.professorbainbridge.com/professorbainbridgecom/2017/05/why-do-law-professors-write-law-review-articles-is-the-wrong-question.html [https://perma.cc/S7QC-UHRD] (responding to LawProfBlawg).

3. Debra Cassens Weiss, *Law Prof Responds After Chief Justice Roberts Disses Legal Scholarship*, A.B.A. J. (Jul. 7, 2011, 10:29 AM), http://www.abajournal.com/news/article/ law_prof_responds_after_chief_justice_roberts_disses_legal_scholarship [https://perma.cc/3XSH-BLFZ]. Weiss quotes Chief Justice Roberts as having said:

> Pick up a copy of any law review that you see, and the first article is likely to be, you know, the influence of Immanuel Kant on evidentiary approaches in 18th Century Bulgaria, or something, which I'm sure was of great interest to the academic that wrote it, but isn't of much help to the bar.

4. *See* discussion *infra* notes 25–40 and accompanying text.

are more experimental in nature.[5] There are compelling reasons for MTurk's popularity: it is both faster and cheaper than most survey or experimental techniques, often allowing researchers to obtain hundreds of results in a few hours for only a few hundred dollars.[6] Despite its relatively recent origins,[7] MTurk has increasingly gained acceptance among legal scholars, and articles relying on MTurk data have been published in some of the leading law reviews.[8]

Questions remain, however, about the quality of the data obtained through MTurk. We believe that two concerns are particularly acute. The first relates to the compensation offered to MTurk participants. While part of what makes MTurk an attractive platform is precisely the fact that it is far cheaper than other available options, this advantage may come with its own nonpecuniary costs. For example, individuals who are being paid substantially below minimum wage may not be particularly invested in the questions and may provide answers that do not reflect their true preferences or beliefs. Moreover, given the extensive literature on the sensitivity of individual behavior to incentives,[9] the *structure* of any compensation offered, in addition to its *level*, is likely to have important implications for participant behavior. In particular, we distinguish between questions and tasks that require individuals to exert effort (that is, to think hard about their answers) and those that require them to pay attention (that is, to read the text of the question carefully). We contend that the optimal compensation structure may vary across these question types.

---

5. *See* discussion *infra* Part II.

6. *See* discussion *infra* Part II.

7. *See* discussion *infra* Part II.

8. Examples of such journals include the *Yale Law Journal*, see, for example, Elizabeth Ingriselli, *Mitigating Jurors' Racial Biases: The Effects of Content and Timing of Jury Instructions*, 124 YALE L.J. 1690 (2015); the *Columbia Law Review*, see, for example, Ryan Calo & Alex Rosenblat, *The Taking Economy: Uber, Information, and Power*, 117 COLUM. L. REV. 1623 (2017); the *Georgetown Law Journal*, see, for example, Justin Sevier, *Popularizing Hearsay*, 104 GEO. L.J. 643 (2016); and the *New York University Law Review*, see, for example David A. Hoffman, *From Promise to Form: How Contracting Online Changes Consumers*, 91 N.Y.U. L. REV. 1595 (2016).

9. *See, e.g.*, N. GREGORY MANKIW, PRINCIPLES OF ECONOMICS 7 (4th ed. 2007) (listing ten key principles of economics, the fourth of which is "people respond to incentives"). Most recently, economists have established that individuals will work longer in response to financial incentives to delay retirement. *See, e.g.*, Kadir Atalay & Garry F. Barrett, *The Impact of Age Pension Eligibility Age on Retirement and Program Dependence: Evidence from an Australian Experiment*, 97 REV. ECON. & STAT. 71 (2014); Courtney C. Coile & Jonathan Gruber, *Future Social Security Entitlements and the Retirement Decision*, 89 REV. ECON. & STAT. 234 (2007).

Our second, and more subtle, concern relates to the way in which the findings of studies relying on MTurk are often presented in law reviews. At present, there appear to be no widely accepted norms regarding what information authors are expected to provide about their empirical methodologies. While the amount of disclosure varies widely across articles, the vast majority provide very little discussion of how the survey or experiment was actually conducted. While there may be good reasons why authors chose to limit this discussion—including a desire not to clutter the body of the article with details that many readers may view as extraneous to the article's main argument—this opacity makes it very difficult for other scholars to interpret or evaluate the results of these studies, limiting their potential impact.

Many law review articles use MTurk to ask questions that are subjective in nature—questions about the respondents' opinions, their feelings about a particular topic, or questions for which there is no obviously correct answer. Even in this context, to the extent that the researcher cares about collecting responses from participants who have paid attention to the questions, the level of compensation may matter. Even if an answer is not wrong per se, answers from inattentive participants may introduce noise or bias to data.

In this Essay, we seek to address these concerns. Using an experimental methodology that we discuss in detail in Section II.A, we first test the implications of varying the compensation structure on the performance of participants recruited using MTurk. We find that both the level and the structure of the offered compensation have substantial implications for performance, and that these implications vary depending on whether the tasks in question primarily require *effort* or *attention*. We discuss these results in Sections II.B and II.C.

Drawing on our experimental results, we then propose a list of best practices that scholars should employ when using MTurk for legal scholarship. These proposals are straightforward to implement, and we believe that doing so would dramatically improve the reliability of articles based on MTurk data.

The remainder of this Essay proceeds as follows. In Part I, we introduce the MTurk platform and discuss its extensive use in legal scholarship. In Part II, we discuss our experiment, including our experimental design and results, and then interpret those results. In Part III, we propose a series of best practices for the use of MTurk in legal scholarship, in light of recent uses of the platform.

I. BACKGROUND AND LITERATURE REVIEW

A. About MTurk

Its popularity among researchers notwithstanding, MTurk was originally designed for a very different purpose. Amazon created MTurk in 2005 as a way to recruit people to help the company identify duplicate pages on its website.[10] At the time, humans could perform these types of tasks more easily than machines. Curiously, its etymology comes from the eighteenth-century chess-playing automaton that was later revealed to be a hoax: a human chess master controlled the machine.[11]

Shortly after MTurk's creation, Amazon opened it up to third parties and the platform has effectively turned into a broker between prospective hirers and prospective workers.[12] Hirers, known as "requesters," can post tasks, known as "Human Intelligence Tasks" ("HITs"), on the platform, along with stated compensation, after which each eligible worker can decide which HITs to accept.[13] Upon completion of the HIT, MTurk handles the payment to the workers on behalf of the requester.[14] Because all workers who register to use the platform are asked to provide demographic information, requesters can also set eligibility criteria for their workers.[15] As the broker between requesters and workers, MTurk takes a commission of the stated compensation: it began with 10% and subsequently raised that to 20% in 2015, charging an additional 20% for tasks involving more than ten people.[16]

Since its creation, MTurk has grown dramatically. By 2007, over one hundred thousand MTurk workers participated from over one hundred countries.[17] In 2017, Amazon reported that total workers had

---

10. *See* Jason Pontin, *Artificial Intelligence, With Help from the Humans*, N.Y. TIMES (Mar. 25, 2007), https://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html [https://perma.cc/C5ZN-HPHX].

11. *See id.*

12. For a description of the registration process, see Adam J. Berinsky et al., *Evaluating Online Labor Markets for Experimental Research: Amazon.com's Mechanical Turk,* 20 POL. ANALYSIS 351, 352–53 (2012).

13. *Id.* at 352.

14. *Id.* at 353.

15. *See id.* at 352.

16. *See* Jillian D'Onfro, *Amazon Just Increased Prices on a Service You've Probably Never Heard of, and Some Researchers Are Really Upset*, BUS. INSIDER (June 23, 2015, 8:39 PM), https://www.businessinsider.com/amazon-mechanical-turk-price-changes [https://perma.cc/LW8R-ZWN5].

17. *See* Pontin, *supra* note 10.

exceeded five hundred thousand.[18] Workers are heavily concentrated in two countries: the United States, with 75%, and India, with 16%.[19]

B. MTurk in Legal Scholarship

Despite its origins in more traditional labor markets, MTurk soon became popular among academics. Traditionally, when conducting research experiments with human subjects, academics interacted personally with those subjects. Members of the general population were invited to participate in some studies, but often researchers drew on undergraduate students. Reliance on this nonrepresentative population prompted concern among some researchers.[20] MTurk afforded researchers a way to address at least three central limitations of traditional research: it made recruiting participants cheaper, made it easier to draw from a diverse population,[21] and made it possible to conduct the experiment in a shorter time frame.

Social science scholars were early adopters of MTurk in their studies. For example, political scientists conducted experiments with MTurk workers to determine factors that influence voting patterns.[22] Economists surveyed MTurk workers for their perception of market

---

18. *See* Djellel Difallah, Elena Filatova & Pano Ipeirotis, Demographics and Dynamics of Mechanical Turk Workers, Address before WSDM 2018: The Eleventh ACM International Conference on Web Search and Data Mining § 1 n.2 (Feb. 2018), https://ipeirotis.org/wp-content/uploads/2017/12/wsdmf074-difallahA.pdf [https://perma.cc/47F2-ZR6N] (noting their own findings that the number of workers is in the 100,000–200,000 range).

19. *See id.* § 3.2 (reporting that aside from Canada (1.1%), the remaining countries each comprise less than 1% of workers).

20. *See* Michael Buhrmester et al., *Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?* 6 PERSP. PSYCHOL. SCI. 3, 4 (2011) ("Commentators have long lamented the heavy reliance on American college samples.").

21. *See* Erin C. Cassese et al., *Socially Mediated Internet Surveys: Recruiting Participants for Online Experiments*, 46 PS: POL. SCI. & POL. 775, 776 (2013) (describing greater diversity of MTurk participation than the typical student convenience sample).

22. *See, e.g.*, Stig Hebbelstrup Rye Rasmussen, *Cognitive Ability Rivals the Effect of Political Sophistication on Ideological Voting*, 69 POL. RES. Q. 773 (2016) (finding through MTurk workers a positive correlation of cognitive ability on ideological voting); Lars J. Lefgren et al., *Effort, Luck, and Voting for Redistribution*, 143 J. PUB. ECON. 89, 95 (2016) (finding that over a quarter of MTurk workers were willing to vote against their self-interest and favor groups who exhibited more effort).

transactions.[23] Psychologists, surveying MTurk workers, explored the origins of social comparison.[24]

As MTurk gained traction in the social sciences, its use spread to legal scholarship. Based on a recent search using Lexis Advance, we identified no fewer than ninety-eight articles involving the use of MTurk participants in either a survey or experimental context.[25] Within the past five years alone, legal scholars have used MTurk to study administrative regulation,[26] alternative dispute resolution,[27] bankruptcy,[28] constitutional law,[29] consumer protection,[30] contracts,[31]

---

23. *See, e.g.*, Sandro Ambuehl et al., *More Money, More Problems? Can High Pay be Coercive and Repugnant?,* 105 AM. ECON. REV. 357, 359 (2015) (finding that MTurk workers viewed in-kind incentives, rather than money, as more ethical).

24. *See, e.g.*, Matthew Baldwin & Thomas Mussweiler, *The Culture of Social Comparison*, PNAS (Sept. 25, 2018), https://www.pnas.org/content/115/39/E9067 [https://perma.cc/S4AK-DP33] (finding that social comparison plays an essential role in the development of social life).

25. To identify the universe of relevant articles, we ran a search on Lexis Advance for the term "MTurk" on December 19, 2018. This search returned 150 hits in the category of "Secondary Materials," which we downloaded and reviewed. Of these 150 articles, 98 involved an original MTurk survey conducted by the author(s).

26. *See, e.g.*, Victor D. Quintanilla, *Taboo Procedural Tradeoffs: Examining How the Public Experiences Tradeoffs Between Procedural Justice and Cost*, 15 NEV. L.J. 882 (2015) (using MTurk to evaluate views on procedural regulation); Edward H. Stiglitz, *Cost-Benefit Analysis and Public Sector Trust*, 24 SUP. CT. ECON. REV. 169 (2016) (finding through MTurk surveys that agency use of cost-benefit analysis increases public sector trust).

27. *See, e.g.*, Victor D. Quintanilla & Alexander B. Avtgis, *The Public Believes Predispute Binding Arbitration Clauses Are Unjust: Ethical Implications for Dispute-System Design in the Time of Vanishing Trials*, 85 FORDHAM L. REV. 2119 (2017) (using MTurk to observe how the public perceives taboo procedural tradeoffs).

28. *See, e.g.*, Dov Cohen et al., *Opposite of Correct: Inverted Insider Perceptions of Race and Bankruptcy*, 91 AM. BANKR. L.J. 623 (2017) (using MTurk to determine perceptions on differences in base rates of chapter 13 bankruptcy filings).

29. *See, e.g.*, Eileen Braman, *Exploring Citizen Assessments of Unilateral Executive Authority*, 50 LAW & SOC'Y REV. 189 (2016) (using MTurk to evaluate the interaction between constitutional considerations and democratic context when evaluating executive authority); David S. Cohen & Jeffrey B. Bingenheimer, *Abortion Rights and the Largeness of the Fraction 1/6*, 164 U. PA. L. REV. ONLINE 115 (2016) (using MTurk to evaluate how the public perceives circuit court doctrine on abortion); Daniel J. Hemel & Lisa Larrimore Ouellette, *Public Perceptions of Government Speech*, 2017 SUP. CT. REV. 33 (2017) (using MTurk to determine how the public perceives monuments as government speech); Jonathon W. Penney, *Chilling Effects: Online Surveillance and Wikipedia Use*, 31 BERKELEY TECH. L.J. 117 (2016) (surveying MTurk workers on their perceptions of privacy through Wikipedia use).

30. *See, e.g.*, Joel R. Reidenberg et al., *Disagreeable Privacy Policies: Mismatches Between Meaning and Users' Understanding*, 30 BERKELEY TECH. L.J. 39 (2015) (surveying MTurk workers for their understanding of privacy policies).

31. *See, e.g.*, David A. Hoffman & Tess Wilkinson-Ryan, *The Psychology of Contract Precautions*, 80 U. CHI. L. REV. 395 (2013) (surveying MTurk workers to understand how perceived finality of the contract influences parties' self-protective behavior).

criminal law,[32] election law,[33] environmental law,[34] evidence,[35] intellectual property,[36] labor and employment,[37] securities,[38] tax,[39] and torts.[40]

As it has become more widely used, MTurk has garnered its share of criticism. Detractors suggest that MTurk is exploitative, describing

---

32. *See, e.g.*, Jane Bambauer, *Defending the Dog*, 91 OR. L. REV. 1203 (2013) (surveying MTurk workers for their views toward narcotics dogs as a means of enforcement); Michael D. Cicchini & Lawrence T. White, *Testing the Impact of Criminal Jury Instructions on Verdicts: A Conceptual Replication*, 117 COLUM. L. REV. ONLINE 22 (2017) (surveying MTurk workers to identify sensitivity to jury instructions); Brandon L. Garrett & Gregory Mitchell, *Forensics and Fallibility: Comparing the Views of Lawyers and Jurors*, 119 W. VA. L. REV. 621 (2016) (using MTurk to gauge how juries perceive forensic evidence in criminal trials); Ingriselli, *supra* note 8 (surveying MTurk workers to evaluate juror bias against criminal defendants); Paul H. Robinson, Geoffrey P. Goodwin & Michael D. Reisig, *The Disutility of Injustice*, 85 N.Y.U. L. REV. 1940 (2010) (using MTurk to understand determinants that influence the public's perceptions of moral credibility in the criminal justice system); Francis X. Shen, *Minority Mens Rea: Racial Bias and Criminal Mental States*, 68 HASTINGS L.J. 1007 (2017) (surveying MTurk workers to understand jury views toward mens rea); Roseanna Sommers, *Will Putting Cameras on Police Reduce Polarization?*, 125 YALE L.J. 1304 (2016) (using MTurk to evaluate public perceptions of police cameras); Avani Mehta Sood, *Cognitive Cleansing: Experimental Psychology and the Exclusionary Rule*, 103 GEO. L.J. 1543 (2013) (surveying MTurk workers for divergence between Fourth Amendment jurisprudence and societal norms).

33. *See, e.g.*, Michael J. Nelson, *Is There a Silver Lining? Dark Money and Support for State Courts*, 67 DEPAUL L. REV. 187 (2018) (using MTurk to understand public attitudes toward dark money in elections).

34. *See, e.g.*, Cass R. Sunstein et al., *How People Update Beliefs About Climate Change: Good News and Bad News*, 102 CORNELL L. REV. 1431 (2017) (using MTurk to understand how the public updates beliefs on global warming); Michael P. Vandenbergh & Kaitlin T. Raimi, *Climate Change: Leveraging Legacy*, 42 ECOLOGY L.Q. 139 (2015) (surveying MTurk to understand public perception of legacy concerns to address climate change).

35. *See, e.g.*, Brandon Garrett & Gregory Mitchell, *How Jurors Evaluate Fingerprint Evidence: The Relative Importance of Match Language, Method Information, and Error Acknowledgment*, 10 J. EMPIRICAL LEGAL STUD. 484 (2013) (using MTurk to understand juror attitudes toward fingerprint identification in criminal trials); Justin Sevier, *Evidentiary Trapdoors*, 103 IOWA L. REV. 1155 (2018) (surveying MTurk workers to explore how jurors perceive inconsistencies between the plain reading of a rule and its application); Sevier, *supra* note 8 (using MTurk to understand how jurors respond to hearsay evidence); Justin Sevier, *Testing Tribe's Triangle: Juries, Hearsay, and Psychological Distance*, 103 GEO. L.J. 879 (2015) (same).

36. *See, e.g.*, Christopher Buccafusco & Paul J. Heald, *Do Bad Things Happen When Works Enter the Public Domain?: Empirical Tests of Copyright Term Extension*, 28 BERKELEY TECH. L.J. 1 (2013) (surveying MTurk to assess views on extension of copyright terms); Gregory N. Mandel et al., *Intellectual Property Law's Plagiarism Fallacy*, 2015 BYU L. REV. 915 (surveying MTurk workers for views on plagiarism in copyright cases); David L. Schwartz & Christopher B. Seaman, *Standards of Proof in Civil Litigation: An Experiment from Patent Law*, 26 HARV. J.L. & TECH. 429 (2013) (using MTurk to evaluate jury instructions' effect on the presumption of validity in patent cases); Christopher Jon Sprigman et al., *What's a Name Worth?: Experimental Tests of the Value of Attribution in Intellectual Property*, 93 B.U. L. REV. 1389 (2013) (using MTurk to understand public perceptions of attribution).

37. *See, e.g.*, Joni Hersh & Jennifer Bennett Shinall, *Something to Talk About: Information Exchange Under Employment Law*, 165 U. PA. L. REV. 49 (2016) (surveying MTurk workers to see how the disclosure of family status affects employment prospects).

it as resembling what a "labor market would look like without minimum wages or labor law protections."[41] The majority of MTurk workers earn considerably less than the federal minimum wage,[42] yet a quarter of MTurk workers identify this work as their primary source of income.[43] One early critic described MTurk as a "virtual sweatshop."[44] Moreover, because they are, by design, independent contractors, MTurk workers do not receive any benefits that employees typically enjoy, such as sick leave, vacation time, parental leave, or an employer-sponsored retirement plan. While competitors professing to treat workers with greater dignity have emerged, [45] MTurk remains the market leader.[46]

Skeptics within academia have also pointed to more methodological concerns about relying on MTurk. Some point to potential overexposure of MTurk workers to academic studies: for example, scholars have found that the median MTurk worker had participated in several studies a week, and hundreds of studies over the

---

38. *See, e.g.*, Jill E. Fisch & Tess Wilkinson-Ryan, *Why Do Retail Investors Make Costly Mistakes? An Experiment on Mutual Fund Choice*, 162 U. PA. L. REV. 605 (2014) (surveying MTurk workers on the effect of information on individuals' retirement planning).

39. *See, e.g.*, Ian Ayres, *Voluntary Taxation and Beyond: The Promise of Social-Contracting Voting Mechanisms*, 19 AM. L. & ECON. REV. 1 (2017) (using MTurk to evaluate public perception of voluntary taxation); Emily Satterthwaite, *Can Audits Encourage Tax Evasion?: An Experimental Assessment*, 20 FLA. TAX REV. 1 (2016) (using MTurk to determine the effect of random audits on tax compliance).

40. *See, e.g.*, John Campbell et al., *Countering the Plaintiff's Anchor: Jury Simulations to Evaluate Damages Arguments*, 101 IOWA L. REV. 543 (2016) (surveying MTurk workers for the effect of anchoring on damages awards); Justin Sevier, *Vicarious Windfalls*, 102 IOWA L. REV. 651 (2017) (using MTurk to evaluate how jurors view vicarious liability when assessing damages).

41. Nancy Folbre, *The Unregulated Work of Mechanical Turk*, N.Y. TIMES: ECONOMIX (Mar. 18, 2003, 6:00 AM), https://economix.blogs.nytimes.com/2013/03/18/the-unregulated-work-of-mechanical-turk [https://perma.cc/KRU3-K38B].

42. *See* Paul Hitlin, *Turkers in this Canvassing: Young, Well-Educated, and Frequent Users*, PEW RES. CTR. (July 11, 2016), http://www.pewinternet.org/2016/07/11/turkers-in-this-canvassing-young-well-educated-and-frequent-users/ [https://perma.cc/E7PD-349V] (reporting that most MTurk workers earn less than $5 an hour, compared with the $7.25 federal minimum wage).

43. *Id.*

44. Katharine Mieszkowski, *I Make $1.45 a Week and I Love It*, SALON (July 24, 2006, 5:00 PM), https://www.salon.com/2006/07/24/turks_3/ [https://perma.cc/UQ6H-E9VM].

45. For example, Daemo was created to serve as an academic crowdwork environment offering a constitution to dynamically reflect worker interests and empower more equitable crowdsourcing. *See As Amazon's 'Mechanical Turks' Push for Better Conditions, a New Platform Emerges to Court Them*, GARTNER (Aug. 24, 2017, 4:32 PM), https://www.cebglobal.com/talentdaily/as-amazons-mechanical-turks-push-for-better-conditions-a-new-platform-emerges-to-court-them/ [https://perma.cc/QYZ2-6BS3].

46. *See* Miranda Katz, *This Startup is Challenging Mechanical Turk—on the Blockchain*, WIRED (Feb. 23, 2018, 7:00 AM), https://www.wired.com/story/this-startup-is-challenging-mechanical-turkon-the-blockchain/ [https://perma.cc/PYG8-4T69] (describing MTurk as the "Xerox of crowdwork").

course of his or her lifetime.[47] This overexposure runs the risk of compromising the integrity of academic surveys. More recently, researchers found that survey responses using MTurk were populated with nonsensical responses to open-ended questions and multiple responses from duplicate GPS locations.[48] This trend, which suggests participation by bots rather than humans, challenges the legitimacy of MTurk as a survey research instrument.[49]

While we recognize the importance of these issues, we examine MTurk from a different perspective. We focus on two very specific concerns that we think are particularly relevant to the use of MTurk by legal scholars and in legal publications. First, our review of the current legal literature revealed that studies rarely explained the reasoning behind the compensation structure offered to participants. Indeed, roughly half of the articles we reviewed did not even disclose the amount of compensation that MTurk workers received, and less than one-third disclosed the duration of the assignment, which would be necessary in order to estimate an imputed hourly rate.[50]

Given both the concerns around exploitation[51] and the well-documented ways in which incentives can affect human behavior, we designed an experiment to test the relationship between compensation

---

47. *See* David G. Rand et al., *Social Heuristics Shape Intuitive Cooperation*, NATURE COMM. 4 (Apr. 22, 2014), http://static1.squarespace.com/static/51ed234ae4b0867e2385d879/t/5356700be4b0b8b008782885/1398173707675/social-heuristics-shape-intuitive-cooperation.pdf [https://perma.cc/3E2L-G6US].

48. *See* Emily Dreyfuss, *A Bot Panic Hits Amazon's Mechanical Turk*, WIRED (Aug. 17, 2018, 11:38 AM), https://www.wired.com/story/amazon-mechanical-turk-bot-panic/ [https://perma.cc/273Q-VNBG].

49. Given the design of our survey, described in detail below, we think it is unlikely that our results are vulnerable to bot participation. Our questions were open ended, and only a very small number of the responses we received were anything other than a facially plausible answer. An examination of IP addresses revealed 962 unique IP addresses out of one thousand respondents. The same IP address is never in the data more than four times (930 appear once, twenty-seven appear twice, four appear three times, and one appears four times). These relatively small numbers of repeat IP addresses are more plausibly explained by multiple people with different accounts working or residing at the same IP address, or perhaps, by a single human holding more than one account.

50. For these purposes, there are at least two different versions of the duration of an assignment, both of which would be relevant for interpreting the implied hourly rate. The first is the amount of time that the researcher *says* that the assignment will take to complete. It is common for researchers to include in the HIT description an estimate of the time required to complete the assignment. Workers may interpret this stated duration as a signal of how long they are expected to spend on the assignment. Moreover, given that this information is provided to potential workers ex ante, it is plausible that they might use it to estimate their implied hourly compensation before deciding whether to accept the HIT. The second is the amount of time workers *actually* spent on the assignment, which is necessary to compute an ex post implied hourly rate.

51. *See* discussion *supra* notes 41–44 and accompanying text.

structures and the performance of MTurk participants. In doing so, we designed the experiment to allow us to distinguish tasks that require individuals to exert *effort* from those that require them to pay *attention*, terms which we discuss in detail in the next Part.[52]

Our second concern relates to the way in which studies relying on MTurk are presented in law reviews. Our review of the current literature makes it clear that there are no widely accepted norms about what kind of information law reviews expect authors to provide regarding their empirical methodologies. In most cases, even if there is a mention of the level of compensation offered, the articles spend very little time discussing how the survey or experiment was actually conducted, making it difficult for readers to evaluate, or even interpret, the results. In order to overcome this difficulty, we propose a series of guidelines for law journals and legal scholars to adopt.

II. OUR EXPERIMENT

Given our review of the use of MTurk in legal scholarship,[53] the goal of our experiment is simple: to test the effects of different compensation structures on the performance of MTurk workers, both in terms of their effort and attention.

A. Experimental Design

To test the effect of compensation structures on the performance of MTurk workers, we administered the same questionnaire four times under four different HITs.[54] Apart from the description of the compensation structure, the HITs were virtually identical.[55] MTurk workers who had already completed one of our HITs were ineligible to

52. *See infra* Part II.

53. *See supra* Section I.B.

54. We considered, but decided against, a survey design that randomized the form of compensation. Our research question examines how different compensation structures influence a respondent's willingness to undertake a task and how well she performs the task. To properly examine this question requires that the respondent know the compensation structure up front.

55. The only other material difference was the statement that if they had taken one of our prior surveys, they were not eligible. They were also told that they would be asked to enter their MTurk WorkerID to help determine eligibility. Any respondent that had completed a prior HIT was deemed ineligible for the final sample. By excluding repeat participation, we avoided bias (presumably upward, as workers improve on their initial performance) in subsequent variations of the test.

take another one.[56] In addition to stating in the HIT that participants must reside in the United States, we employed the MTurk worker requirement function and required that MTurk participants be geographically located in the United States.[57] All four of our HITs were posted between 10:00 a.m. and 11:00 a.m. eastern time (7:00 a.m. and 8:00 a.m. pacific time) on either a Tuesday or a Thursday in December 2018.[58] After each HIT was posted, we waited until we had 250 eligible respondents[59] before canceling the HIT.[60]

The four different treatment groups allowed us to vary both the base pay and the existence of a bonus for correct answers. Group 1—the "low" group—was offered $0.50 for completing the questionnaire. Group 2—the "low plus incentive" group—was offered $0.25 for completing the questionnaire, plus a bonus of $0.10 per correct answer. Because there

---

56. One potential risk in doing this is that, to the extent that this exclusion affects the composition of our sample, MTurk workers who are predisposed to participating in academic studies—so called "superturkers"—are disproportionately likely to be in Group 1. We address this concern through the following: in our main specifications, we control for demographic variables such as education and income in an effort to make apples-to-apples comparisons. Of course, the possibility remains that individuals who most enjoy participating in academic studies differ along dimensions that are unobservable to us. However, these are precisely the MTurk workers who may have had prior exposure to some of our questions or styles of question, and/or are disproportionately likely to perform well on our questionnaire. This selection effect should therefore bias our results toward finding higher performance in Group 1, and lower performance in later groups such as Group 4. In fact, as discussed in more detail in Section II.B, we find the opposite pattern. *See* discussion *infra* Section II.B. This suggests that, if anything, our results *understate* the differences between groups that would be observed in the absence of this selection effect.

57. We stated in the HIT that workers must be at least eighteen years old and reside in the United States to participate. In addition to this, we used the worker requirement function because of evidence that a substantial proportion of the workers in MTurk are located outside the United States. *See* Difallah et al., *supra* note 18, at 3. Given that we were offering substantially above market compensation in Groups 3 and 4, we were concerned that, in the absence of this worker requirement, MTurk workers outside the United States might be tempted to misstate their residency in order to participate, thereby polluting our results.

58. The specific dates were December 6, 11, 13 and 18, 2018. In order to reduce the risk of contamination between treatments, we required that one full business day pass between when the last HIT was cancelled and the next HIT was posted. One HIT (Group 2) completed before midnight the day it was posted, which allowed us to post the next HIT (Group 3) two days later. The other three HITs were completed by mid-afternoon the day after they were posted. In total, it took roughly twenty-five hours for Group 1 to be complete, thirteen hours for Group 2, twenty-seven hours for Group 3, and twenty-nine hours for Group 4.

59. The MTurk HIT contained a link to a Qualtrics questionnaire. Once a participant had completed the survey, she was given a unique response code by Qualtrics and asked to enter it into the HIT. We matched the response codes provided by the MTurk workers with the original response codes provided to them in the Qualtrics questionnaire. Only responses that matched were counted as eligible responses.

60. We had a few extra to ensure that we would end up with 250 eligible responses. In all four treatments, we ended up with a few more than 250 eligible responses. We kept the first 250 eligible responses (by date and time submitted) and discarded the rest.

were ten questions in the questionnaire, the total potential compensation for participants was $1.25. Group 3—the "high" group—was offered $1.50 for completing the questionnaire. Finally, Group 4—the "high plus incentive" group—was offered $1.25 for completing the questionnaire, plus a bonus of $0.10 per correct answer, for a total possible compensation of $2.25. The compensation structure is summarized in Table 1.

TABLE 1: COMPENSATION STRUCTURE BY TREATMENT GROUP

Group	Description	Base Compensation	Maximum Bonus
Group 1	Low, No Incentive	$0.50	—
Group 2	Low, Plus Incentive	$0.25	$1.00
Group 3	High, No Incentive	$1.50	—
Group 4	High, Plus Incentive	$1.25	$1.00

We advised potential participants that we expected the task to take approximately ten minutes to complete.[61] This duration corresponds to an implied hourly rate of $3.00 for Group 1 and $9.00 for Group 3. Anecdotally, the "market" hourly rate on MTurk is approximately $6.00 per hour (or $1.00 per ten-minute task). We selected the compensation structure of Groups 1 and 3 to correspond to "below market" and "above market" rates on MTurk, respectively.

After the consent page, our questionnaire contained a page collecting demographic characteristics, including gender, birth year, educational attainment, and income.[62] Respondents then proceeded to the third page of the questionnaire, which contained the ten substantive questions, the order of which was randomized. Four of the questions involved solving mathematical problems. One related to numeracy, one to algebra, one to geometry, and one to order of operations.[63] These questions are objective in nature and require mathematical computations of varying levels of difficulty. The other six involved basic logic. While these questions are also objective in nature, answering

61. As we discuss in more detail below, the average participant spent somewhat less than ten minutes completing the questionnaire. The average time spent was between 6.5 and 6.75 minutes for Groups 1 through 3, and about nine minutes for Group 4, using the winsorized variable. *See infra* Section II.B.

62. We also collected information on whether the individual was born in the U.S., state of residence, race, whether the individual is Hispanic, and employment status.

63. The first three of these were based on ninth-grade mathematics questions. The fourth was based on fifth-grade mathematics.

them correctly requires a careful reading of the question rather than mathematical computations. All but one of the questions were free-response questions, where respondents were required to type in their answers.[64] As we discuss in more detail in Section II.C, if anything, open-text fields are likely to mitigate any tendency towards inattention, since participants cannot quickly click through the way they could with a multiple-choice questionnaire.

We distinguish our use of the term *attention* from the way others use it with respect to survey validity. Researchers, to ensure the validity of a survey, may implement an *attention check* to identify careless or unengaged respondents and allow researchers to screen them from the analyzed data.[65] In general, they are interested in a binary determination—that is, whether the respondent is paying attention or not. In contrast, we are examining a more nuanced conception, where the type of answer provides insight into the respondent's continuous level of attention. Our conception of attention uses a higher bar than a typical attention-check question to reflect the context in which MTurk is being used. In the context of legal scholarship, researchers use MTurk to ask questions that are reasonably complex, and require a nontrivial amount of attention to understand. As such, the level of attention should be more than simply confirming that a human has glanced at the questions.

B. Results

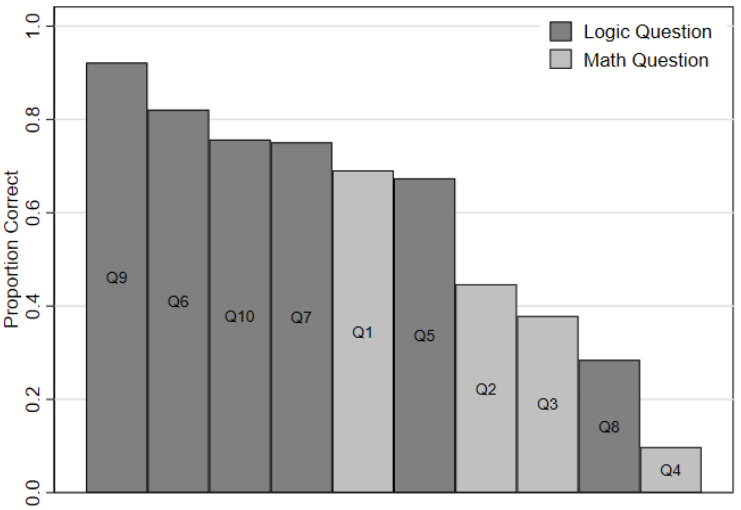
We begin by plotting the overall proportion of correct answers for each question and present the results in Figure 1. Figure 1 shows that there is substantial heterogeneity in accuracy across questions. Whereas about 90% of respondents answered question 9 correctly, only about 10% gave the correct answer to question 4. Overall, our respondents were more successful in answering the logic questions than the math questions, although this is not universally true—the second most difficult question, as measured by the number of respondents who answered it incorrectly, was a logic question.

---

64. The exception was the numeracy question, which asked respondents: "Which of the following numbers is the largest?" Respondents were required to select from among six options. The full question text is reproduced in Appendix A.

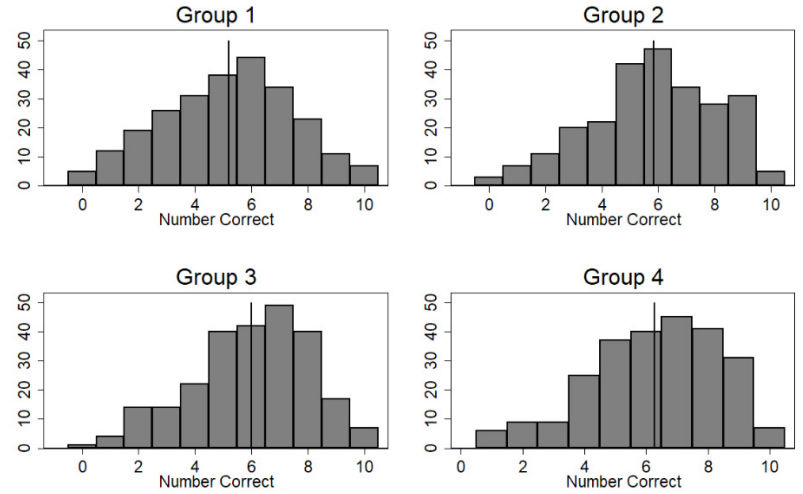
65. One such example of an attention check is the following question: "There are five choices below. Please mark the middle choice." *See* Franki Y.H. Kung et al., *Are Attention Check Questions a Threat to Scale Validity?*, 67 APPLIED PSYCHOL. 264, 265 (2018) (describing the purpose of attention checks).

FIGURE 1: PROPORTION OF RESPONDENTS WHO ANSWERED EACH QUESTION CORRECTLY



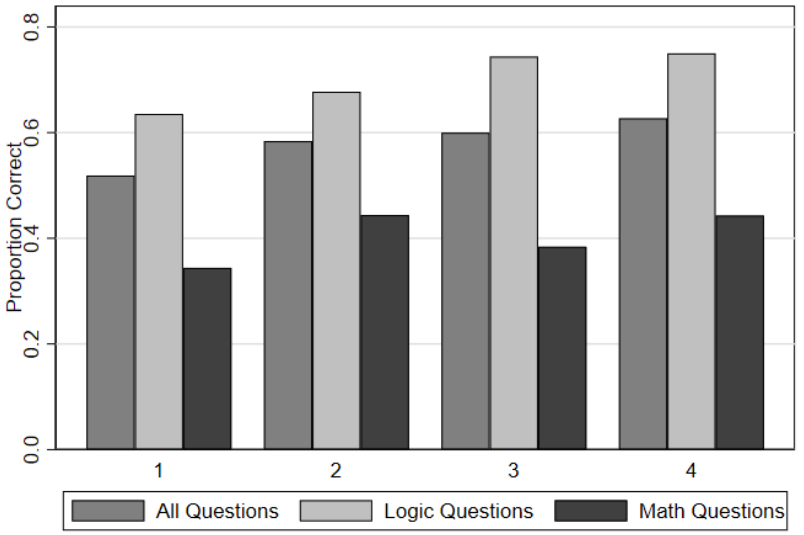
We then compare the overall performance across the four groups in Figure 2. Already, with this completely nonparametric approach, there appear to be clear differences between the four groups. The scores increase monotonically from Group 1 to Group 4. The histograms also reveal that the shape of the distributions vary across the four groups, suggesting that while different compensation systems improve performance, the effects may differ across questions. We then extend this analysis and present the average raw performance of each group, as well as their performance on the math and logic questions, in Figure 3.

FIGURE 2: DISTRIBUTION OF CORRECT ANSWERS, BY GROUP



The vertical line represents the average within each group.

FIGURE 3: AVERAGE PROPORTION OF CORRECT ANSWERS, BY GROUP AND TYPE OF QUESTION



Despite our experimental design, it is always possible that there could be other differences between the four groups. To account for these differences as much as possible, we also estimate the relative performance of each group using a multivariate regression, which allows us to control for the respondents' self-reported age, gender, educational attainment, and income. The estimated differences in performance are presented in Table 2. Details on the regression specification, as well as the full list of results, are reserved for the appendix. As it turns out, the addition of controls does not materially change the relative performance across groups.

TABLE 2: PERFORMANCE DIFFERENCES ACROSS TREATMENT GROUPS—ALL QUESTIONS

	Group 1	**Group 2**	**Group 3**
Group 2	0.575**		
p-value	(0.003)		
Group 3	0.753***	0.178	
p-value	(0.000)	(0.350)	
Group 4	1.047***	0.472*	0.293
p-value	(0.000)	(0.014)	(0.124)

*** p<0.001, ** p < 0.01, * p<0.05

Each cell presents the difference in coefficients between the group indicated in the row and the group indicated in the column. So, for example, the first cell shows that, on average, the number of correct questions for participants in Group 2 was 0.575 higher than it was for participants in Group 1. p-values of F-tests under the null hypothesis that the two coefficients are the same are presented below in parentheses.

As Table 2 makes clear, participants in Groups 2, 3, and 4 substantially outperform participants in Group 1. Group 4 has the highest overall performance, and its performance differential is impressive. Participants in Group 4 score, on average, 1.05 points higher than those in Group 1. Given that this is a test out of ten points, this improvement represents over 10% of the *total* possible score. Even more impressive, this differential represents about 20% of the average score of participants in Group 1.[66] Group 4 also substantially outperforms Group 2, although the magnitude of this outperformance is less than half as large. While the point estimate on the difference

66. The average raw score of participants in Group 1 was 5.188 points (out of a possible ten points).

between Group 4 and Group 3 is positive, the difference is not quite statistically significant.

Even though the addition of controls has little effect on the relative performance across groups, these multivariate regressions have a second benefit, which is that they allow us to estimate performance differentials across the demographic characteristics. We discuss these differences in more detail below in Section II.C.5.

Next, we divide the questions into math questions and logic questions and perform the same analysis. The results are presented in Table 3. Panel A presents the results relating to the four math questions, and Panel B presents the results relating to the six logic questions. The full set of coefficient estimates is presented in Appendix B, Table 4.

TABLE 3: PERFORMANCE DIFFERENCES ACROSS TREATMENT GROUPS—BY QUESTION TYPE

Panel A: Math Questions			
	Group 1	**Group 2**	**Group 3**
Group 2	0.393***		
p-value	(0.000)		
Group 3	0.165	-0.229*	
p-value	(0.115)	(0.027)	
Group 4	0.421***	0.028	0.256*
p-value	(0.000)	(0.789)	(0.013)
Panel B: Logic Questions			
	Group 1	**Group 2**	**Group 3**
Group 2	0.182		
p-value	(0.168)		
Group 3	0.589***	0.407**	
p-value	(0.000)	(0.002)	
Group 4	0.626***	0.444***	0.037
p-value	(0.000)	(0.001)	(0.777)

*** p<0.001, ** p < 0.01, * p<0.05

Each cell presents the difference in coefficients between the group indicated in the row and the group indicated in the column. Panel A presents the results for the four math questions, and Panel B presents the results for the six logic questions. p-values of F-tests under the null hypothesis that the two coefficients are the same are presented below in parentheses.

Table 3 demonstrates an interesting dimension of heterogeneity. Specifically, the incentivized groups (Groups 2 and 4) seem to perform substantially better on the math questions than the non-incentivized

groups (Groups 1 and 3). For example, Group 2 substantially outperforms Group 3 in this subsample, despite the fact that, as shown in Table 2, there is no difference in the overall performance of the two groups. Group 4 also substantially outperforms Group 3 in this treatment, despite the fact that the difference was not statistically significant overall. While the point estimates are lower than they were in Table 2, this decrease is partially related to the fact that there were only four math questions. In fact, the coefficient estimates are quite large relative to the means—for example, the 0.23 point difference between Group 3 and Group 2 represents a performance gap of about 15% relative to the average performance in Group 3.

Panel B demonstrates the opposite pattern. Here, the difference appears to be between the high base groups (Groups 3 and 4) and the low base groups (Groups 1 and 2). For example, both Groups 3 and 4 substantially outperform Group 2 (as well as Group 1). In contrast, the differential between Groups 1 and 2 is much smaller in this subsample and is not statistically significant.

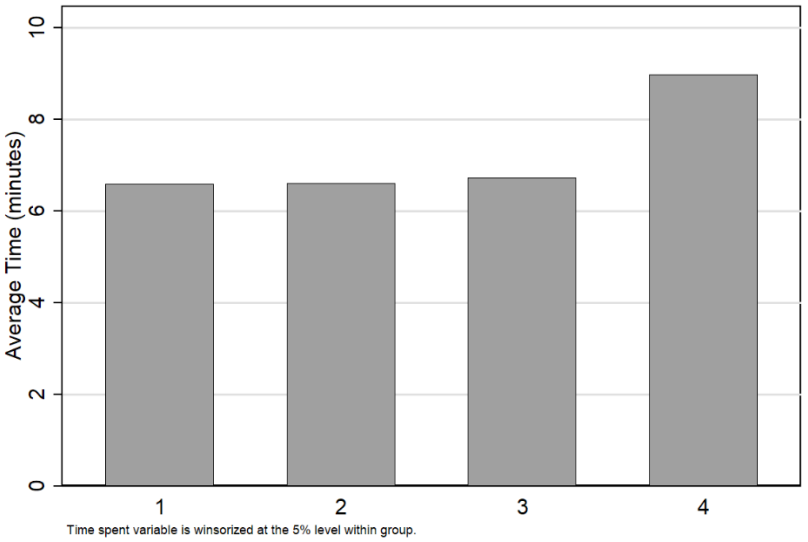
So far, the results in Table 2 and Table 3 suggest a straightforward story: the performance of MTurk workers is sensitive to compensation structure, but in a very specific way. Performance on questions that involve exerting effort—that is, the math questions—is sensitive to the bonus payments. In contrast, performance on questions that simply involve paying attention—basic logic questions—is sensitive to the base pay offered. Stated another way: bonuses motivate workers to exert more effort, while higher base pay increases their level of attention. We discuss this phenomenon in greater detail in Section II.C.

We also investigate the relationship between compensation, time spent, and performance. To account for the fact that some respondents may have opened the questionnaire and then returned to complete it later, we winsorize the time spent variable at the 5% level.[67] We begin with Table 4, which summarizes the average time spent by group. The most noticeable aspect of this figure is that the amount of time spent between Groups 1 through 3 is virtually identical; the differences among these groups are not statistically significant. Participants in Group 4, in contrast, did spend substantially more time

67. Winsorizing involves recoding the first 5% of the values to the fifth percentile, and the top 5% to the values of the ninety-fifth percentile. This form of censoring helps to mitigate the potentially distortionary effect of outliers. *See* Alan Reifman & Kristina Keyton, *Winsorize*, *in* ENCYCLOPEDIA OF RESEARCH DESIGN 1637 (Neil J. Salkind ed., 2012).

completing the questionnaire. This difference is highly statistically significant.[68]

FIGURE 4: TIME SPENT ANSWERING THE QUESTIONNAIRE, BY GROUP



The results so far suggest that the performance differential across groups will not be fully captured by time spent completing the questionnaire. To confirm this, we include the time spent variable in the regression. This inclusion allows us to disentangle the direct relationship between the compensation structure and performance from any indirect effect that flows through time the worker spent completing the assignment. While the coefficient on time spent was positive and statistically significant, the magnitude was quite small—approximately 10% of the size of the performance differential between Groups 1 and 2, and an even smaller percentage of the difference between Group 1 and Groups 3 or 4. Perhaps even more importantly, the inclusion of the time spent variable causes the coefficients on the group variables to change only very slightly, indicating that the performance differential observed between groups is, at most, only

68. Specifically, using this measure in pairwise comparisons, the difference between the average time spent between Group 4 and each of Group 1, Group 2, and Group 3 is statistically significant at the 0.001 level.

weakly related to any differences in time spent across groups. The full set of coefficient estimates is presented in Appendix B, Table 5.

Finally, we investigate whether the characteristics of the respondents in the groups differed in any systematic ways. To that end, we compare the observable demographic characteristics across the four groups. In general, the demographic point estimates are very similar. The biggest difference is in educational attainment,[69] where we find that Group 1 had significantly more highly educated respondents (respondents with more than a bachelor's degree) than the other three groups. As mentioned in Section III.A, this may be related to the fact that MTurk workers who had participated in one of our HITs were precluded from participating in subsequent HITs. To the extent that very highly educated individuals might be more inclined to participate in an academic study, this may not be surprising.

One potential concern that this difference in educational attainment raises is that there could also be selection on other unobservable factors. While not ideal, to the extent that any selection exists, we would expect those individuals to perform disproportionately well on the questionnaire due to some combination of experience and aptitude. In other words, it is likely to bias our results towards finding *higher* performance in Group 1, and *lower* performance in later groups (such as Group 4). Given the likely direction of this bias, if anything, our results are likely to *understate* the differences between groups that would be observed in the absence of this selection effect.

C. Interpreting Our Results

In this Section, we interpret our results and discuss four implications of our findings. First, we discuss the marginal effects of incentives versus higher base pay. Next, we discuss the implications of using time spent as a proxy for effort. Third, we discuss the costs associated with each of the compensation structures, as well as the average cost per correct question. We then discuss the differences in these measures between the math and the logic questions. Finally, we

---

69. There were only four other pairwise comparisons that were statistically distinguishable at the 5% level. Given that there are a total of seventy-two pairwise comparisons, we would expect that, on average, 3.6 comparisons would be statistically significant at the 5% level by random chance. Observing four comparisons that are statistically distinguishable at the 5% level is therefore entirely consistent with random chance, and does not represent a concern. The full set of estimates is in Appendix B, Table 7.

discuss the relationship between performance and demographic variables in our experiment.

1. Incentives v. Base Pay

As mentioned in Section II.B, the results in Table 2 and Table 3 suggest that the compensation structure has differential effects between different types of questions. In particular, we found that participants' performance on questions that involve exerting effort—that is, on questions involving solving mathematical problems—is highly sensitive to the payment of bonuses. In contrast, we found that participants' performance on questions that involve paying close attention to the question—that is, on questions involving logic—is highly sensitive to the level of base pay offered. This observation is particularly noteworthy given that, in contrast to a large number of questionnaires fielded on MTurk, our design forced respondents to enter typed text rather than simply clicking on an answer from a multiple-choice list. To the extent that this requirement affected our results, it is likely to mitigate any general tendency towards inattention.[70] That we still find differences between groups, particularly with respect to the logic questions, is therefore even more noteworthy.

We believe that the most natural interpretation of this result is that MTurk workers respond differently to the payment of incentives than they do to higher base pays. Whereas incentives seem to induce workers to exert effort, it does not seem to induce them to read questions more carefully. One potential explanation for this differential effect is that the MTurk workers interpret a high base pay as a signal that the requester (that is, the experimenter) places a high value on the task. This then leads to two non-exclusive potential inferences. The first is that respondents believe that, because the requester places a high value on the task, she may have higher standards for the work performed and is therefore more likely to reject the work performed. Respondents dislike this outcome, as it results in nonpayment. Given that hirers on MTurk can restrict participation by worker reputation,[71]

---

70. Indeed, while participants could enter any text they wanted, the respondents do not appear to have entered text randomly. For example, on several occasions, respondents typed out answers along the lines of "I don't know" or "I don't remember the formula."

71. *See* Eyal Peer, *Reputation as a Sufficient Condition for Data Quality on Amazon Mechanical Turk*, 46 BEHAV. RES. METHODS 1023 (2014) (observing that worker reputation on MTurk closely correlates with worker performance).

it follows that MTurk workers care about reputation, even beyond the fact that they are not compensated for rejected work.

A second non-exclusive interpretation is that participants believe that the highly compensated task is a precursor to a subsequent task (or several additional tasks). For example, MTurk allows requesters to assign workers to "Qualification Types" and includes a score from zero to one hundred for each Qualification Type.[72] The requester can then use these Qualification Types later for subsequent tasks.[73] If MTurk workers believe that above-market wages could suggest that the requester is really seeking to qualify workers for later tasks, they might be inclined to answer more carefully in anticipation of potentially lucrative future opportunities.[74] Under both of these explanations, MTurk workers are behaving strategically and trying to infer the underlying motivations or desires of the requester based on the above-market compensation. This is consistent with the observed non-naivety of many MTurk workers.[75]

In contrast, bonuses do not seem to induce workers to perform better at the basic logic questions (that is, those that rely on carefully reading the question), but do seem to induce them to do better at the math questions (that is, those that rely on effort). This difference suggests that bonuses, while improving performance, may not necessarily induce increased *attention* among participants. Accordingly, the efficacy of bonuses is likely higher for sweat-of-the-brow type questions where effort matters.

2. Time Spent as a Proxy for Effort

A second major implication of our findings is that the amount of time that an MTurk participant spends on a task may be a poor proxy for effort. We can see this clearly in two different ways. First, we see that between Groups 1 and 2, and Groups 1 and 3, there is a substantial improvement in average performance, with no accompanying increase in average time spent. Moreover, despite the performance differential across question types between Groups 2 and 3, there is no difference in time spent. Both of these findings suggest that any differences in

---

72. *See Amazon Mechanical Turk: Requester User Interface Guide*, AMAZON WEB SERVICES 52–-54 (2019), https://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/amt-ui.pdf# ManagingQualificationTypes [https://perma.cc/HKZ3-C9Q4].

73. *See id.*

74. While we think that this is plausible, it is, of course, somewhat speculative.

75. *See supra* note 47 and accompanying text.

performance cannot be attributed to the amount of time the MTurk participants spent answering the questions.

Second, as discussed in Section II.B, we found that including time spent as a control variable hardly changed the relationship between group and performance, suggesting that the differences between groups in terms of performance are largely unrelated to any differences in time spent. Taken together, these results suggest that using time spent as a proxy for effort may be problematic.

The issue of time as a function of effort is a familiar concept in the legal profession. Hourly billing is a long-established practice by which lawyers charge for their services.[76] Prior to hourly billing, lawyers were compensated primarily through standardized fees. As legal work became more complex, and the time incurred more variable, the legal profession borrowed from the accounting world to introduce hourly billing.[77] In the aftermath of the 2008 recession, and with the emergence of legal technology, law firms have explored other alternatives. For example, clients at large law firms began to resist the traditional hourly billing model, resulting in firms offering flat fees, contingency fees, and "collars" (an agreed-upon range of fees where the counsel receives a bonus if the actual fees fall below the range and the client receives a discount if the actual fees fall above the range).[78]

3. Costs Associated with Each Group

A third implication of our results is that, as the saying goes, "you get what you pay for."[79] In particular, as we move from Group 1 to Group 2 or 3, and from Group 2 (and, to a lesser extent, Group 3) to Group 4, performance increases. At the same time, of course, the average cost also increases. Indeed, the cost ends up rising faster than the performance, meaning that the cost per correct answer also

---

76. For an interesting early discussion of hourly billing, see SPECIAL COMM. OF ECON. OF LAW PRACTICE FOR THE AM. BAR ASS'N, THE 1958 LAWYER AND HIS 1938 DOLLAR (1958).

77. *See* Herbert M. Kritzer, *Lawyers' Fees and the Holy Grail: Where Should Clients Search for Value?*, 77 JUDICATURE 187, 187 (1994).

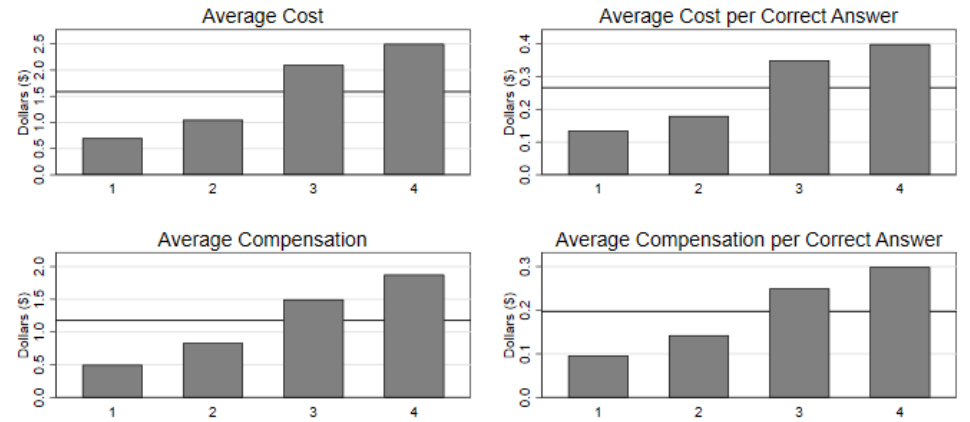
78. *See* Jonathan D. Glater, *Billable Hours Giving Ground at Law Firms*, N.Y. TIMES (Jan. 29, 2009), https://www.nytimes.com/2009/01/30/business/30hours.html [https://perma.cc/T82B-X7PD] (describing client opposition to hourly rates); Catherine Ho, *Law Firms Look for Alternatives to the Billable Hour*, WASH. POST (Apr. 15, 2012) https://www.washingtonpost.com/business/economy/law-firms-look-for-alternatives-to-the-billable-hour/2012/04/15/gIQAeyW9JT_story.html [https://perma.cc/U285-B7HX] (describing alternative fee arrangements).

79. *See* KURT VONNEGUT, CAT'S CRADLE 128 (Dial Press Trade Paperbacks 2010) (1963).

increases as we move across Groups. Figure 5 summarizes these results.

As Figure 5 makes clear, the largest jump is between Group 2 and Group 3, where we find that both the average cost and the average cost per correct answer almost double (from $1.05 to $2.10, and from $0.18 to $0.35, respectively). Part of this is due to the fee structure employed by MTurk. While HITs with ten or more assignments (which includes HITs containing surveys with ten or more respondents) are charged a 40% fee, bonuses are only charged a 20% fee.[80] While this accounts for some of the gap, that portion is relatively small: looking instead at average total compensation and average compensation per correct answer, the figures for Group 2 and Group 3 are $0.83 and $1.50, and $0.14 and $0.25, respectively. Overall, the average compensation and the average compensation per correct answer in Group 2 is roughly 55% of that in Group 3. Both cost and compensation per correct answer rise even further when we move to Group 4, where we find that the cost per correct answer is about $0.40 (and the compensation per correct answer is about $0.30).

FIGURE 5: COSTS AND COMPENSATION FOR ALL QUESTIONS, BY GROUP



The horizontal line represents the average across all four groups.

The more subtle implication of these results—going beyond the "you get what you pay for" maxim—is that the marginal cost of

80. *Pricing*, AMAZON MECHANICAL TURK, https://www.mturk.com/pricing (last visited Sept. 10, 2019) [https://perma.cc/JM9A-CXSX].

increased performance is quite high. Our inference is that there may be an upper bound to the performance that can be reasonably expected from MTurk participants. One might be able to dramatically improve performance by paying MTurk workers substantially more than typical rates. But doing so, depending on the assignment, may distort findings, either by drawing on an anomalistic population or encouraging effort that is incongruous with the real world.

4. Heterogeneity in Costs Across Question Types

The differences between Groups 2 and 3, discussed in Section II.C.3, suggest that bonuses may be a more cost-effective way to increase performance overall than a higher base pay. In fact, we can push this analysis further by distinguishing between math and logic questions. To do so, we repeat the analysis in Figure 5 separately for these two groups of questions.[81] The results are presented in Figure 6 and Figure 7 for the math questions and the logic questions, respectively.

81. In order to do this, we must attribute the costs and compensation between the two groups of questions. For Groups 1 and 3 (where there is no incentive component), we simply divide the total cost and compensation between the questions on a pro rata basis. Because there were four math questions and six logic questions, this resulted in attributing 40% of the cost and the compensation to the math questions, and 60% to the logic questions. For Groups 2 and 4 (where there was an incentive component in addition to the base component), we attributed the base compensation in the same way. We then added the incentive component based on the number of math questions that the individual answered correct to compute the cost and compensation for the math questions. The cost and compensation for the logic questions was computed analogously.

FIGURE 6: COSTS AND COMPENSATION FOR MATH QUESTIONS, BY GROUP

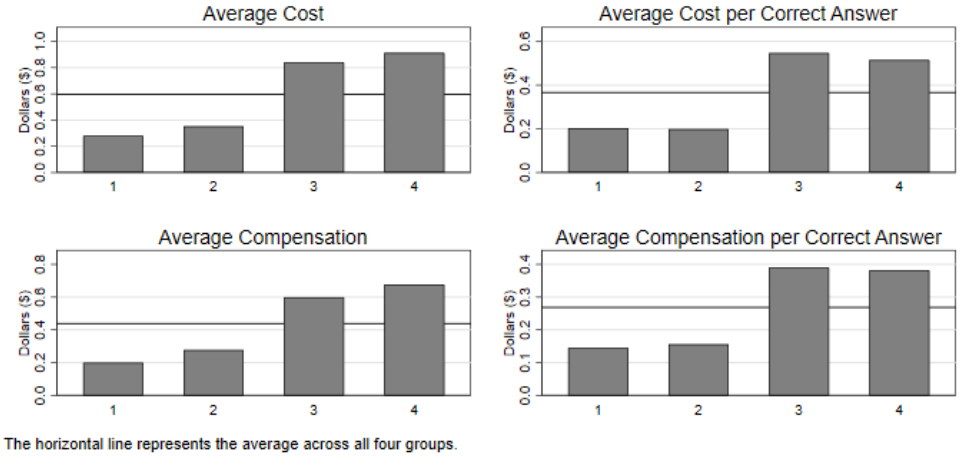
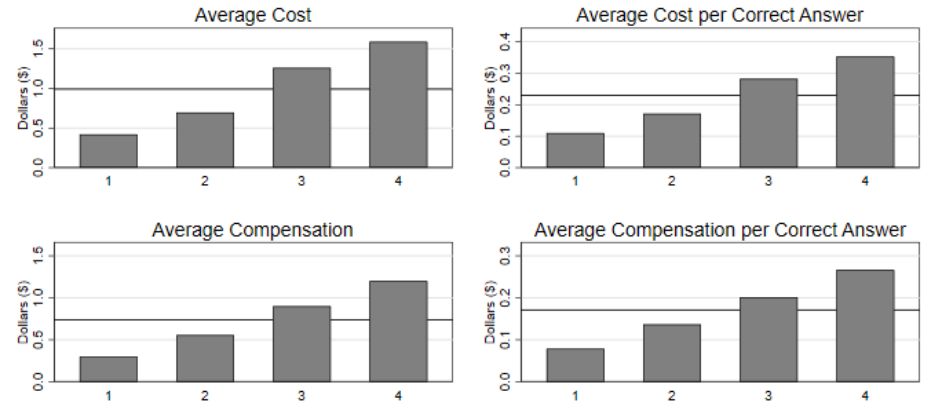


Figure 6 shows that, when we restrict attention to the math questions, the average cost (and compensation) increase in moving from Group 1 to Group 2 is offset by the increase in correct answers. As a result, the average costs (and compensation) per correct answer are virtually identical. In contrast, the average costs (and compensation) per correct answer jump dramatically in moving from Group 2 to Group 3, actually overshooting the values in Group 4. Figure 7, which is limited to the logic questions, tells a different story: the pattern displayed in Figure 7 is generally similar to that in Figure 5, albeit somewhat more muted. We discuss the implications of this further in Sections III.C and III.D.

FIGURE 7: COSTS AND COMPENSATION FOR LOGIC QUESTIONS, BY GROUP



The horizontal line represents the average across all four groups.

5. Relationship Between Demographics and Performance

A final implication of our results is that demographic variables of MTurk participants may represent a rather blunt measure of participant ability. To see this, we note that many of the demographic variables that one might think of as relevant predictors of performance on math and logic questions (including educational attainment, employment status, and race) are not statistically significant. This result holds even after controlling for the participant's compensation group.

We did find other demographic characteristics that correlated with performance. Some variables, including age, were statistically significant, although not necessarily in the direction that one might predict a priori. Others, such as whether the respondent identified as Hispanic, might have been correlated with English language skills, which is something that we did not directly ask about in the survey. Given the rich literature on the returns to education,[82] for example, the

82. *See, e.g.*, GARY S. BECKER, HUMAN CAPITAL: A THEORETICAL AND EMPIRICAL ANALYSIS, WITH SPECIAL REFERENCE TO EDUCATION (3d ed. 1993); David Card, *Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems*, 69 ECONOMETRICA 1127 (2001); James J. Heckman et al., *Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond*, *in* 1 HANDBOOK OF THE ECONOMICS OF EDUCATION (Eric A. Hanushek & Finis Welch eds., 2006); Philip Oreopoulos & Uros Petronijevic, *Making College Worth It: A Review of Research on the Returns to Higher Education* (Nat'l Bureau Econ. Research, Working Paper No. 19053, 2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2269476 [https://perma.cc/E37H-X5WF].

overall non-significance of demographic factors suggests that the type of workers who choose to work within MTurk may not be representative of the broader population. The full set of coefficient estimates is presented in Appendix B, Table 6.

Our findings suggest that the efficacy of an MTurk survey depends on the population group that the survey is designed to capture. Our results provide evidence that MTurk participants reflect a broad distribution of the population. For each compensation structure in our experiment, some respondents answered every question correctly. At the same time—with the exception of the high plus incentive group—some respondents answered every question *incorrectly*.

Our results also reveal two other phenomena: One, MTurk is less effective at drawing from a narrowly defined subset of the population. Each compensation structure generated a wide range of responses. Two, our results suggest that there is an upper bound in the performance of respondent pools recruited using MTurk. Depending on the subset of the population one is trying to reach, MTurk respondents may not be a valid comparison.

The upshot is that our findings have direct relevance to the use of surveys in legal research. In some contexts, MTurk may provide a germane comparison. For example, an MTurk survey may go a long way in informing our understanding of jury deliberations, as both MTurk and juries draw widely from the general population.[83] Conversely, asking MTurk participants to evaluate promulgations by the Environmental Protection Agency may yield less insight into how administrative agencies approach rulemaking (it may, however, still say something about how a layperson thinks about such matters).

Ultimately, our findings suggest that MTurk is neither a panacea nor a scourge for survey research. It is simply a mechanism to capture the views of a subset of the population, large or small. It works best when the MTurk participants accord with the desired population subset. When MTurk participants have little in common with the desired population subset, however, it effectively compares apples with oranges. Ultimately, legal scholarship using MTurk turns on the

83. Scholars note that jurors reach different results when deliberating individually than collectively. *See* David Austen-Smith & Jeffrey S. Banks, *Information Aggregation, Rationality, and the Condorcet Jury Theorem*, 90 AM. POL. SCI. REV. 34, 34–35 (1996).

assumption that it provides a credible proxy for the desired population subset. Stated in the negative: garbage in, garbage out.[84]

III. BEST PRACTICES

We now discuss the implications of our results for other studies. Overall, while there are undoubtably benefits to using MTurk, our results suggest that it should be used with caution. In this Part, we draw on the results of our experiment and suggest several best practices that researchers should adopt in using MTurk for legal and social science scholarship. In particular, we discuss five specific and concrete suggestions that come out of our study. These suggestions are intended to strengthen both the quality of the underlying studies and the interpretability of any resulting papers based on such studies.

A. Provide Minimum Disclosure of Research Methodology

A thorough review of published articles in the legal literature that relied on MTurk samples revealed that most provide only a high-level discussion of the research methodology. While this may be for the sake of parsimony, these details are frequently important for interpreting and evaluating the results in the paper. For example, most studies we reviewed do not report the compensation provided to participants. Given our experimental results—that responses are highly sensitive to compensation—participant compensation (both in terms of level and bonus structure, if any) should be viewed as a material fact and therefore should be clearly disclosed and discussed in the description of the research methodology.

This minimum disclosure should extend beyond the compensation structure. Many of the studies we reviewed omitted any discussion of a large number of relevant features, including the number of participants (both the number initially sought and the number included in the final analysis), the amount of time participants spent completing the questionnaire (both the amount of time participants were told that the questionnaire should take to complete and the average amount of time actually spent), the way the task was described to potential participants on the MTurk platform, when the HIT was

84. The origins of the terms are credited to IBM programmer George Feuchsel. *See What is Garbage in Garbage Out?*, WISEGEEK, http://www.wisegeek.org/what-is-garbage-in-garbage-out.htm (last visited Sept. 10, 2019) [https://perma.cc/TF5R-6AH4].

posted on MTurk, when the last eligible participant completed the HIT (and, by implication, the total amount of time between when the HIT was posted and when the last eligible participant completed it), and any qualifications required of MTurk workers (such as, for example, their location).[85] Because all of these features of the study are known to the researchers at the time that the study is being fielded, reporting this information is virtually costless. To the extent that researchers are concerned about clogging their papers with methodological details, this information can always be relegated to footnotes or to a methodological appendix. This approach would allow authors to maintain a degree of parsimony in the main text, while also preserving access to important information about the research methodology.

B. Ensure Robustness by Varying Compensation

In our review of the literature, we found only two studies that fielded their questionnaires with different compensation levels or structures, out of the ninety-eight that we reviewed.[86] Given our finding that respondents are sensitive to compensation, using only a single compensation level or structure leaves open the possibility that the results reflect the compensation structure more than they do the question of interest.

A second suggestion that comes out of our experiment is that researchers should field their questionnaire more than once, varying the compensation provided to participants each time. This would act as a robustness check—to the extent that the results are similar, this would help to ensure that the results are not artifacts of the

---

85. Note that we do not mean simply stating the requirement, since potential respondents may misstate their eligibility in order to participate. Rather, we mean Amazon requirements, which preclude ineligible workers based on information that Amazon has about the individual. Researchers should clearly state any eligibility requirements, as well as which requirements were screened for thorough self-reporting and which were imposed by qualifications.

86. One of these studies raised the compensation offered from $0.50 to $0.75 "to expedite data collection," Christopher T. Robertson et al., *Perceptions of Efficacy, Morality, and Politics of Potential Cadaveric Organ-Transplantation Reforms*, 77 L. & CONTEMP. PROBS. 101, 129 (2014), presumably because the authors felt that there was insufficient interest in participating in their survey at the lower compensation. The authors removed respondents who completed the survey in under 6.99 minutes, implying that the *maximum* hourly rate offered to respondents was about $4.29 under the lower compensation, and $6.43 under the higher compensation. The mean hourly rate was presumably substantially less than this. The second offered two different levels of compensation—$0.25 and $0.50. Garrett & Mitchell, *supra* note 35, at 492. While the authors did not discuss their rationale for offering two different compensation levels, they reported that "the same patterns hold when we . . . compare the participants receiving 25 cents vs. 50 cents." *Id.*

compensation structure. This, in turn, would lend support to the idea that there may be some external validity to the results. To the extent that differently compensated participants produce different results, understanding this would give researchers an opportunity to explore the phenomenon of interest more deeply.

C. Recognize that Time Spent May Be a Poor Proxy for Effort

Until recently, the legal profession had embraced the hourly billing model, which treated lawyers' time as a proxy for effort. Economic and technological forces have compelled firms—large and small—to revisit this maxim. In addition to the alternative fee arrangements discussed above, law firms are increasingly willing to discount the cost of their services in an effort to attract and retain clients, further weakening the hourly billing model.[87] Our experiment reaches a similarly nuanced conclusion that time may be a poor proxy for effort.

This is not to say that time conveys no information. Very low time spent on a survey may be a credible signal that the MTurk worker did not exert a good-faith effort on an assignment. We came across MTurk studies that excluded responses where the worker spent less than a threshold amount of time.[88]

At the same time, we found that workers who spent more time on the questions did not perform meaningfully better. Workers who spent more time did score higher, but this difference was very small and not statistically significant. Differences in compensation (that is, membership in Group 2, Group 3, or Group 4 versus Group 1) had an effect that was at least ten times greater than an additional minute of time. Our takeaway here is that researchers should report the time workers spend on assignments, but be wary of using it as a proxy.

D. Distinguish Between Subjective and Objective Questions, and Tailor Compensation Accordingly

A fourth recommendation is that researchers should distinguish between subjective questions (questions where researchers are asking

---

87. *See, e.g.*, Roy Strom, *How the Am Law 100 Left $4.4B on the Table in 2018*, AM. LAW. (Apr. 23, 2019, 9:50 AM), https://www.law.com/americanlawyer/2019/04/23/clients-are-demanding-discounts-and-leaving-law-firms-at-a-loss/ [https://perma.cc/X86P-8NV4].

88. *See, e.g.*, Cicchini & White, *supra* note 32, at 28 (rejecting responses from those who spent fewer than three minutes on the survey).

participants for their opinions) and objective questions (where there is a correct answer that the researcher can determine ex ante). It bears repeating that our study looks only at objective questions. At the same time, we ask questions that correlate with effort (math) and attention (logic). It is our study's observations on the latter characteristic—attention—that we believe can shed some light on subjective questions.

Our results clearly show that incentives, coupled with a low base, provide a cost-effective way to improve participant performance. In a setting where the researchers are interested in posing questions with objective answers, and/or where the goal is to induce participants to exert effort in solving problems, this type of compensation structure may be preferable.

In the majority of the studies we reviewed, the researchers were interested in participants' answers to subjective questions. Sometimes this manifests itself as a simple survey asking participants about their beliefs or preferences.[89] Others are designed as experiments—for example, a study related to jury instructions might investigate whether a change in jury instructions makes the respondent more likely to rule in favor of the defendant,[90] or a study related to taxpayer audits might investigate whether the way audits are conducted (randomly or based on past behavior) affects cheating.[91]

In these types of studies, the researcher is interested in eliciting the respondents' honest and thoughtful answers and reactions. As a result, the researcher may be more interested in inducing the respondent to pay attention—and to answer the question attentively—than she is in inducing the respondents to exert effort to solve a problem. The subjective nature of the question means that there is no single "correct" answer. That said, one would still like respondents to pay attention to the question, and think carefully before providing a response.

In these settings, our results show that a higher base compensation—substantially above the common rates of $0.25 to $1.00 per ten minutes (for an implied hourly rate of $1.50 to $6.00)—may be required to induce participants to answer attentively.[92] While such

---

89. *See, e.g.*, Bambauer, *supra* note 32, at 1204–06; Cass R. Sunstein, *People Prefer System 2 Nudges (Kind Of)*, 66 DUKE L.J. 121, 155 (2016).

90. *See, e.g.*, Cicchini & White, *supra* note 32, at 23.

91. *See e.g.*, Satterthwaite, *supra* note 39, at 8–9.

92. Indeed, even with the higher compensation, a substantial fraction of participants *still* failed to answer our logic questions correctly, indicating that they were still not answering attentively.

higher compensation rates do cut against the cost effectiveness of MTurk as a research platform, our evidence shows that they also substantially improve performance, making the results of studies that rely on it far more reliable. We believe that this is a worthwhile tradeoff.

E. Recognize a Potential Upper Bound in the Quality of MTurk Participants

A fifth recommendation is that researchers should recognize a potential upper bound to the quality of responses achievable by MTurk participants. Whether the pool of MTurk participants is representative of the broader U.S. population remains an open question: some studies have found it to be fairly representative,[93] while others have found it to be less so.[94] Our study does not address the question of whether the pool of MTurk participants is broadly representative, as this was never part of our research design. However, we do find evidence that the performance of participants drawn from MTurk may be bounded at a relatively modest level. For example, even under our most generous compensation structure, the average number of correct questions was less than 6.3 out of 10.

Average performance on the math questions was even more disappointing: even in the highest compensated group, the average score was about 1.772 out of 4 (or about 44%).[95] Restricting attention to respondents who reported having at least an associate degree does nothing to improve average performance: in fact, the average performance actually falls by the very tiny amount of 0.004 (to 1.768 out of 4).[96] This is despite the fact that the math questions are drawn from the curriculum for ninth grade and below.

---

93. *See, e.g.*, Berinsky et al., *supra* note 12, at 352 (finding that MTurk respondents are more representative of the U.S. population than in-person convenience samples, but less representative than subjects in web-based panels or national probability samples such as the Current Population Survey).

94. *See, e.g.*, Gabriele Paolacci & Jesse Chandler, *Inside the Turk: Understanding Mechanical Turk as a Participant Pool*, 23 ASS'N FOR PSYCHOL. SCI. 184, 184–85 (2014) (finding that workers, while diverse, are not representative of the populations from which they are drawn).

95. The performance differential on the math questions between respondents with an associate degree or higher and respondents with less than an associate degree is not statistically significant at the 5% (or even the 10%) level.

96. Restricting attention to participants with at least a bachelor's degree also makes very little difference—the number of correct answers out of four rises to 1.858 out of four (or 46%). The performance differential on the math questions between respondents with a bachelor's degree or higher and respondents with less than a bachelor's degree is not statistically significant at the 5% (or even the 10%) level.

The fact that the average quality of the responses is somewhat disappointing is not necessarily fatal to all applications of MTurk for legal research. Rather, whether the upper bound is problematic will depend on the specifics of the question at issue. For example, this may not be a serious problem in the context of a study designed to simulate the behavior of jurors if the researcher believes that the typical juror is drawn from a similar pool.[97] On the other hand, this may be a bigger problem in the case of a study seeking to simulate the behavior of a more sophisticated pool of individuals than those available through MTurk.

F. Create Objective Measures to Supplement Self-Reported Demographic Information

A final, related recommendation from our study is that researchers should exercise caution in using demographic variables to control for factors such as participant ability, skill, or "quality." As discussed in Section II.C.5, our study shows that standard demographic variables (such as education and age) do not perform as one would typically expect, suggesting that they may make for poor control variables. This, in turn, may create interpretative challenges and, at the extreme, cast doubt on certain findings.

As such, even if the pool of MTurk respondents may be appropriate for the question of interest,[98] it is worthwhile to construct an objective baseline measure of the quality of each respondent. For example, a researcher could include objective questions within the survey instrument, which would enable her to construct a standard, objective measure for each respondent. These measures could then supplement other control variables, such as demographic variables, in the analysis. Demographic measures are often poor proxies for the underlying measure of interest (for example, ability), exacerbated by unobservable selection for MTurk participation. For these reasons, simply relying on the standard demographic variables may be problematic. This reinforces our suggestion above that a more

---

97. *See, e.g.*, Svein Magnussen et al., *Beliefs about Factors Affecting the Reliability of Eyewitness Testimony: A Comparison of Judges, Jurors and the General Public*, 24 APPLIED COGNITIVE PSYCHOL. 122, 130 (2010) (describing how the jurors performed comparably to the general public with respect to the reliability and credibility of eyewitness testimony).

98. *See* discussion *supra* Section III.E.

appropriate measure of "quality" is one that is constructed from the inclusion of objective questions in the survey instrument itself.[99]

CONCLUSION

MTurk plays an increasingly prominent role in legal scholarship, providing scholars with access to individual responses quickly and cheaply. While it offers several attractive features, we identify two major concerns regarding the current use of MTurk in legal scholarship. We design an experiment and demonstrate that the performance of MTurk participants is highly sensitive to the compensation structure offered and that this sensitivity depends on the nature of the tasks involved. We draw on our experience and results to propose a series of best practices for legal scholars and law reviews to adopt in studies involving MTurk participants going forward. Our hope is that adoption of the guidelines will help to mitigate potential criticisms of studies relying on MTurk, thereby fostering richer and more fruitful scholarly discussions.

99. *See* discussion *supra* Section III.E.

APPENDIX A: QUESTION TEXT

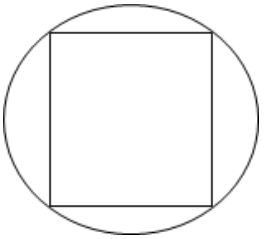
Question 1 [math]: Which of the following numbers is the largest?

- 5/7
- -0.6
- -1/6
- 0.004
- 6/8
- -2/7

Question 2 [math]: If $4m - 2 = 5m+7$, m = ?

Question 3 [math]: What is $2 \times (8 - 3)^2 + \sqrt{4} \div 2$?

Question 4 [math]: The figure below depicts a square and a circle. The square has all its corners on the circle. The area of the square is equal to 400 square inches. What is the area of the circle?



Question 5 [logic]: If you're running a race and you pass the person in second place, what place are you in?

Question 6 [logic]: A farmer had 15 sheep and all but 8 died. How many are left?

Question 7 [logic]: Emily's father has three daughters. The first two are named April and May. What is the third daughter's name?

Question 8 [logic]: How many cubic feet of dirt are there in a hole that is 3 feet deep x 3 feet wide x 3 feet long?

Question 9 [logic]: A man was born in 1955, how old is he on his 18th birthday?

Question 10 [logic]: How many three cent stamps are in a dozen?

APPENDIX B: ADDITIONAL EMPIRICAL RESULTS

TABLE 4: FULL REGRESSION ESTIMATES

	(1) All Questions	(2) Math Questions	(3) Logic Questions
Group 2	0.575**	0.393***	0.182
	(0.192)	(0.104)	(0.132)
Group 3	0.753***	0.165	0.589***
	(0.192)	(0.104)	(0.132)
Group 4	1.047***	0.421***	0.626***
	(0.191)	(0.104)	(0.131)
Female	-0.350*	-0.278***	-0.072
	(0.137)	(0.074)	(0.094)
Associate or Bachelor's Degree	0.014	0.131	-0.118
	(0.163)	(0.089)	(0.112)
Graduate Degree	-0.125	0.154	-0.279+
	(0.228)	(0.124)	(0.157)
High Income	0.113	0.151	-0.038
	(0.187)	(0.102)	(0.129)
Age: [30–33]	0.316	0.008	0.308*
	(0.194)	(0.106)	(0.133)
Age: [34–41]	0.586**	-0.007	0.593***
	(0.182)	(0.099)	(0.125)
Age: [42–100]	0.773***	0.166	0.607***
	(0.192)	(0.104)	(0.132)
Working Now	0.072	-0.062	0.134
	(0.186)	(0.101)	(0.127)
Born in the US	-0.409	-0.230	-0.179
	(0.367)	(0.199)	(0.252)
Black	-0.641**	-0.058	-0.583***
	(0.240)	(0.130)	(0.165)
Asian	-0.222	0.143	-0.366+
	(0.309)	(0.168)	(0.212)
Other Race	-0.323	-0.108	-0.215
	(0.371)	(0.201)	(0.255)
Hispanic	-1.584***	-0.588***	-0.997***
	(0.237)	(0.128)	(0.162)
Constant	5.548***	1.662***	3.886***
	(0.462)	(0.251)	(0.317)
Number of Observations	986	986	986
R^2	0.122	0.068	0.135
Adjusted R^2	0.108	0.053	0.121

Standard errors in parentheses. *** p<0.001, ** p < 0.01, * p<0.05, + p>0.1

This table presents coefficient estimates of OLS regressions. The dependent variable in Column (1) is the total number of questions answered correctly (out of a possible ten). The dependent variable in Column (2) is the total number of math questions answered correctly (out of a possible four). The dependent variable in Column (1) is the total number of logic questions answered correctly (out of a possible six). High income is a dummy equal to one if the respondent reported a total household income before taxes during the past twelve months of at least $100,000. White is the omitted category for race. Other Race includes respondents who selected multiple categories.

TABLE 5: FULL REGRESSION ESTIMATES, WITH AND WITHOUT TIME SPENT VARIABLE

	(1) All Questions	(2) All Questions	(3) All Questions
Group 2	0.575**	0.574**	0.589**
	(0.192)	(0.190)	(0.191)
Group 3	0.753***	0.741***	0.752***
	(0.192)	(0.190)	(0.191)
Group 4	1.047***	0.925***	1.015***
	(0.191)	(0.191)	(0.191)
Female	-0.350*	-0.359**	-0.347*
	(0.137)	(0.136)	(0.136)
Associate or Bachelor's Degree	0.014	-0.02	0.010
	(0.163)	(0.162)	(0.163)
Graduate Degree	-0.125	-0.146	-0.113
	(0.228)	(0.226)	(0.227)
High Income	0.113	0.112	0.117
	(0.187)	(0.185)	(0.186)
Age: [30–33]	0.316	0.300	0.352+
	(0.194)	(0.193)	(0.194)
Age: [34–41]	0.586**	0.602***	0.608***
	(0.182)	(0.180)	(0.181)
Age: [42–100]	0.773***	0.749***	0.808***
	(0.192)	(0.190)	(0.192)
Working Now	0.072	0.088	0.075
	(0.186)	(0.184)	(0.185)
Born in the US	-0.409	-0.422	-0.409
	(0.367)	(0.364)	(0.365)
Black	-0.641**	-0.704**	-0.680**
	(0.240)	(0.238)	(0.239)
Asian	-0.222	-0.178	-0.200
	(0.309)	(0.306)	(0.308)
Other Race	-0.323	-0.302	-0.292
	(0.371)	(0.367)	(0.369)
Hispanic	-1.584***	-1.631***	-1.608***
	(0.237)	(0.235)	(0.236)
Time Spent (winsorized)		0.050***	
		(0.011)	
Time Spent (not winsorized)			0.009**
			(0.003)
Constant	5.548***	5.257***	5.443***
	(0.462)	(0.462)	(0.461)
Number of Observations	986	986	986
R^2	0.122	0.14	0.132
Adjusted R^2	0.108	0.125	0.117

Standard errors in parentheses. *** p<0.001, ** p < 0.01, * p<0.05, + p>0.1

This table presents coefficient estimates of OLS regressions. The dependent variable is the total number of questions answered correctly (out of a possible ten). In Column (1), we repeate the analysis in Table 4. In Column (2), we add a control for the amount of time the participant spent completing the survey, winsorized at the 5% level (Time Spent (winsorized)). For robustness, in Column (3), we repeat the analysis with the raw time elapsed variable (Time Spent (not winsorized)). High income is a dummy equal to one if the respondent reported a total household income before taxes during the past twelve months of at least $100,000. White is the omitted category for race. Other Race includes respondents who selected multiple categories.

TABLE 6: FULL REGRESSION ESTIMATES, WITH AND WITHOUT GROUP DUMMIES

	(1)	(2)	(3)	(4)	(5)	(6)
	All Questions		Math Questions		Logic Questions	
Group 2		0.575**		0.393***		0.182
		(0.192)		(0.104)		(0.132)
Group 3		0.753***		0.165		0.589***
		(0.192)		(0.104)		(0.132)
Group 4		1.047***		0.421***		0.626***
		(0.191)		(0.104)		(0.131)
Female	-0.314*	-0.350*	-0.261***	-0.278***	-0.053	-0.072
	(0.139)	(0.137)	(0.075)	(0.074)	(0.095)	(0.094)
Associate or Bachelor's Degree	-0.050	0.014	0.109	0.131	-0.160	-0.118
	(0.165)	(0.163)	(0.089)	(0.089)	(0.114)	(0.112)
Graduate Degree	-0.270	-0.125	0.109	0.154	-0.379*	-0.279+
	(0.230)	(0.228)	(0.124)	(0.124)	(0.158)	(0.157)
High Income	0.120	0.113	0.151	0.151	-0.031	-0.038
	(0.190)	(0.187)	(0.103)	(0.102)	(0.130)	(0.129)
Age: [30–33]	0.285	0.316	-0.004	0.008	0.290*	0.308*
	(0.197)	(0.194)	(0.106)	(0.106)	(0.135)	(0.133)
Age: [34–41]	0.552**	0.586**	-0.022	-0.007	0.574***	0.593***
	(0.184)	(0.182)	(0.099)	(0.099)	(0.126)	(0.125)
Age: [42–100]	0.737***	0.773***	0.156	0.166	0.581***	0.607***
	(0.195)	(0.192)	(0.105)	(0.104)	(0.134)	(0.132)
Working Now	0.091	0.072	-0.062	-0.062	0.153	0.134
	(0.188)	(0.186)	(0.101)	(0.101)	(0.129)	(0.127)
Born in the US	-0.401	-0.409	-0.246	-0.230	-0.156	-0.179
	(0.372)	(0.367)	(0.201)	(0.199)	(0.256)	(0.252)
Black	-0.679**	-0.641**	-0.079	-0.058	-0.600***	-0.583***
	(0.243)	(0.240)	(0.131)	(0.130)	(0.167)	(0.165)
Asian	-0.282	-0.222	0.112	0.143	-0.394+	-0.366+
	(0.313)	(0.309)	(0.169)	(0.168)	(0.215)	(0.212)
Other Race	-0.389	-0.323	-0.141	-0.108	-0.247	-0.215
	(0.376)	(0.371)	(0.203)	(0.201)	(0.258)	(0.255)
Hispanic	-1.636***	-1.584***	-0.611***	-0.588***	-1.024***	-0.997***
	(0.240)	(0.237)	(0.130)	(0.128)	(0.165)	(0.162)
Constant	6.200***	5.548***	1.950***	1.662***	4.250***	3.886***
	(0.448)	(0.462)	(0.242)	(0.251)	(0.307)	(0.317)
Number of Observations	986	986	986	986	986	986
R^2	0.094	0.122	0.047	0.068	0.106	0.135
Adjusted R^2	0.082	0.108	0.034	0.053	0.094	0.121

Standard errors in parentheses. *** p<0.001, ** p < 0.01, * p<0.05, + p>0.1

This table presents coefficient estimates of OLS regressions. The dependent variable in Columns (1) and (2) is the total number of questions answered correctly (out of a possible ten). The dependent variable in Columns (3) and (4) is the total number of math questions answered correctly (out of a possible four). The dependent variable in Columns (5) and (6) is the total number of logic questions answered correctly (out of a possible six). High income is a dummy equal to one if the respondent reported a total household income before taxes during the past twelve months of at least $100,000. White is the omitted category for race. Other Race includes respondents who selected multiple categories.

TABLE 7: DEMOGRAPHIC BALANCE

	Group 1 v. Group 2	Group 1 v. Group 3	Group 1 v. Group 4	Group 2 v. Group 3	Group 2 v. Group 4	Group 3 v. Group 4
Female	0.046	0.022	0.064	-0.024	0.018	0.042
	(0.045)	(0.044)	(0.045)	(0.045)	(0.045)	(0.045)
Bachelor's Degree	-0.030	-0.022	-0.030	0.009	0.001	-0.008
	(0.045)	(0.045)	(0.044)	(0.044)	(0.044)	(0.044)
Graduate Degree	-0.076*	-0.121***	-0.085*	-0.045	-0.009	0.036
	(0.035)	(0.033)	(0.034)	(0.030)	(0.032)	(0.029)
High Income	-0.011	-0.012	-0.013	-0.001	-0.003	-0.001
	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)	(0.033)
Age (years)	0.136	-0.500	-0.788	-0.636	-0.924	-0.288
	(0.975)	(0.934)	(0.909)	(0.950)	(0.926)	(0.882)
Working Now	-0.048	-0.020	0.024	0.028	0.072*	0.044
	(0.034)	(0.033)	(0.031)	(0.035)	(0.033)	(0.032)
Born in the US	-0.012	0.008	0.008	0.020	0.020	0.000
	(0.019)	(0.017)	(0.017)	(0.018)	(0.018)	(0.016)
White	0.072*	0.068	0.056	-0.004	-0.016	-0.012
	(0.035)	(0.036)	(0.036)	(0.033)	(0.033)	(0.034)
Black	-0.064*	-0.056*	-0.016	0.008	0.048	0.040
	(0.026)	(0.026)	(0.029)	(0.022)	(0.025)	(0.025)
Asian	-0.004	-0.004	-0.024	0.000	-0.020	-0.020
	(0.022)	(0.022)	(0.020)	(0.021)	(0.020)	(0.020)
Other Race	-0.004	-0.008	-0.016	-0.004	-0.012	-0.008
	(0.018)	(0.018)	(0.017)	(0.017)	(0.016)	(0.016)
Hispanic	-0.032	-0.028	-0.040	0.004	-0.008	-0.012
	(0.027)	(0.028)	(0.027)	(0.026)	(0.025)	(0.025)

Standard errors in parentheses. *** p<0.001, ** p < 0.01, * p<0.05

This table presents pairwise comparisons of each demographic variable across treatment groups. Each demographic variable with the exception of Age is a dummy variable. The reported values are therefore the difference between the average proportion of respondents with the relevant characteristic in the first group, compared to the second group. The reported values are estimated by estimating an OLS regression with a dummy variable representing the second group listed on a sample that includes respondents from the first and second group. High income is a dummy equal to one if the respondent reported a total household income before taxes during the past twelve months of at least $100,000. Other Race includes respondents who selected multiple categories.