Comparing age differences in cognition, personality, and political orientation across six online recruitment platforms

Michael S. Cohen[1,2], Karolina M. Lempert[1,3], David A. Wolk[4], Joseph W. Kable[1*]

[1] Department of Psychology, University of Pennsylvania, Philadelphia, PA
[2] Department of Psychology, University of Chicago, Chicago, IL
[3] Gordon F. Derner School of Psychology, Adelphi University, Garden City, NY
[4] Department of Neurology and Penn Memory Center, Hospital of the University of Pennsylvania, Philadelphia, PA.

Author Note

The authors thank Laura Germine and Corianna Moffatt from the Many Brains Project for their assistance with setting up the cognitive tasks and for providing the Testmybrain normative data, Samuel Gosling, Jeff Potter, and Chris Soto for providing access to the GPIPP personality dataset and for guidance regarding that dataset, and Aaron Moss from CloudResearch for advice about data collection platforms. Supplemental results, data and code from this project are available at https://osf.io/2qd7x/ .

[*]Address correspondence to:
Joseph W. Kable
Department of Psychology
University of Pennsylvania
Philadelphia, PA 19104
E-mail: kable@psych.upenn.edu

Abstract

Aging is typically associated with characteristic changes to cognitive function, personality, and political orientation. While online data collection for psychological research has greatly increased in frequency, there has been little systematic examination of whether online samples are appropriate for studying aging. Here, we examine whether typical age-related differences in cognitive function, personality and political orientation are replicated in online samples. We measured cognitive performance using tests of vocabulary, processing speed, memory, and attention developed by the TestMyBrain Project; personality using the Big Five Inventory; and political orientation following the ANES survey format. At least 200 participants each were sampled from three crowdsourcing websites (Amazon MTurk, CloudResearch MTurk Toolkit, and Prolific) and three panel recruitment websites (Lucid, CloudResearch PrimePanels, and Qualtrics Panels). On all six platforms, consistent with established norms, age was positively correlated with vocabulary performance and negatively correlated with processing speed. Additionally, in all six samples, consistent with prior studies, age was associated with higher Agreeableness, lower Neuroticism, and greater political conservatism. There were some differences between crowdsourced and panel samples, however. Performance on cognitive measures was broadly better for crowdsourced samples. The correlations between age and Openness and Extraversion differed between crowdsourced and panel samples, with trends in panel samples likely more comparable to the ground truth. Finally, MTurk produced some discrepant effects of age on cognition relative to other crowdsourced platforms. Beyond these differences, though, our results are broadly encouraging for the prospect of studying aging via online experiments.

Introduction

Over the past decade, researchers in the social sciences have made increasing use of online data collection. Online data collection has several advantages over in-person data collection, including the ability to collect larger and more diverse samples, greater convenience for researchers and participants, and reduced costs. The COVID-19 pandemic, which greatly reduced opportunities to safely collect in-person data, further accelerated this shift. Still, concerns have been raised about the representativeness of older adults who participate in online studies (e.g., Ogletree & Katz, 2021). There is a need for more work systematically comparing performance across age groups among participants recruited online to test whether online recruitment yields comparable age differences to more traditional recruitment methods.

One major distinction in online recruitment is between crowdsourcing and panel services. On crowdsourcing platforms, such as Amazon MTurk, CloudResearch MTurk Toolkit, and Prolific, participants are presented with a list of available studies. The details of each study are provided, including the tasks participants will complete, the required time commitment, and the compensation amount (in cash). People can then choose to participate in a specific study. Online panels instead build on the approach of market research panels that predated online data collection. These services differ from marketplaces in that they approach participants who are already part of a panel to participate in a specific study. The amount and type of participant compensation is controlled by the panel service. Panels typically have access to a much larger and more diverse sample of participants than crowdsourcing sites, which can allow for greater representativeness, more naivete with regard to experimental manipulations, and more precise targeting (Chandler et al., 2019; Moss et al., 2023). Panels have notable limitations, however. For instance, panel services often limit studies to approximately 20 minutes or less. Additionally,

panel services work with many panel providers, who bid on each study, meaning that multiple sources can be aggregated for a single study and also that different studies with the same panel service do not necessarily draw upon the same sources (Moss et al., 2023). Three of the most widely-used online panel recruitment platforms are Lucid (which in late 2021 was absorbed into another company, Cint), CloudResearch's Prime Panels, and Qualtrics Panels.

Validation of paid online recruitment platforms

Putting aside the question of age differences, many studies have examined other aspects of the validity of online data collection since it first became widespread. Several studies have focused on Amazon's Mechanical Turk ("MTurk"), the first platform to be widely used for recruiting research participants. For instance, Berinsky et al. (2012) replicated characteristic biases in judgment and decision-making tasks in MTurk samples. They also found that MTurk samples are diverse demographically, but do lean more liberal politically, and, at that time, had very few older adults. Other studies have examined whether presentation or response time effects can be reliably measured online. Here, Crump et al. (2013) found that tasks manipulating stimulus presentation times or measuring response times show comparable results in MTurk samples as in traditional samples, with limited exceptions.

In more recent years, alternative platforms have become available that address one of the limitations of MTurk, which is that Amazon does little screening of MTurk workers. The primary mechanism by which workers are screened is rejection rate on prior tasks. However, academic researchers are typically expected to compensate all participants who complete a study regardless of data quality; thus, MTurk accounts generating poor quality data (e.g., bots or inattentive participants) may still have a low rejection rate. Prolific and CloudResearch aim to address this concern. Prolific has its own participant pool, while the CloudResearch MTurk

Toolkit ("CR Toolkit") applies quality filters to the existing MTurk participant pool. (Note that since the data reported here were collected, CloudResearch has begun to phase out CR Toolkit in favor of a new platform, Connect, which, like Prolific, maintains its own participant pool.) The steps implemented by both Prolific and CR Toolkit appear to improve data quality relative to MTurk (e.g., Peer et al., 2021; Hauser et al., 2022; Douglas et al., 2023; Stagnaro et al., 2024). Broadly, Prolific and CR Toolkit participants show better performance on a range of measures, including effect sizes on established tasks, attention check pass rate, consistency of survey responses, and compliance with instructions.

One study found Lucid panels to be more representative on demographic and personality measures than MTurk, while on both platforms classic judgment and decision-making findings replicated (Coppock & McClellan, 2019). Chandler et al. (2019) similarly found that classic behavioral effects replicated in a Prime Panels sample, which was also closer to representative demographically than MTurk and had less prior exposure to common research questionnaires. Peer et al. (2021) found little difference in data quality between MTurk and Qualtrics Panels. Douglas et al. (2023) found, however, that Qualtrics Panels participants performed more poorly on attention check measures than those recruited via CR Toolkit or Prolific, but somewhat better than those recruited from MTurk. Stagnaro et al. (2024) recently assessed many platforms on both data quality and representativeness, generally finding a tradeoff between the two. Specifically, data collected via Prolific and CloudResearch (both Connect and CR Toolkit) showed higher data quality but lower representativeness, while Lucid showed the reverse. MTurk was an exception, showing both low data quality and poor representativeness.

Another consideration is that data quality on different platforms may change over time. For instance, contrary to more recent work indicating better data quality on Prolific and CR

Toolkit, an earlier study by Peer et al. (2017) found no difference in data quality between Prolific and MTurk. There is also evidence for discrete changes in data quality online. For example, some researchers have reported reductions in attentiveness and cognitive performance on both MTurk (Arechar & Rand, 2021) and Lucid (Ternovski & Orr, 2022) when comparing data collected in early 2020 before the COVID pandemic to data collected after the onset of the pandemic through mid-2021. Other researchers observed a notable increase in the frequency of low-quality responses on MTurk beginning in mid-2018 (e.g., Chmielewski & Kucker, 2019; Kennedy et al., 2020), which affected the validity of results (e.g., Ophir et al., 2019). Demographic distortions in the Prolific participant pool emerged after a blogger promoted the service in July 2021 as a money-making opportunity to her audience of mostly young women (Charalambides, 2021). Given these examples, periodic reassessment of data quality on online platforms is advisable.

Testing older adults online

The above studies, however, have not addressed whether samples recruited online reveal the same characteristic patterns of age differences in cognition, personality and political orientation that have been observed in traditionally recruited samples. On one hand, older adults may show worse performance on cognitive tests self-administered at home via a computer or tablet than when a researcher or clinician administers a test in an office. Causes here could include increased cognitive load due to lack of familiarity with the technology, stereotype threat, or other factors that impair performance in older but not young adults. At the same time, it is easy to imagine that older adults who volunteer to participate via online crowdsourcing services are drawn from a subset of the population with higher socioeconomic status and/or cognitive abilities than the typical young adult on the same platforms. Older adult samples in cognitive

aging studies are traditionally recruited using methods such as in-person contacts, newsletters, and doctor's offices, which may achieve a relatively more representative sample of the older adult population. Thus, prior findings that online data is valid across the population at large does not address the issue of whether basic differences between older and young adults replicate when both age groups are recruited online.

Several studies have examined the reliability of self-administered online tests, relative to cognitive tests administered in-person, in older adults (e.g., Cyr, Romero, & Galin-Corini, 2021; Backx, Skirrow, Dente, Barnett, & Cormack, 2020; Feenstra, Murre, Vermeulen, Kieffer, & Schagen, 2018; Assmann et al., 2016; Eve & de Jager, 2014). These studies focus on important questions about possible variability in performance due to reduced technical literacy in older adults and find that performance on computerized tests is generally in alignment with more traditional methods of administration. Other studies have examined age differences online, finding consistency between normative data and large participant samples recruited online. Tasks that have been examined in this manner include vocabulary, digit symbol coding, and working memory (Harthorne & Germine, 2015), sustained attention (Fortenbaugh et al., 2015), as well as working memory span, feature binding, and prospective memory (Logie & Maylor, 2009). In these studies, however, participants were not financially compensated. Thus, these studies do not address possible biases in samples from paid recruitment platforms.

There are three studies, to our knowledge, that have examined cognitive performance in older adults recruited through paid online platforms. One study compared an MTurk sample with older adults recruited through a traditional longitudinal study (Ogletree & Katz, 2021). Here, the MTurk group showed better performance on analogical reasoning and verbal fluency tasks, suggesting that older adults who sign up for online studies may be a higher-functioning group

than the general population. At the same time, Bui, Myerson, and Hale (2015) reported typical age-related declines in processing speed, and expected modulations of this effect, in three sets of participants recruited from MTurk. The Bui et al. study used a similar approach as the present work but was more limited in the scope of tasks addressed and in only examining a single platform. Finally, Greene and Naveh-Benjamin (2022) showed that for one specific memory task, effects of age on performance were comparable between a sample collected via Prolific during the COVID-19 pandemic and an in-lab sample collected pre-pandemic.

None of these prior studies provide a systematic evaluation of age differences on measures of cognition, personality, and political orientation across different online recruitment platforms. The present study provides such an evaluation, examining age differences across six different paid online recruitment platforms, including both crowdsourced and panel populations, on a variety of measures, including cognitive performance, personality, and political preferences. We therefore aim to provide important validation for researchers who wish to examine aging via online studies in future work.

Measures

We chose measures that have demonstrated reliable and reproducible differences between young and older adults in order to determine whether those effects replicate across six online platforms.

Cognitive Performance

To assess cognitive performance, we selected four tasks: vocabulary, digit symbol coding, paired associate memory, and sustained attention. The vocabulary and digit symbol coding tasks used here have been validated in a large online sample as showing comparable age-related differences as the Wechsler Adult Intelligence Scale (WAIS) intelligence test norms. The

vocabulary measure shows a characteristic increase in performance with age, while the digit symbol coding measure shows a characteristic decrease in performance with age (Hartshorne & Germine, 2015). We also examined paired associate memory, a type of memory test that typically shows age-related decreases (e.g., Old & Naveh-Benjamin, 2008). The final task, the gradual-onset continuous performance task (gradCPT), measures individual differences in sustained attention. Fortenbaugh et al. (2015) found that overall accuracy in this task, represented by d', does not show a linear main effect across the adult life span; instead, it gradually increases through young adulthood to a peak in middle age (early 40s), then gradually decreases. Strategy, a second measure quantified using response criterion, shows a negative linear effect with age, with older participants becoming more cautious in responding. While age differences on the gradCPT are not as straightforward as for the other cognitive measures, we chose it to assess whether there are age-related differences in the attentiveness of participants recruited from different paid online participant pools.

Personality

We chose the Big Five Inventory (BFI) to assess personality. Two cross-sectional studies administered this measure to very large samples of participants who signed up voluntarily for an online personality test and found a clear pattern of age differences (Srivastava, John, Gosling, & Potter, 2003; Soto, John, Gosling, & Potter, 2011). Both studies draw on data from the Gosling-Potter Internet Personality Project (GPIPP), but from different time windows: Srivastava et al. (2003) included data collected between 1998 and 2000, while Soto et al. (2011) reported data collected from 2003 to 2009. Both studies showed positive associations with age for Agreeableness and Conscientiousness, and a negative association between age and Neuroticism, from young adulthood to late middle age (age 60 or 65). Extraversion did not show a linear effect

with age in either study. Openness is the only factor to show different effects in the two studies, with Srivastava et al. (2003) finding a decrease while Soto et al. (2011) found an increase with age across the adult life span. Other studies of personality and age, using traditionally recruited samples, have typically observed similar positive associations between age and Agreeableness and negative associations between age and Neuroticism. There are some differences from the GPIPP online samples, however, as traditional studies have consistently found age-related decreases in Extraversion and Openness, as well as a less-consistent age-related decrease in Conscientiousness, especially past age 75 (Donnellan & Lucas, 2008; Graham et al., 2020; Roberts et al., 2006).

Political orientation

To assess political ideology and political party affiliation, we adopted the question format of the American National Election Studies (ANES), which provides data on these questions going back to 1972. One study used the ANES to examine changes over time within and between generations (Fisher, 2020). This analysis showed that beginning in 2008, through at least 2016, the most recent dataset included in that study, each younger generation was more liberal than the one preceding it. There was some evidence for such a relationship in prior decades as well, but less consistently so. This analysis also showed age-related increases in conservatism and Republican Presidential vote share over time within each generation. Other data has similarly suggested increased conservatism with age within generations among Americans (e.g., Peltzman, 2019; Peterson, Smith, & Hibbing, 2020) and in other countries as well (e.g., Tilley & Evans, 2014). Thus, there is good reason to believe that at the time data were collected for the present study, age would be associated with an increase in ideological conservatism and Republican party identity. Here, we analyze effects of age on the 2020 ANES survey as a baseline and

compare it to the six paid online participant pools, aiming to ground future online studies of how age affects political behavior.

Method

Participants

Complete datasets (i.e., participants passing all attention checks and completing all core measures) were obtained from 1829 participants across six recruitment platforms: three crowdsourcing sites (Amazon MTurk, CloudResearch MTurk Toolkit (CR Toolkit), and Prolific) and three panel recruitment platforms (Lucid, Prime Panels, and Qualtrics Panels). Data were not recorded from participants who failed the simple attention checks described below or who did not complete all parts of the study. Of the participants who did complete all parts of the study, 71 were excluded: 56 for entering an age and birth year that did not match, 13 who did not provide enough information to assess their level of education, and 2 who completed the full study twice through both MTurk and CR Toolkit. For participants who completed one or more parts of the study more than once, only the first responses were included for analysis. The final sample included data from 1758 participants.

Participation was limited to U.S. residents who were using a desktop or laptop computer. For MTurk and CR Toolkit, an 85% Amazon MTurk approval threshold was used. For CR Toolkit, we additionally applied the default “CloudResearch Approved Participants” filter. All crowdsourced participants were paid $4 each, while compensation for panel participants was determined by the platform. The intended sample was 210 participants per platform, with an even distribution across the adult age range; Table 1 shows the actual distribution of participants by age. The total sample was 52.1% female, 47.5% male, and 0.4% other/non-binary/did not

report gender (see Table 1 for gender distribution by platform). Data were collected in spring and early summer of 2021, which provides an important reference point for assessing the quality of data collected online at a specific time point relative to the COVID-19 pandemic (see Table 1 for precise dates of data collection for each platform). Costs for each platform are reported in Supplemental Table 1.

Procedure

In order to begin the study, participants had to provide informed consent, and to successfully complete three simple attention checks (selecting an odd number from three choices, selecting a picture of a cat from three choices, and an automated Captcha). They then completed the 44-item Big Five Inventory (BFI; John & Srivastava, 1999) and provided their year of birth. Participants then completed cognitive tasks on Testmybrain.org in the following order: Vocabulary, Paired Associate Memory encoding, Digit Symbol Coding, Paired Associate Memory test, and gradual onset Continuous Performance Test (gradCPT). Vocabulary and Paired Associate Memory were scored as the proportion of items correct. Digit symbol coding was scored as number of items correct. GradCPT provided measures of accuracy (d') and criterion. Participants had the opportunity to view their scores on these four tests after all four were completed. Finally, participants provided demographic and other information, including age (compared against year of birth as an attention check), gender, race/ethnicity, education, income, household size, ZIP code, political party and ideology. Education level was collected categorically, which we later converted to years of education (see Supplemental Table 2). Years of education were manually coded for nine participants who responded "other" to education level and provided additional information.

Political party and ideology were collected in the format of the American National Election Studies (ANES). A single 7-point scale was used for political ideology, ranging from "Extremely Liberal" to "Extremely Conservative". For political party, responses were converted to a 7-point scale based on two questions. First, participants were asked whether they think of themselves as Republican, Democrat, Independent, Other, or No Preference. If they chose "Republican" or "Democrat", a second question asked whether they are "Strong" or "Not Very Strong" in that affiliation; these responses form the ends of the 7-point scale. If they chose another response, they were then asked whether they are closer to the Republican party, the Democratic party, or neither; these responses form the 3 intermediate points of the 7-point scale.

Norms for cognitive measures were provided by the Many Brains Project, corresponding to people of all ages and locations who voluntarily took the same tests for no compensation at Testmybrain.org. We limited the norm data to participants over the age of 18, to correspond with our sample, yielding the following sample sizes: Vocabulary (n = 33321), Digit Symbol Coding (n = 6862), Paired Associate Memory (n = 9085), GradCPT (n = 19023). Normative data for personality measures were taken from the GPIPP dataset, incorporating all data up to 03/25/2015. We included participants who reported living in the United States and ranging in age from 18 to 85, encompassing 2,669,696 data points. Finally, normative data on political ideology and partisan orientation were taken from the 2020 ANES dataset available via the https://electionstudies.org web site. As each norm dataset used somewhat different categories for level of education, the coding for years of education differed slightly for each dataset; conversions from categories to years for each dataset are reported in Supplemental Table 2.

We first examined regressions predicting outcome measures across all six platforms and within each platform, including age and years of education as predictor variables. We then

examined models across all 6 platforms with regressors of age (mean-centered), platform type (crowdsourcing vs. panel), and the interaction between platform type and age, with years of education as a control variable. Main effects indicate differences in cognition/personality/political orientation by platform type at mean age, while interactions reflect that effects of age differ by platform type.

Then, to determine whether models adding terms for main effects of specific platform and interactions between platform and age were merited, we used ANOVA to compare the proportion of variance explained by models with only terms for platform type to models with additional terms for each specific platform. Reference conditions in these models were arbitrarily chosen to be MTurk for crowdsourced data and Lucid for panels. Results from models accounting for effects of individual platforms are only reported when the ANOVA showed that they explain additional variance.

Finally, we compared the data from each individual platform to the normative data. Here, normative data for a given measure was concatenated with the data that we collected from all six platforms. Main effects of age (mean-centered) and years of education were modeled, as were 6 dummy regressors coding main effects for each platform, and 6 regressors coding the interaction between each platform and age.

Regression models were generally fit using the python smf.ols function, part of the statsmodels module. For regressions that included ANES data on political orientation, we used weighted least squares regression via the smf.wls function. This allowed us to incorporate the weights provided in the ANES dataset to better approximate the actual U.S. population; the data that we collected were all given a weight of 1. For plots showing data from each platform, our data were combined with the normative data, and variance based on years of education was

removed by running a regression with a single predictor variable of education for each outcome measure. The mean fitted value from each regression was then added to the residuals, in order to match the original scale. Linear regression lines indicating effects of age on these residuals were computed separately for the normative data and for data that we collected within each platform.

Results

Cognitive Performance Measures

Examining each platform individually, we found the expected age differences in vocabulary and processing speed on all platforms, and the expected age differences in memory and sustained attention on most platforms. Figure 1 shows the relationship between age and each cognitive performance measure on each platform while controlling for level of education. Statistics for effects of age when regressing age and education on each outcome measure within each platform are reported in Table 2. Within each platform, age was positively associated with vocabulary performance and negatively associated with digit symbol coding performance, as expected from prior work. Age was also a negative predictor of paired associate memory, as expected based on prior work, on four of the six platforms, with the two null results on MTurk and CR Toolkit. GradCPT showed expected effects of age on CR Toolkit and Prolific, i.e., no linear effect of age on d' and a negative effect of age on criterion. A different pattern was apparent on MTurk, which showed a strong positive correlation between age and d' and no age differences in criterion, as well as Qualtrics Panels, which showed a marginal positive effect of age on d' with no effect of age on criterion. Effects of both d' and criterion were null on Lucid and Prime Panels.

Regressions combining data across all platforms showed the expected age-related increase in vocabulary (b = 0.0046, t = 17.27, p < 0.001), age-related decrease in digit-symbol coding (b = -0.5005, t = -27.31, p < 0.001), and age-related decrease in paired-associate memory (b = -0.0023, t = -12.42, p < 0.001). The gradCPT showed no effect of age on d' (b = -0.0009, t = -0.72, p = 0.47), as expected, and showed the expected decrease in criterion with age (b = -0.0049, t = -6.68, p < 0.001).

Comparisons Across Platforms

Comparing platform type (crowdsourced vs. panels), crowdsourced samples showed better overall performance than panels on vocabulary (b = -0.1003, t = -9.20, p < 0.001), digit symbol coding (b = -7.2686, t = -9.59, p < 0.001), paired associate memory (b = -0.076, t = -9.92, p < 0.001), and gradCPT d' (b = -0.636, t = -12.30, p < 0.001), as well as a higher gradCPT criterion (b = -0.203, t = -6.58, p < 0.001). There were no significant interactions between platform type and age for any of the cognitive measures, though marginal trends were present for vocabulary (b = 0.0011, t = 1.90, p = 0.058), reflecting a trend towards a larger increase with age for panels, and gradCPT criterion (b = 0.0030, t = 1.86, p = 0.063), indicating that crowdsourced samples showed a larger decrease with age in the tendency to respond impulsively.

Within crowdsourced platforms, model-comparison ANOVAs indicated platform-specific effects in vocabulary (F(4, 566) = 9.92, p < 0.001), digit symbol coding (F(4, 566) = 3.74, p = 0.005), and gradCPT d' (F(4, 566) = 5.95, p < 0.001), but not paired associate memory (F(4, 566) = 1.46, p = 0.21) or gradCPT criterion (F(4, 566) < 1). These platform-specific effects reflected a different pattern in MTurk participants compared to CR Toolkit and Prolific. The CR Toolkit sample showed better overall performance than MTurk on vocabulary (b = .0876, t =

4.58, $p < 0.001$), digit symbol coding (b = 2.4432, $t = 2.02$, $p = 0.044$), and gradCPT d' (b = 0.2419, $t = 2.53$, $p = 0.012$). The Prolific sample showed better overall performance than MTurk on vocabulary (b = 0.0766, $t = 4.05$, $p < 0.001$), but not on any other measures. Compared to MTurk, CR Toolkit and Prolific also displayed less of an increase in vocabulary performance with age (CR Toolkit: b = -0.0041, $t = -3.72$, $p < 0.001$; Prolific: b = -0.0033, $t = -2.93$, $p = 0.004$), a greater decrease in digit symbol coding with age (CR Toolkit: b = -0.2186, $t = -3.15$, $p = 0.002$; Prolific: b = -0.1857, $t = -2.61$, $p = 0.009$), and a less positive effect of age on gradCPT d' (CR Toolkit: b = -0.0190, $t = -3.45$, $p = .001$; Prolific: b = -0.0198, $t = -3.51$, $p < .001$). Prolific also showed a greater decline in paired associate memory than MTurk (b = -0.0022, $t = -2.06$, $p = 0.040$).

Within panels, model-comparison ANOVAs only indicated platform-specific effects for vocabulary ($F(4, 1178) = 2.53$, $p = 0.039$). There were no differences between panel platforms for any of the other accuracy-based measures (all $F < 1$), nor for gradCPT criterion ($F(4, 1178) = 1.40$, p = .23). For vocabulary, Prime Panels did not differ from Lucid in overall performance (b = 0.0264, $t = 1.60$, $p = 0.11$), but did show less of an increase in vocabulary performance with age (b = -0.0021, $t = -2.30$, $p = 0.022$). There were no differences between Qualtrics and Lucid in vocabulary performance or the effect of age on vocabulary (all $t < 1$).

Comparisons with Normative Data

We compared the data from each of the six platforms to norms provided from Testmybrain (Figure 1 – dashed lines). Detailed results and statistics are reported in the Supplement. Table 3 shows the direction of any overall differences between each platform and the normative data at mean age, and the direction of any differences in the effect of age between each platform and the normative data. Across all of the five cognitive measures, overall

performance was generally poorer for MTurk and the three panel platforms (Lucid, Prime Panels, Qualtrics Panels) relative to the norm data. CR Toolkit and Prolific broadly showed equivalent performance relative to the norms, except on paired associate memory, where all platforms showed poorer performance than the norms. Differences in the effects of age broadly do not show a consistent pattern, though on the two measures where the strongest age effects would be expected, CR Toolkit, Prolific, and Prime Panels all show a more negative effect of age, compared to the norms—that is, a weaker than expected increase with age for vocabulary, and a stronger than expected decrease with age for digit symbol coding.

Personality Measures

Within each platform, the effects of age on personality were largely as expected, though with differences between platforms for Extraversion and Openness (Figure 2; Table 4). Specifically, within each of the six recruitment platforms, age reliably predicted personality scores in the positive direction for Agreeableness and Conscientiousness, and in the negative direction for Neuroticism. Extraversion correlated positively with age on two of the three marketplace platforms, MTurk and CR Toolkit, but showed no correlation with age on the three panel platforms or Prolific. Correlations with Openness were inconsistent across platforms –an age-related decrease on Qualtrics Panels, a marginal age-related increase on MTurk, and null effects on other platforms.

Combining across all platforms, we found expected main effects of age; older participants showed decreased Neuroticism (b = -0.0185, t = -16.14, $p < 0.001$), increased Agreeableness (b = 0.0104, t = 11.93, $p < 0.001$), and increased Conscientiousness (b = 0.0112, t = 12.39, $p < .001$). We also observed increased Extraversion with age (b = 0.0043, t = 3.89, $p <$ 0.001) and decreased Openness with age (b = -0.0027, t = -3.02, p = .003), though as noted

above and explored more below, these two effects were not consistent between different platforms in our dataset.

Comparisons Across Platforms

Comparing between platform types, panel participants reported higher Neuroticism (b = 0.2672, $t = 5.52$, $p < 0.001$), lower Openness (b = -0.2608, $t = -6.92$, $p < 0.001$), lower Agreeableness (b = -0.1217, $t = -3.29$, $p = 0.001$), and lower Conscientiousness (b = -0.2199, $t = -5.76$, $p < 0.001$), but no difference in Extraversion ($|t| < 1$), compared to crowdsourced samples. There were also interactions between platform type and age for Extraversion (b = -0.0141, $t = -5.70$, $p < 0.001$), Neuroticism (b = 0.0053, $t = 2.06$, $p = 0.039$), Openness (b = -0.0079, $t = -3.97$, $p < 0.001$), and Conscientiousness (b = -0.0048, $t = -2.37$, $p = 0.018$), but not for Agreeableness ($|t| < 1$). Breaking down the interactions, in crowdsourced samples, there were positive effects of age on Openness (b = 0.0046, $t = 2.58$, $p = 0.010$) and Extraversion (b = 0.0131, $t = 5.76$, $p < 0.001$), while in panel recruitments, there was a negative effect of age on Openness (b = -0.0034, $t = -3.05$, $p = 0.002$) and no effect of age on Extraversion ($|t| < 1$). The other two significant interactions reflected differences in the magnitude but not the direction of the age effect. Effects of age on Neuroticism were more negative for crowdsourced samples (crowdsourced: b = -0.0242, $t = -10.36$, $p < 0.001$; panels: b = -0.0189, $t = -13.51$, $p < 0.001$), and effects of age on Conscientiousness were more positive for crowdsourced samples (crowdsourced: b = 0.0163, $t = 8.87$, $p < 0.001$; panels: b = 0.0114, $t = 10.29$, $p < 0.001$).

Within crowdsourced platforms, model-comparison ANOVAs indicated platform-specific effects in Neuroticism ($F(4, 566) = 9.69$, $p < 0.001$), Agreeableness ($F(4, 566) = 5.77$, $p < 0.001$), and Conscientiousness ($F(4, 566) = 11.80$, $p < 0.001$), but not Extraversion ($F(4, 566) = 1.20$, $p = .31$) or Openness ($F(4, 566) < 1$). The CR Toolkit sample, compared to the MTurk

sample, showed overall lower Neuroticism (b = -0.2611, t = -2.65, p = 0.008) and higher Conscientiousness (b = 0.2083, t = 2.71, p = 0.007), but no overall difference in Agreeableness (b = 0.1164, t = 1.53, p = .13). In contrast, relative to MTurk, the Prolific sample showed higher Neuroticism (b = 0.2632, t = 2.69, p = 0.007), lower Conscientiousness (b = -0.2778, t = -3.64, p < 0.001), and lower Agreeableness (b = -0.1799, t = -2.39, p = 0.017). Relative to MTurk, the CR Toolkit sample also showed less decrease in Neuroticism with age (b = 0.0152, t = 2.68, p = 0.008), less increase in Conscientiousness with age (b = -.0101, t = -2.28, p = 0.023), and less increase in Agreeableness with age (b = -0.0098, t = -2.25, p = 0.025). The Prolific sample also showed less increase in agreeableness with age than MTurk (b = -0.0111, t = -2.47, p = 0.014), but no difference from MTurk in effects of age on Neuroticism (t < 1) or Conscientiousness (b = -0.0052, t = -1.15, p = .25).

Within panel recruitments, model-comparison ANOVAs only indicated platform-specific effects for Agreeableness (F(4, 1178) = 2.47, p = 0.043). Overall Agreeableness was marginally higher for Prime Panels compared to Lucid (b = 0.1015, t = 1.96, p = .051), but there was no difference between Prime Panels and Lucid in the effect of age ($|t|$ < 1). In contrast, there was a greater increase in Agreeableness with age for Qualtrics Panels, relative to Lucid (b = 0.0054, t = 2.16, p = 0.031), but no difference between Qualtrics Panels and Lucid in overall Agreeableness (t < 1).

Comparisons with Normative Data

We compared the data from each of the six platforms to the normative GPIPP dataset (Figure 2 – dashed lines). Detailed results and statistics are reported in the Supplement. Table 5 shows the direction of any differences between each platform and the normative data at mean age, and the direction of any differences in the effect of age between each platform and the

normative data. Compared to the GPIPP sample, all platforms showed lower levels of Extraversion, all platforms except Prolific showed lower Openness, all platforms except CR Toolkit showed lower Agreeableness, and all platforms showed higher Neuroticism except for CR Toolkit, which showed lower Neuroticism, and Prime Panels, which showed no difference. Differences in Conscientiousness were more variable. The effect of age on Extraversion was more positive in all three crowdsourced samples than in the normative data, but was not different from the normative data in any of the three panel platforms. The effect of age on Neuroticism was more negative on all six platforms than in the normative data, and the effects of age on Agreeableness were more positive than in the normative data for all platforms except Prolific. Openness showed a more positive effect of age on MTurk than in the normative data, but a more negative effect of age in samples from Lucid and Qualtrics, with other platforms not differing. Finally, effects of age on Conscientiousness were more positive than the norms for MTurk, Prolific, and Qualtrics but matched the norms on other platforms.

Political Orientation Measures

Within each of the six platforms, we found the expected increases with age in both conservative political ideology and Republican party affiliation (Figure 3; Table 6). Combining across all platforms, there was an age-related increase on both conservative ideology (b = 0.0295, $t = 12.02$, $p < 0.001$) and Republican party affiliation (b = 0.0286, $t = 9.98$, $p < 0.001$).

Comparisons Across Platforms

Comparing platform types, panel recruitments exhibited more conservative ideological orientation overall than crowdsourced samples (b = 0.3667, $t = 3.56$, $p < 0.001$), but there were no differences in political partisanship (b = 0.1510, $t = 1.24$, $p = 0.216$). Neither ideology nor partisanship showed an interaction between platform type and age ($t < 1$). Furthermore,

according to model-comparison ANOVAs, there were no platform-specific effects in ideology or partisanship in the crowdsourced samples (both $F < 1$), nor were there platform-specific effects in ideology ($F < 1$) or partisanship ($F(4, 1176) = 1.88$, $p = 0.11$) in the panel recruitments.

Comparisons with Normative Data

We used the 2020 ANES survey as a normative dataset here, and compared the data from each of the six platforms to the ANES. Detailed results and statistics are reported in the Supplement. Table 7 shows the direction of any differences between each platform and the ANES sample. In all three crowdsourced samples, compared to the ANES, participants were both more ideologically liberal and more likely to identify as Democrats. Panel participants were also more likely to identify as Democrats than ANES participants, but in most cases did not differ in political ideology. Age effects generally did not differ significantly from the ANES sample, with the exception of Lucid, which showed a larger effect of age on both political ideology and partisan identity compared to the ANES sample.

Discussion

This study aimed to determine whether expected age differences in cognitive performance, personality, and political orientation would be observed among participants recruited from paid online platforms. The most striking feature of these data is that age effects consistent with prior work were observed across all six platforms for the majority of measures examined. Regarding cognition, effects on vocabulary and digit symbol tests, which prior work has demonstrated show strong positive and negative age effects, respectively, were particularly consistent and robust. Paired associate memory additionally decreased with age, as would be expected based on prior work, in a manner that did not reliably differ between different

platforms. Sustained attention similarly showed the expected change towards more cautious responding with age, in a manner that differed only slightly between platforms. Regarding personality, Neuroticism showed an expected negative correlation with age, while Agreeableness showed an expected positive correlation with age, and these effects were apparent in data from every platform tested. Conscientiousness also showed a consistent positive correlation with age across all platforms, consistent with large samples of unpaid online participants in the GPIPP dataset (Srivastava et al., 2003; Soto et al., 2011), though not necessarily with in-person samples (cf., Graham et al., 2020). Finally, regarding political orientation, self-reported political ideology and partisan identity showed reliable conservative shifts with age, consistent with prior work and with the 2020 ANES survey data.

We also identified some broad differences between crowdsourced samples and panel recruitments. Although crowdsourced and panel platforms showed similar age differences in cognitive performance, participants in the crowdsourced samples scored better on all four cognitive tasks than those in the panel samples. Additionally, the panel samples all showed reliably worse performance than the TestMyBrain norm data for each of the four cognitive tasks. It is not clear if these differences reflect a motivational bias such that participants in panels are less inclined to stay focused on online cognitive tasks, or if they reflect a selection bias such that participants in panels are truly less cognitively skilled. Further work would be needed to address these issues.

In contrast to the cognitive variables, age effects for many of the Big Five personality variables did differ between crowdsourced and panel samples. The two most notable differences were for Extraversion, which increased in age in crowdsourced samples but showed no effect in panel samples, and for Openness, which increased with age in crowdsourced samples and

decreased in panel samples. The increase in Extraversion with age in crowdsourced samples is particularly anomalous, while the lack of a reliable effect of age in panel samples is more consistent with prior work. The prior literature is less clear regarding the effects of age on Openness. On one hand, the more recent published study from the GPIPP dataset (Soto et al., 2011), and the larger GPIPP sample covering data up to 2015 that we used as a normative comparison, both showed positive effects of age on Openness. On the other hand, many other studies have observed a negative effect of age on Openness, including the earlier subset of the GPIPP data (Srivastava et al., 2003), various in-person samples (e.g., Donnellan & Lucas, 2008; Terraciano et al., 2005), and meta-analyses of longitudinal studies (Roberts et al., 2006; Graham et al., 2020). On balance, we would tentatively conclude that a negative relationship between age and Openness is more likely the ground truth. Thus, the results for both Extraversion and Openness would suggest that, relative to crowdsourced samples, panels may be a more representative cross-section of the population for examining cross-sectional effects of age on personality.

Beyond these broad differences between crowdsourced and panel platforms, we also found that MTurk was a notable outlier among crowdsourced platforms. The MTurk sample showed poorer overall performance on all four cognitive tasks compared to the other two crowdsourced platforms, CR Toolkit and Prolific. MTurk also showed significantly or marginally worse performance than the norms on all four cognitive tasks, while CR Toolkit and Prolific did not show a clear deviation from the norms across tasks. The MTurk sample also exhibited a different pattern of age effects on cognitive tasks. One clear deviation was that older adults on MTurk showed better sustained attention than younger adults on MTurk. MTurk was the only platform on which participants showed a reliable increase with age in gradCPT d’ and

the only platform to show a more positive effect of age on this task than in the Testmybrain norms. Additionally, MTurk participants showed a greater increase in vocabulary performance with age, and less of decrease in digit symbol coding performance with age, than the other two marketplace platforms; both effects suggest relatively higher cognitive performance in older adults on MTurk. These data alone do not clarify whether these differences are due to older adults, younger adults, or both, on MTurk. However, as discussed above, others have seen poorer performance and higher rates of inattentive responding on MTurk compared to other platforms, and both CloudResearch and Prolific actively work to maintain high-quality participant pools. Since more of the participants on these platforms are younger, studies that do not stratify by age recruit largely young adults. Thus, these interactions between crowdsourced platform and age are likely to be driven at least in part by relatively poorer performance among young adult participants. In addition, recruiting an age-stratified sample was more expensive on MTurk than on the other crowdsourced platforms, as MTurk required both a full 40% platform fee plus a separate fee to stratify by age, rather than the discounted 20% fee applied otherwise, including in CR Toolkit. All of these factors suggest that researchers should be cautious about using MTurk to conduct online studies of cognitive aging.

One area where all online platforms deviated from a population sample was in political orientation. Participants on all platforms were at least marginally more likely to identify as Democrats compared to the ANES population sample, and participants on the three crowdsourced platforms were also more likely to identify as ideologically liberal than the ANES sample. This extends prior findings from the early days of online data collection that online participants tend to be more liberal than the population at large (e.g., Berinsky et al., 2012). Still,

participants on all platforms showed the expected relationship between age and increasing conservatism.

Ultimately, these results tell a broadly optimistic story regarding the ability to use online participant recruitment to test questions related to cognitive aging. All of the tested platforms could provide substantial samples of individuals across the lifespan at a lower cost in money and time than a comparable in-person sample. The data from participants on those platforms was also generally consistent with age differences previously established in in-person samples. Thus, we conclude that age differences can be assessed using online participant recruitment, though different platforms may be better suited for different questions. Specifically, filtered crowdsourced samples, such as CR Toolkit and Prolific, may be slightly better suited to cognitive studies, whereas panel samples may be slightly better suited for studies or personality or political orientation where population representativeness is important.

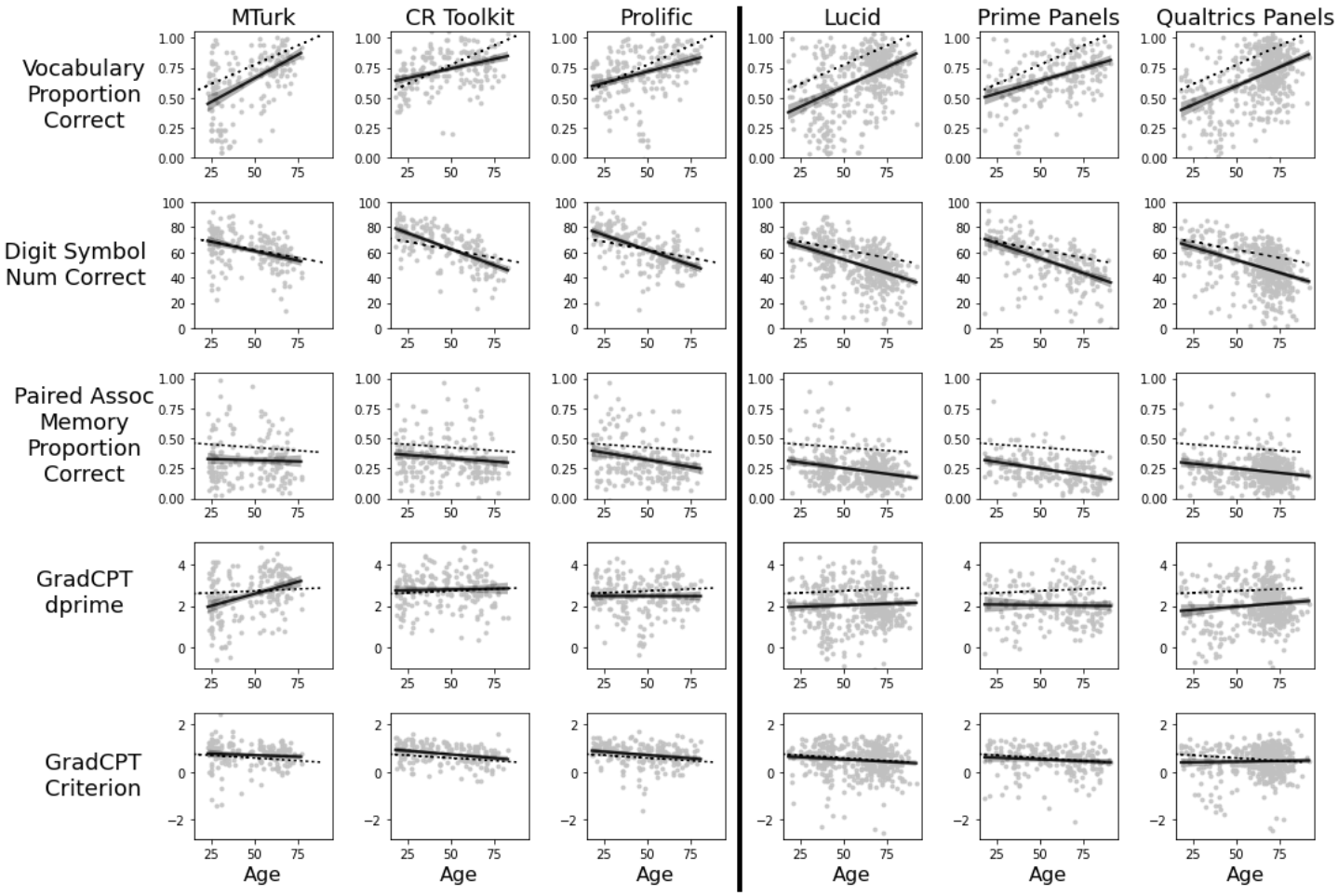

Figure 1. Relationships between age and cognitive measures by platform, controlling for level of education, with solid lines showing linear regression effects and 95% confidence intervals within each platform. Dashed lines represent regression effects from Testmybrain's large norming samples, controlling for level of education.

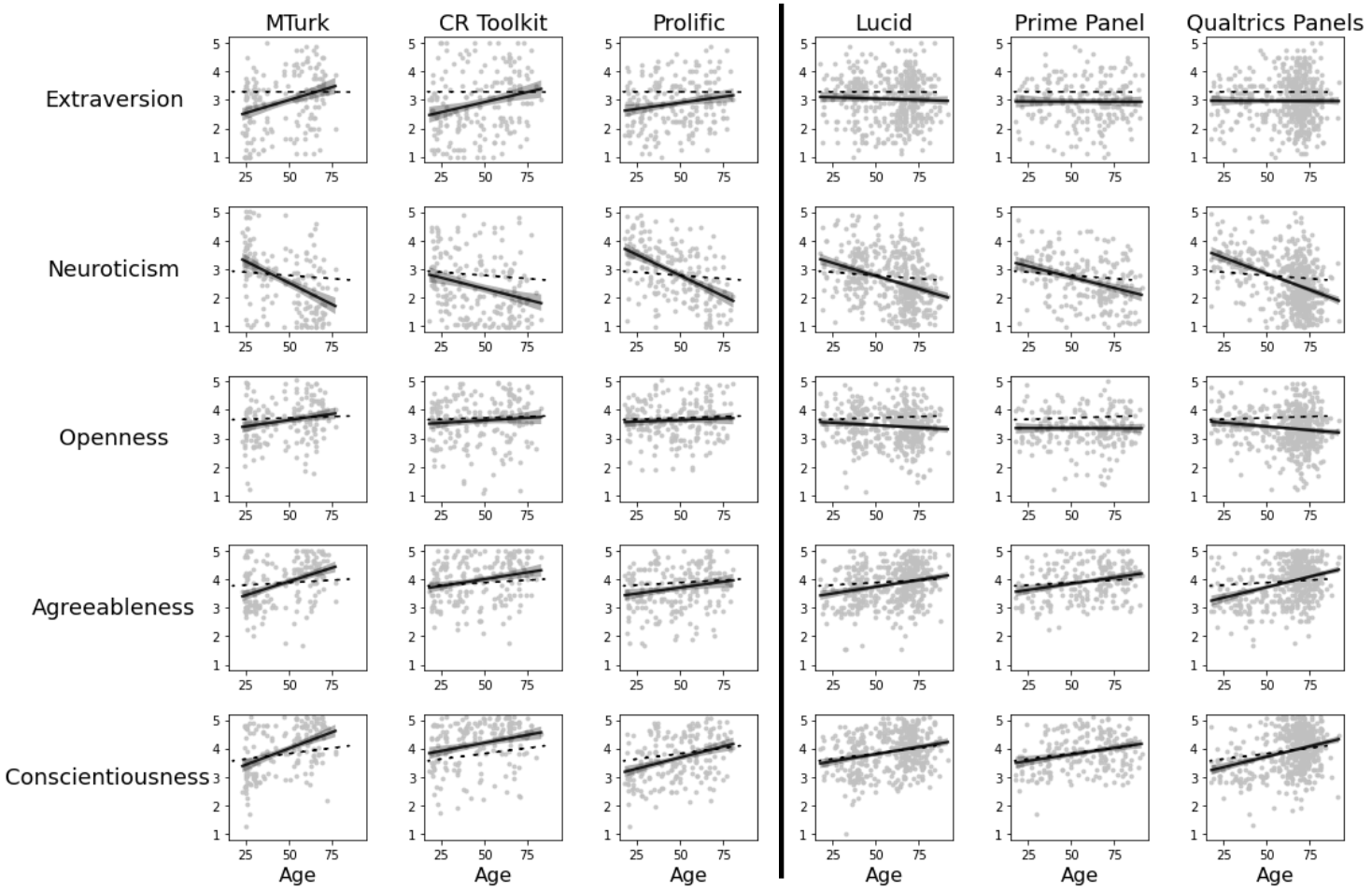

Figure 2. Relationships between age and personality measures by platform, controlling for level of education, with solid lines showing linear regression effects and 95% confidence intervals shown within each platform. Dashed lines represent regression effects from the large GPIPP personality dataset, controlling for level of education.

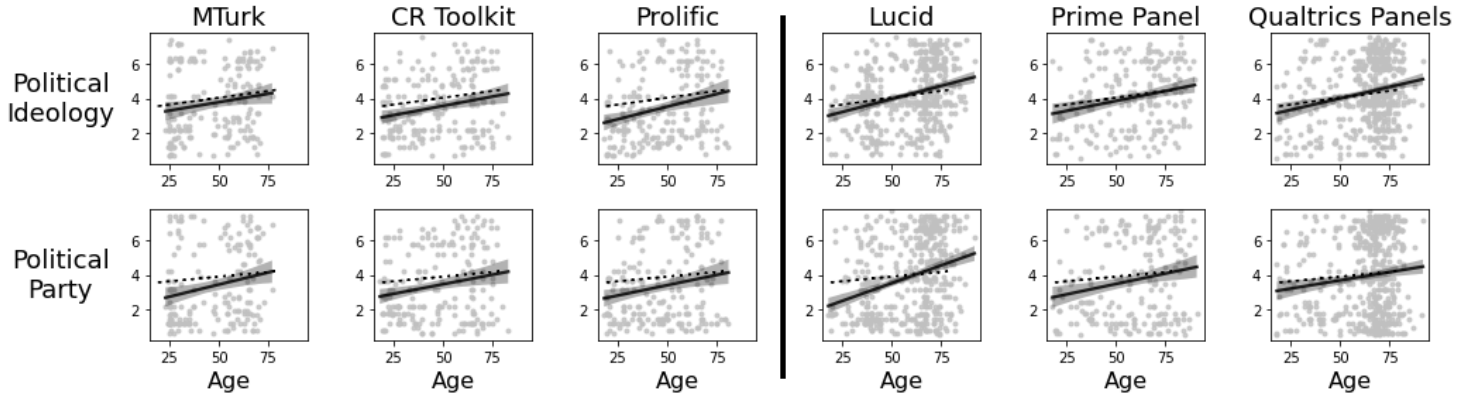

Figure 3. Relationships between age and political measures by platform, controlling for level of education, with solid lines showing linear regression effects and 95% confidence intervals shown within each platform. Dashed lines represent regression effects from the ANES 2020 sample, controlling for level of education.

Table 1. Demographic information by platform

Platform		MTurk	CR Toolkit	Prolific	Lucid	Prime Panels	Qualtrics Panels
n by age group	Age 18-29	47	49	45	23	20	28
	Age 30-39	29	34	35	61	30	30
	Age 40-49	13	26	29	60	21	33
	Age 50-59	29	22	39	51	36	44
	Age 60-69	47	39	35	120	36	180
	Age 70-79	10	24	18	113	33	172
	Age 80+	0	2	1	28	35	31
n by gender	Female	102	98	106	188	115	307
	Male	72	96	94	268	96	209
	Other Gender / No Response	1	2	2	0	0	2
n by education level	Did not finish high school	0	0	1	4	3	11
	High school	21	31	19	70	36	103
	Some College	33	52	40	90	36	108
	Associate's Degree	17	11	15	51	21	56
	Bachelor's Degree	69	83	73	125	62	141
	Master's Degree	32	15	41	87	45	85
	Doctoral Degree	2	3	11	26	8	12
	Other	1	1	2	3	0	2
Dates tested		4/27/21 – 4/28/21	4/27/21 – 4/28/21	4/27/21 – 4/28/21	4/27/21 – 4/29/21	6/4/21 – 6/13/21	7/1/21 – 7/5/21

Table 2. Effects of age on cognitive performance measures within each platform, controlling for level of education. Reported p values for each measure include a Bonferroni-Holm correction for multiple comparisons across the 6 platforms.

	MTurk			**CR Toolkit**			**Prolific**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Vocabulary	**0.0072**	**7.42**	**< 0.001*** **	**0.0032**	**5.58**	**< 0.001*** **	**0.0042**	**5.53**	**< 0.001*** **
Digit Symbol Coding	**-0.3098**	**-5.28**	**< 0.001*** **	**-0.5189**	**-12.83**	**< 0.001*** **	**-0.4802**	**-10.08**	**< 0.001*** **
Paired Assoc Memory	-0.0003	-0.33	0.74	-0.0012	-1.79	0.15	**-0.0024**	**-3.43**	**0.002** **
GradCPT d'	**0.0197**	**4.24**	**< 0.001*** **	0.0015	0.45	1	0.0016	0.44	1
GradCPT criterion	-0.0027	-1.12	0.53	**-0.0064**	**-4.27**	**< 0.001*** **	**-0.0053**	**-2.75**	**0.033***

	Lucid			**Prime Panels**			**Qualtrics Panels**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Vocabulary	**0.0065**	**10.73**	**< 0.001*** **	**0.0043**	**7.10**	**< 0.001*** **	**0.0062**	**11.42**	**< 0.001*** **
Digit Symbol Coding	**-0.4270**	**-10.65**	**< 0.001*** **	**-0.4860**	**-9.25**	**< 0.001*** **	**-0.4066**	**-9.79**	**< 0.001*** **
Paired Assoc Memory	**-0.0019**	**-5.75**	**< 0.001*** **	**-0.0021**	**-5.73**	**< 0.001*** **	**-0.0015**	**-4.74**	**< 0.001*** **
GradCPT d'	0.0023	0.87	1	0.0002	0.58	1	0.0064	2.55	0.056 ~
GradCPT criterion	-0.0037	-2.16	0.12	-0.0037	-1.82	0.21	0.0011	0.64	0.53

Table 3. Direction of effects relative to Testmybrain norms (marginal effects are in parentheses). Up arrows for main effects indicate better overall performance than the norms, while down arrows indicate poorer performance. Similarly, for age effects, up arrows indicate more a positive relationship between age and performance for a given platform relative to the norms, while down arrows indicate the reverse.

	MTurk		CR Tookit		Prolific		Lucid		Prime Panels		Qualtrics Panels	
	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age
Vocabulary	↓	—	—	↓	—	↓	↓	—	↓	↓	↓	—
Digit Symbol	(↓)	↑	↑	↓	(↑)	↓	↓	—	↓	↓	↓	—
Paired Assoc Memory	↓	—	↓	—	↓	—	↓	—	↓	—	↓	—
GradCPT d’	↓	↑	—	—	(↓)	—	↓	—	↓	(↓)	↓	—
GradCPT criterion	—	(↑)	↑	—	↑	—	↓	(↑)	(↓)	(↑)	↓	↑

Table 4. Effects of age on personality measures within each platform, controlling for level of education. Reported p values for each measure include a Bonferroni-Holm correction for multiple comparisons across the 6 platforms.

	MTurk			**CR Toolkit**			**Prolific**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Extraversion	**0.0196**	**4.49**	**< 0.001*** **	**0.0137**	**3.30**	**0.006** **	0.0076	2.18	0.12
Neuroticism	**-0.0300**	**-6.79**	**< 0.001*** **	**-0.0152**	**-3.71**	**< 0.001*** **	**-0.0285**	**-8.30**	**< 0.001*** **
Openness	0.0081	2.46	0.075 ~	0.0033	1.03	0.92	0.0020	0.71	0.95
Agreeableness	**0.0190**	**5.90**	**< 0.001*** **	**0.0093**	**3.25**	**0.003** **	**0.0083**	**2.67**	**0.008** **
Conscientiousness	**0.0209**	**6.34**	**< 0.001*** **	**0.0115**	**3.91**	**< 0.001*** **	**0.0163**	**5.32**	**< 0.001*** **

	Lucid			**Prime Panels**			**Qualtrics Panels**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Extraversion	-0.0016	-0.74	1	-0.0005	-0.20	1	-0.0001	-0.06	1
Neuroticism	**-0.0182**	**-7.89**	**< 0.001*** **	**-0.0155**	**-5.39**	**< 0.001*** **	**-0.0226**	**-9.71**	**< 0.001*** **
Openness	-0.0032	-1.90	0.23	-0.0003	-0.14	0.95	**-0.0051**	**-2.69**	**0.044***
Agreeableness	**0.0093**	**5.34**	**< 0.001*** **	**0.0091**	**4.21**	**< 0.001*** **	**0.0147**	**8.20**	**< 0.001*** **
Conscientiousness	**0.0099**	**5.69**	**< 0.001*** **	**0.0095**	**4.17**	**< 0.001*** **	**0.0146**	**7.65**	**< 0.001*** **

Table 5. Direction of effects on personality measures relative to GPIPP norms (marginal effects are in parentheses). Arrows are defined as described above for Table 3.

	MTurk		CR Tookit		Prolific		Lucid		Prime Panels		Qualtrics Panels	
	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age
Extraversion	↓	↑	↓	↑	↓	↑	↓	—	↓	—	↓	—
Neuroticism	↑	↓	↓	↓	↑	↓	↑	↓	—	↓	↑	↓
Openness	↓	↑	(↓)	—	—	—	↓	↓	↓	—	↓	↓
Agreeableness	↓	↑	—	↑	↓	—	↓	↑	(↓)	↑	↓	↑
Conscientiousness	—	↑	↑	—	↓	↑	—	—	—	—	↓	↑

Table 6. Effects of age on political preference measures within each platform, controlling for level of education. Reported p values for each measure include a Bonferroni-Holm correction for multiple comparisons across the 6 platforms.

	MTurk			**CR Toolkit**			**Prolific**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Political Ideology	**0.0227**	**2.51**	**0.013***	**0.0217**	**3.16**	**0.004****	**0.0298**	**3.72**	**0.001*****
Political Party	**0.0303**	**3.05**	**0.013***	**0.0221**	**2.79**	**0.013***	**0.0236**	**2.68**	**0.013***

	Lucid			**Prime Panels**			**Qualtrics Panels**		
	b	t	Corrected p	b	t	Corrected p	b	t	Corrected p
Political Ideology	**0.0301**	**6.04**	**< 0.001*****	**0.0227**	**3.31**	**0.003****	**0.0268**	**4.87**	**< 0.001*****
Political Party	**0.0409**	**6.97**	**< 0.001*****	**0.0245**	**2.97**	**0.013***	**0.0191**	**2.89**	**0.013***

Table 7. Direction of political orientation effects relative to ANES norms (marginal effects are in parentheses). Arrows are defined as described above for Table 3.

	MTurk		CR Toolkit		Prolific		Lucid		Prime Panels		Qualtrics Panels	
	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age	Main Effect	Age
Political Ideology	↓	—	↓	—	↓	(↑)	—	↑	(↓)	—	—	(↑)
Political Party	↓	—	↓	—	↓	—	↓	↑	↓	—	(↓)	—

Declarations

Conflicts of interest
None of the authors report any conflicts of interest.

Ethics approval
The study was reviewed and approved by the University of Pennsylvania IRB under protocol #828391.

Consent to participate
Informed consent was obtained from all individual participants included in the study.

Consent to publish
No individual participant results are included in the manuscript, and all identifying information is removed from data shared in the OSF repository.

Availability of data and materials
All data that were collected as part of this project are shared on OSF at https://osf.io/2qd7x/ . The tasks and normative data from Testmybrain/Many Brains Project, and the normative data from GPIPP and ANES, are not shared in our OSF repository because we are not the creators/owners of those materials.

Code availability
Analysis code is shared in the OSF repository at https://osf.io/2qd7x/ . Specifically, an input file and script that generate all statistical analyses in the paper, except those involving normative data, are shared. Input files and scripts to generate versions of the figures without normative data are also shared. Finally, we provide the full data processing script for transparency, but without the input files needed to run it, for privacy reasons (i.e., this script uses subject IDs from the platforms to unite different raw data sources) and due to the data ownership concerns on normative data noted above.

Open Practices Statement
See above for availability of data and materials and code availability. This study was not preregistered.

References

Arechar, A. A., & Rand, D. G. (2021). Turking in the time of COVID. *Behavior Research Methods, 53,* 2591-2595.

Assmann, K. E., Bailet, M., Lecoffre, A. C., Galan, P., Hercberg, S., Amieva, H., & Kesse-Guyot, E. (2016). Comparison between a self-administered and supervised version of a web-based cognitive test battery: Results from the NutriNet-Santé cohort study. *Journal of Medical Internet Research, 18,* e68.

Backx, R., Skirrow, C., Dente, P., Barnett, J. H., Cormack, F. K. (2020). Comparing web-based and lab-based cognitive assessment using the Cambridge Neuropsychological Test Automated Battery: A within-subjects counterbalanced study. *Journal of Medical Internet Research, 22,* e16792.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis, 20,* 351-368.

Bui, D. C., Myerson, J., & Hale, S. (2015). Age-related slowing in online samples. *The Psychological Record, 65,* 649-655.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior Research Methods, 21,* 2022-2038.

Chmielewski, M., & Kucker, S. C. (2019). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11,* 464-473.

Charalambides, N. (2021). We recently went viral on TikTok - here's what we learned. Retrieved from: https://www.prolific.co/blog/we-recently-went-viral-on-tiktok-heres-what-we-learned

Coppock, A., & McClellan, O. A. (2019). Validating the demographic, political, psychological, and experimental results obtained from a new source of online survey respondents. *Research and Politics, 6(1).*

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8,* e57410.

Cyr, A. A., Romero, K., Galin-Corini, L. (2021). Web-based cognitive testing of older adults in person versus at home: Within-subjects comparison study. *JMIR Aging, 4,* e23384.

Donnellan, M. B., & Lucas, R. E. (2008). Age differences in the Big Five across the life span: Evidence from two national samples. *Psychology and Aging, 23,* 558-566.

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS ONE, 18,* e0279720.

Eve, C. T., & de Jager, C. A. (2014). Piloting and validation of a novel self-administered online cognitive screening tool in normal older persons: the Cognitive Function Test. *International Journal of Geriatric Psychiatry, 29,* 198-206.

Feenstra, H. E. M., Murre, J. M. J., Vermeulen, I. E., Kieffer, J. M., Schagen, S. B. (2018). Reliability and validity of a self-administered tool for online neuropsychological testing: The Amsterdam Cognition Scan. *Journal of Clinical and Experimental Neuropsychology, 40,* 253-273.

Fisher, P. (2020). Generational cycles in American politics, 1952–2016. *Society, 57*, 22–29.

Fortenbaugh, F. C., DeGutis, J., Germine, L., Wilmer, J. B., Grosso, M., Russo, K., & Esterman, M. (2015). Sustained attention across the life span in a sample of 10,000: Dissociating ability and strategy. *Psychological Science, 26,* 1497-1510.

Graham, E. K., Weston, S. J., Gerstorf, D., Yoneda, T. B., Booth, T., Beam, C. R., Petkus, A. J., Drewelies, J., Hall, A. N., Bastarache, E. D., Estabrook, R., Katz, M. J., Turiano, N. A., Lindenberger, U., Smith, J., Wagner, G. G., Pedersen, N. L., Allemand, M., Spiro, A., Deeg, D. J. H., Johansson, B., Piccinin, A. M., Lipton, R. B., Schaie, K. W., Willis, S., Reynolds, C. A., Deary, I. J., Hofer, S. M., Mroczek, D. K. (2020). Trajectories of Big Five personality traits: A coordinated analysis of 16 longitudinal samples. *European Journal of Personality, 34,* 301-321.

Greene, N. R., & Naveh-Benjamin, M. (2022). Online experimentation and sampling in cognitive aging research. *Psychology and Aging, 37,* 72-83.

Hartshorne, J. K., & Germine, L. T. (2015). When does cognitive functioning peak? The asynchronous rise and fall of different cognitive abilities across the life span. *Psychological Science, 26,* 433-443.

Hauser, D. J., Moss, A. J., Rosenzweig, C., Jaffe, S. N., Robinson, J., & Litman, L. (2022). Evaluating CloudResearch's Approved Group as a solution for problematic data quality on MTurk. *Behavior Research Methods.*

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods, 8,* 614–629.

Logie, R. H., & Maylor, E. A. (2009). An Internet study of prospective memory across adulthood. *Psychology and Aging, 24*(3), 767–774.

Moss A. J., Hauser, D. J., Rosenzweig, C., Jaffe, S., Robinson, J., & Litman, L. Using Market-Research Panels for Behavioral Science: An Overview and Tutorial. *Advances in Methods and Practices in Psychological Science*. 2023;6(2).

Ogletree, A. M., & Katz, B. (2021). How do older adults recruited using MTurk differ from those in a national probability sample?. *The International Journal of Aging and Human Development, 93,* 700-721.

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23*, 104–118.

Ophir, Y., Sisso, I., Asterhan, C. S. C., Tikochinski, R., & Reichart, R. (2019). The Turker blues: Hidden factors behind increased depression rates among Amazon's Mechanical Turkers. *Clinical Psychological Science, 8,* 65-83.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70,* 153-163.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54,* 1643-1662.

Peltzman, S. (2019). Political ideology over the life course. Available at: https://ssrn.com/abstract=3501174

Peterson, J. C., Smith, K. B., & Hibbing, J. R. (2020). Do people really become more conservative as they age?. *The Journal of Politics, 82,* 600-611.

Roberts, B. W., Walton, K. E., & Veichtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin, 132,* 1-25.

Soto, C., John, O. P., Gosling, S. D., & Potter, J. (2011). Age differences in personality traits from 10 to 65: Big Five domains and facets in a large cross-sectional sample. *Journal of Personality and Social Psychology, 100,* 330-348.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*, 1041–1053.

Stagnaro, M.N., Druckman, J.N., Berinsky, A.J., Arechar, A.A., Willer, R., & Rand, D.G. (2024). Representativeness versus response quality: Assessing nine opt-in online survey samples. PsyArXiv preprint. DOI: https://doi.org/10.31234/osf.io/h9j2d

Ternovski, J., & Orr, L. (2022). A Note on increases in inattentive online survey-takers since 2020. *Journal of Quantitative Description: Digital Media, 2,* 1–35.

Terracciano, A., McCrae, R. R., Brant, L. J., Costa, P. T. (2005). Hierarchical Linear Modeling analyses of the NEO-PI–R scales in the Baltimore Longitudinal Study of Aging. *Psychology and Aging, 20,* 493-506.

Tilley, J. & Evans, G. (2014). Ageing and generational effects on vote choice: Combining cross-sectional and panel data to estimate APC effects. *Electoral Studies, 33,* 19-27.

Supplemental Results

In regressions run to compare data from each of the six platforms to normative data for that measure (as the reference condition), the norm dataset for a given task was combined with the data that we collected for that task. Twelve dummy regressors were included to reflect platform main effects and each platform's interaction with age. Main effects of age and education were also modeled. These regressions show that at baseline, i.e., based on the norm data, we see the expected positive effect of age on vocabulary (b = 0.0067, t = 90.48, $p < 0.001$) and negative effects of age on digit symbol coding (b = -0.3899, t = -35.44, $p < 0.001$) and paired-associate memory (b = -0.0014, t = -8.36, $p < 0.001$). There was also a positive effect of age on gradCPT d' (b = 0.0042, t = 9.96, $p < 0.001$), and age predicted a more conservative gradCPT criterion (b = -0.0060, t = -24.84, $p < .001$). Higher level of education also predicted better performance on vocabulary (b = 0.0100, t = 22.28, $p < 0.001$), digit symbol coding (b = 0.5726, t = 11.37, $p < 0.001$), paired-associate memory (b = 0.0136, t = 17.40, $p < 0.001$), and gradCPT d' (b = 0.0438, t = 21.22, $p < 0.001$), and a more conservative criterion (b = -0.0071, t = -5.99, $p < 0.001$). Statistics representing main effects of each platform, and interactions between platform and age, are shown in Supplemental Table 3.

Similar regressions were run to compare personality measures to normative data from the GPIPP dataset incorporating all data up to 03/25/2015. All participants who reported living in the United States and ranging in age from 18 to 85 were included, encompassing 2,669,696 data points. At baseline, we see negative effects of age on extraversion (b = -0.0002, t = -5.06, $p < 0.001$) and on neuroticism (b = -0.0054, t = -118.19, $p < 0.001$). We see positive effects of age on agreeableness (b = 0.0041, t = 111.40, $p < 0.001$), openness (b = 0.0023, t = 61.19, $p < 0.001$), and conscientiousness (b = 0.0093, t = 243.11, $p < 0.001$). Education was associated with

positive effects on extraversion (b = 0.0015, $t = 5.27$, $p < 0.001$), openness (b = 0.0359, $t = 164.63$, $p < 0.001$), and conscientiousness (b = 0.0222, $t = 98.21$, $p < 0.001$), and with negative effects on neuroticism (b = -0.0099, $t = -36.36$, $p < 0.001$) and agreeableness (b = -0.0095, $t = -43.23$, $p < 0.001$). Statistics representing main effects of each platform, and interactions between platform and age, are shown in Supplemental Table 4.

Similar regressions were run to compare data on political orientation and ideology to normative data provided in the 2020 ANES dataset. For political ideology, the baseline ANES sample showed a strong positive main effect of age (b = 0.0176, $t = 15.20$, $p < 0.001$) and a strong negative main effect of education (b = -0.1189, $t = -14.95$, $p < 0.001$), indicating that older participants and those with lower levels of education tended to be more ideologically conservative. With political party identity as the outcome measure, the baseline ANES sample again showed a positive effect of age (b = 0.0137, $t = 8.78$, $p < 0.001$) and a negative effect of education (b = -0.1307, $t = -12.23$, $p < 0.001$), indicating that older and less well-educated participants were also more likely to identify as Republican. Statistics representing main effects of each platform, and interactions between platform and age, are shown in Supplemental Table 5.

Supplemental Table 1. Cost of data collection on each platform.

	Cost
Amazon MTurk	$1,270.40
CR Toolkit	$1,209.83
Prolific	$1,120.00
Lucid	$871.50[a]
Prime Panels	$1,009.33
Qualtrics Panels	$2,102.50[b]

[a] Minimum commitment (for current and future studies) = $1,500

[b] Includes $1,000 required fee for integration with non-Qualtrics data collection platform

Supplemental Table 2. Coding of educational attainment in years based on categorical responses in each dataset

Primary data		**Testmybrain**		**GSIPP**		**ANES**	
Category	**Years**	**Category**	**Years**	**Category**	**Years**	**Category**	**Years**
Did not finish high school	11	Primary	5	Currently in high school	11	8th grade or less	8
High school	12	Middle school	8	Did not finish high school	11	Completed between 9-12 grades, no diploma	11
Some college	13	High school	12	Completed high school	12	High school diploma	12
Associate's Degree	14	Some college	13	Currently in college	13	High school diploma with nonacademic training	13
Bachelor's Degree	16	Technical	13	Some college	13	Some college or Associate's degree	13
Master's Degree	18	College	16	Completed Bachelor's degree	16	Bachelor's degree	16
Doctoral Degree	20	Graduate Degree	18	Currently in graduate school	16	Advanced degree	18
				Graduate or professional degree	18		

Supplemental Table 3. Main effects and age differences relative to Testmybrain norms for each cognitive measure.

	Vocabulary					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.1324**	**-7.53**	**< 0.001*****	0.0010	1.16	0.25
CR Toolkit	0.0112	0.69	0.49	**-0.0034**	**-4.49**	**< 0.001*****
Prolific	-0.0171	-1.04	0.30	**-0.0027**	**-3.43**	**0.001****
Lucid	**-0.1863**	**-12.12**	**< 0.001*****	0	-0.16	0.87
Prime Panels	**-0.1052**	**-5.18**	**< 0.001*****	**-0.0023**	**-3.32**	**0.001****
Qualtrics Panels	**-0.1771**	**-10.06**	**< 0.001*****	-0.0004	-0.73	0.47

	Digit Symbol Coding					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	-1.8980	-1.68	0.093 ~	**0.1134**	**2.04**	**0.041***
CR Toolkit	**3.8845**	**3.71**	**< 0.001*****	**-0.1304**	**-2.68**	**0.007****
Prolific	1.7365	1.65	0.099 ~	**-0.1060**	**-2.09**	**0.037***
Lucid	**-6.4831**	**-6.55**	**< 0.001*****	-0.0328	-0.92	0.36
Prime Panels	**-4.0709**	**-3.13**	**0.002****	**-0.1032**	**-2.32**	**0.020***
Qualtrics Panels	**-6.4809**	**-5.74**	**< 0.001*****	-0.0164	-0.45	0.65

	Paired Assoc Memory					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.1219**	**-6.43**	**< 0.001*****	0.0011	1.29	0.20
CR Toolkit	**-0.0879**	**-5.03**	**< 0.001*****	0.0003	0.35	0.73
Prolific	**-0.0827**	**-4.69**	**< 0.001*****	-0.0010	-1.29	0.20
Lucid	**-0.1604**	**-9.67**	**< 0.001*****	-0.0005	-0.80	0.43
Prime Panels	**-0.1562**	**-7.19**	**< 0.001*****	-0.0009	-1.20	0.23
Qualtrics Panels	**-0.1656**	**-8.75**	**< 0.001*****	0	-0.17	0.87

	GradCPT d'					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.5081**	**-6.28**	**< 0.001*****	**0.0186**	**5.31**	**< 0.001*****
CR Toolkit	0.1165	1.57	0.12	-0.0027	-0.88	0.38
Prolific	**-0.1623**	**-2.16**	**0.031***	-0.0043	-1.33	0.18
Lucid	**-0.6611**	**-9.40**	**< 0.001*****	-0.0015	-0.66	0.51
Prime Panels	**-0.5834**	**-6.36**	**< 0.001*****	-0.0051	-1.84	0.066 ~
Qualtrics Panels	**-0.7991**	**-9.98**	**< 0.001*****	0.0022	0.98	0.33

	GradCPT criterion					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	0.0608	1.32	0.19	0.0037	1.87	0.062 ~
CR Toolkit	**0.1746**	**4.12**	**< 0.001*****	-0.0002	-0.14	0.89
Prolific	**0.1293**	**3.01**	**0.003****	0.0001	0.06	0.95
Lucid	**-0.0932**	**-2.32**	**0.020***	0.0023	1.86	0.063 ~
Prime Panels	-0.1027	-1.96	0.050 ~	0.0029	1.83	0.067 ~
Qualtrics Panels	**-0.2718**	**-5.95**	**< 0.001*****	**0.0071**	**5.49**	**< 0.001*****

Supplemental Table 4. Main effects and age differences relative to GPIPP norms for each personality measure.

	Extraversion					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.6646**	**-7.35**	**< 0.001*****	**0.0184**	**4.87**	**< 0.001*****
CR Toolkit	**-0.6526**	**-7.89**	**< 0.001*****	**0.0143**	**4.33**	**< 0.001*****
Prolific	**-0.5643**	**-6.72**	**< 0.001*****	**0.0088**	**2.53**	**0.011***
Lucid	**-0.2051**	**-2.62**	**0.009****	-0.0016	-0.67	0.50
Prime Panels	**-0.3424**	**-3.35**	**0.001****	0.0001	0.02	0.99
Qualtrics Panels	**-0.3199**	**-3.59**	**< 0.001*****	0.0001	0.04	0.97

	Neuroticism					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**0.2435**	**2.76**	**0.006****	**-0.0245**	**-6.64**	**< 0.001*****
CR Toolkit	**-0.2526**	**-3.14**	**0.002****	**-0.0100**	**-3.12**	**0.002****
Prolific	**0.4851**	**5.93**	**< 0.001*****	**-0.0237**	**-7.04**	**< 0.001*****
Lucid	**0.2485**	**3.26**	**0.001****	**-0.0126**	**-5.48**	**< 0.001*****
Prime Panels	0.1520	1.53	0.13	**-0.0100**	**-3.42**	**0.001****
Qualtrics Panels	**0.4210**	**4.85**	**< 0.001*****	**-0.0171**	**-7.25**	**< 0.001*****

	Openness					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.2093**	**-2.96**	**0.003****	**0.0064**	**2.15**	**0.032***
CR Toolkit	-0.1180	-1.82	0.068 ~	0.0013	0.52	0.60
Prolific	-0.0754	-1.15	0.25	0.0000	0.01	0.99
Lucid	**-0.1396**	**-2.28**	**0.023***	**-0.0056**	**-3.05**	**0.002***
Prime Panels	**-0.3166**	**-3.96**	**< 0.001*****	-0.0023	-0.98	0.33
Qualtrics Panels	**-0.1509**	**-2.16**	**0.031***	**-0.0074**	**-3.89**	**< 0.001*****

	Agreeableness					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.2700**	**-3.79**	**< 0.001*****	**0.0149**	**4.99**	**< 0.001*****
CR Toolkit	0.0014	0.02	0.98	**0.0054**	**2.09**	**0.036***
Prolific	**-0.2649**	**-4.01**	**< 0.001*****	0.0044	1.60	0.11
Lucid	**-0.2579**	**-4.19**	**< 0.001*****	**0.0053**	**2.83**	**0.005****
Prime Panels	-0.1407	-1.75	0.08 ~	**0.0047**	**2.00**	**0.045***
Qualtrics Panels	**-0.3863**	**-5.51**	**< 0.001*****	**0.0106**	**5.53**	**< 0.001*****

	Conscientiousness					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	-0.0966	-1.32	0.19	**0.0131**	**4.26**	**< 0.001*****
CR Toolkit	**0.3136**	**4.68**	**< 0.001*****	0.0021	0.79	0.43
Prolific	**-0.2752**	**-4.05**	**< 0.001*****	**0.0067**	**2.38**	**0.017***
Lucid	-0.0422	-0.67	0.51	0.0007	0.36	0.72
Prime Panels	-0.0468	-0.57	0.57	0.0002	0.09	0.93
Qualtrics Panels	**-0.2373**	**-3.29**	**0.001****	**0.0053**	**2.69**	**0.007****

Supplemental Table 5. Main effects and age differences relative to ANES for political measures.

	Political Ideology					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.3096**	**-2.32**	**0.021***	0.0020	0.27	0.79
CR Toolkit	**-0.5196**	**-4.17**	**< 0.001*****	0.0041	0.62	0.54
Prolific	**-0.5444**	**-4.38**	**< 0.001*****	0.0117	1.71	0.088 ~
Lucid	-0.0963	-1.13	0.26	**0.0128**	**2.66**	**0.008****
Prime Panels	-0.2302	-1.88	0.060 ~	0.0056	0.90	0.37
Qualtrics Panels	-0.0750	-0.81	0.42	0.0091	1.80	0.071 ~

	Political Party					
	Main Effect			**Age**		
	b	t	p	b	t	p
MTurk	**-0.4568**	**-2.54**	**0.011***	0.0146	1.45	0.15
CR Toolkit	**-0.4518**	**-2.69**	**0.007****	0.0085	0.97	0.34
Prolific	**-0.5079**	**-3.04**	**0.002****	0.0101	1.09	0.27
Lucid	**-0.3466**	**-3.06**	**0.002****	**0.0276**	**4.32**	**< 0.001*****
Prime Panels	**-0.4333**	**-2.71**	**0.007****	0.0105	1.31	0.19
Qualtrics Panels	-0.2348	-1.94	0.052 ~	0.0054	0.83	0.41