Current Psychology (2022) 41:2241–2249
https://doi.org/10.1007/s12144-020-00741-0

Examining broad intellectual abilities obtained within an mTurk internet sample

Zachary C. Merz[1] · John W. Lace[2] · Alexander M. Eisenstein[2]

Published online: 22 April 2020

Abstract
Widely used in social science research, samples of participants obtained via Amazon's Mechanical Turk (mTurk) tend to be representative across many sociodemographic variables. However, to date, no research has investigated and reported the global cognitive ability level (i.e., intelligence) of samples obtained via mTurk. The present study contributes to the literature by investigating a previously well-validated, public domain measure of cognitive ability in a sample of American adults recruited via mTurk. As part of a larger cross-sectional, survey-based study, four hundred thirty-four (434) Americans (*M* age = 37.86; 35.7% men) completed a demographic questionnaire and the 16-item International Cognitive Ability Resource, Sample Test (ICAR-16). Results revealed a normal distribution of ICAR-16 scores across the current sample. Additionally, total scores were positively correlated with participants' level of education, income, and self-estimated intelligence, but did not significantly correlate with participant age. No gender differences were identified on ICAR-16 total scores. Finally, ICAR-16 scores did not significantly differ from normative data derived from its validation study. These results suggested that American mTurk samples may be representative of the broader population in terms of global cognitive ability, and that the ICAR-16 is likely a reasonable, psychometrically sound, and inexpensive measure of global cognitive ability appropriate for use in mTurk samples.

Keywords mTurk · Cognitive ability · ICAR-16 · Psychometric intelligence

As the internet has ushered in new ways for humans to communicate in the twenty-first century, so has it provided novel data collection methods for behavioral and psychological scientists. Using internet samples has frequently been found to be more efficient and economical than in-person lab samples (Gosling et al. 2004), especially for survey-based studies. One method for data collection, crowdsourcing, or the use of large numbers of people to accomplish a task over the internet, has been steadily gaining popularity across many disciplines. However, as crowdsourcing methods increase in popularity, assessing internal and external validity characteristics of data collected through these methods remains important to ensure the generalizability of results. The current study sought to expand on the knowledge regarding representativeness and generalizability of data collected via Amazon Mechanical Turk (mTurk), a commonly utilized crowdsourcing platform.

The mTurk platform was originally intended as a way for companies, or "requesters" in mTurk parlance, to recruit users, or "workers", to complete "human intelligence tasks" (HITs) that were thought to be too complex for computers. Examples of HITs included transcribing spoken language recordings (Marge et al. 2010) and evaluating Wikipedia article quality (Kittur et al. 2008). The utility of this service was soon recognized by social science researchers, who realized they could create HITs in the form of surveys, questionnaires, or experimental manipulations, post them to mTurk, and use the built-in services to collect data.

Data obtained via mTurk have been shown to have several advantages to other online collection methods as the system integrates recruitment, data collection, and participant compensation in one user-friendly system (Buhrmester et al. 2011), thereby making it more cost-effective and user-friendly than traditional participant pools (Berinsky et al.

✉ Zachary C. Merz
zachary_merz@med.unc.edu

[1] Department of Physical Medicine and Rehabilitation, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[2] Department of Psychology, Saint Louis University, Saint Louis, MO, USA

2012). This approach allows researchers with limited funds to conduct studies that may not have been affordable, or allows them to access a larger number of participants than would be otherwise impossible (Johnson and Borden 2012).

As interest in mTurk has grown among researchers, questions about the reliability and validity of procured data have arisen. Multiple studies have examined characteristics of worker samples, with findings revealing high rates of internal consistency, inter-rater reliability, and high test-retest reliability (Berinsky et al. 2012; Buhrmester et al. 2011; Mason and Suri 2011; Mason and Watts 2009). Further examinations suggested that mTurk, as well as other online labor markets, can protect data quality from noise resulting from experimenter and "John Henry" effects (i.e., the tendency for participants in a control group to perceive themselves at a disadvantage and work harder in order to overcome this perceived deficit) because workers complete HITs in private (Horton et al. 2011).

To better assure that erroneous responses and invalid data are reduced, mTurk requesters (i.e., researchers) can rate a worker (i.e., participant) based on his or her satisfactory completion of a HIT. Requesters can exclude workers based on this approval rating and studies have recommended a 95% approval rating for ensuring quality data (Peer et al. 2014), while also ensuring that each worker may only complete the task a single time. A legitimate question is whether workers discuss studies with each other, labeled "cross-talk", due to the existence of mTurk discussion forums. However, this appears infrequent as only 26% of mTurk workers reported knowing another individual utilizing the platform, and only 13% reported seeing a discussion about study content in an online forum (Chandler et al. 2012). Readers are encouraged to review Mason and Suri (2011) for recommendations on how to mitigate data quality pitfalls on mTurk. Although more data would help solidify its reputation, these studies suggest appropriate reliability and validity characteristics of data obtained from mTurk. However, the question of whether data from mTurk is externally valid and generalizable to a larger population remains.

It has been argued that journal editors and publishers consider the importance of external validity as secondary to that of measures of internal validity, and convenience samples are therefore often written off for having less than optimally generalizable results (Landers and Behrend 2015). For example, the heterogeneity of most social science research participant pools, often described as WEIRD (i.e., Western, educated, industrialized, rich, and democratic), is a barrier to generalizing results to a larger population that by and large would not be described that way (Henrich et al. 2010a, b). Because mTurk is an online service and available worldwide to anyone with an internet connection, it has the potential for a more diverse worker pool than other types of convenience samples. As recent use of mTurk in research has accelerated, the question of whether the worker samples are representative of populations of interest needs to be examined (Harms and DeSimone 2015).

Multiple studies have compared demographic information of mTurk workers to the general public of the United States of America (USA). Workers come predominantly from the USA and India (Ipeirotis 2010, Eriksson and Simpson 2010), and mTurk allows for the recruitment of workers from preselected specific geographic locations. There is consensus that among USA samples, workers are more ethnically, socio-economically and geographically diverse than traditional laboratory participants, mainly because most lab samples are comprised of undergraduate psychology students (Berinsky et al. 2012; Buhrmester et al. 2011). American mTurk samples tend to be younger than other internet-procured samples, but older than traditional American college samples, and slightly more female than traditional (i.e., non-Internet based) samples (Buhrmester et al. 2011; Paolacci et al. 2010; Ipeirotis 2010). Workers tend to be more educated than the general public (Paolacci et al. 2010; Ipeirotis 2010, Ross et al. 2009). Despite the higher levels of education, income distributions of USA workers are skewed to lower wages than is representative of the USA population, which may indicate some motivation for participation in mTurk, given that participants are paid for their time (Paolacci et al. 2010). MTurk workers also show greater diversity of employment fields than the general population, but are disproportionately dominated by individuals in the tech industries, and have large numbers of participants who described being under- or unemployed (Keith and Harms 2016). Workers also tended to be single and childless (Ipeirotis 2010). As such, the demographic characteristics of mTurk workers support the argument that this population should represent the general public more accurately than other data collection pools, particularly university-based undergraduate samples (Burnham et al. 2018).

Additionally, researchers in the realm of psychology would benefit from known comparisons between performance of mTurk workers and other participants in psychological research settings. Classic experiments from the fields of behavioral economics and psycholinguistics have been successfully replicated on mTurk, indicating that workers make decisions similar to traditional laboratory samples (Horton et al. 2011; Sprouse 2011; Suri and Watts 2011). Results from other simple behavioral tasks have also been compared to those from in-person lab participants to show no significant differences (Casler et al. 2013; Eriksson and Simpson 2010; Paolacci et al. 2010).

Despite several efforts to better understand common demographics of individuals completing research tasks on mTurk, less research comparing mTurk workers to traditional participants has been performed in the area of cognitive abilities. Using a technique called the Instructional Manipulation Check (IMC; Oppenheimer et al. 2009), researchers have produced mixed results regarding attentiveness of mTurk samples. Hauser and Schwarz (2016) found that workers performed better on the IMC and had a larger effect in responses to text manipulation than student samples, but another study

using the IMC found that workers performed more poorly than a student sample, although the second study did not exclude non-native English speakers or non-USA participants, which may confound results (Goodman et al. 2013). Crump et al. (2013) replicated a series of classic cognitive behavioral tasks on mTurk. They found that workers performed statistically similarly to lab participants in the Stroop test, task-switching, flanker task, and other measurements of response time in relation to visual attention (Crump et al. 2013).

As the effort to utilize online and computer-based methods of cognitive assessment increases, these psychological dimensions will undoubtedly continue to be examined. However, as it stands, many of these studies have focused on cognitive domains of processing speed and attention and we are not aware of any studies which have attempted to shed light on the broad, collective intellectual abilities of mTurk workers. While cognitive domains of processing speed and attention may correlate with intellectual abilities (e.g., Ren et al. 2018), they remain distinct constructs deserving of individual attention in psychological research (Sternberg 2019). It is important to examine the distribution of intellectual abilities among mTurk workers as another indication of their representativeness of the general population. As such, the current study sought to provide initial data surrounding broad intellectual functioning within a large group of mTurk workers using a well-validated, public domain measure of general cognitive ability.

Consistent with previous research, we hypothesized that while overall intelligence scores would not yield observable gender differences (Halpern and LaMay 2000), differences would emerge across individual subdomains, with more spatial tasks favoring males (Halpern and Collaer 2005; Reilly et al. 2017). We further hypothesized that scores would be normally distributed across our obtained sample, commensurate with normative data reported for the measure of interest, and correlated with constructs known to covary with intelligence, thus demonstrating continued support for the external validity and generalizability of data obtained via this platform.

Method

Participants

Participants were recruited via Amazon mTurk's online crowdsourcing platform and participated as part of a larger study. As described above, this method of obtaining psychological data has been shown to be more biopsychosocially diverse compared to samples of university students, specifically in terms of age, gender, racial/ethnic identity, socioeconomic status, and geographic representation (Buhrmester et al. 2011; Casler et al. 2013; Paolacci & Chandler 2014; Sprouse 2011). Further studies have concluded that data obtained via mTurk procedures exhibited data quality often exceeding data collected via traditional methods (Kees et al. 2017).

Participants were residents of the USA aged between 18 and 89 years. Stipulations were put in place within the mTurk platform so that no worker could complete the study instrument more than one time. Sample characteristics are outlined in Table 1.

Measures

Demographic Questionnaire All participants completed a questionnaire aimed at gathering common demographic information used to examine the generalizability of the obtained sample to the general population. Demographic information included participant's age, gender, years of completed education, ethnic identity, approximate annual income, and residing geographic region of the USA.

International Cognitive Ability Recourse (ICAR) The ICAR (Condon and Revelle 2014) was created as a public-domain assessment of cognitive abilities for use within clinical and research pursuits. The original 60-item ICAR was composed of questions designed to assess four distinct domains of cognition: letter and number series (i.e., pattern recognition; 9 items), matrix reasoning (i.e., nonverbal abstract reasoning;

Table 1 Sample characteristics

Variable	Mean (SD)
Age	37.86 (13.77)
Years of Education	15.62 (2.66)
Income (USD)	57,429 (44,767)
Gender	% of sample
Male	35.7
Female	63.8
Other	0.5
Race/Ethnicity	
Caucasian	75
Asian/Asian-American	8
Black/African-American	7
Hispanic or Latinx	6
Multiracial	3
Other	1
Geographic Region	
Southeast	26
Midwest	24
Northeast	21
West	16
Southwest	13

SD = Standard Deviation; USD = United States Dollars
$N = 434$

11 items), verbal reasoning (i.e., general knowledge, vocabulary, and logic; 16 items), and three-dimensional rotation (i.e., visuospatial/perceptual abilities; 24 items). The individual validation sample for this measure included 96,958 individuals across 199 countries; readers are referred to Condon and Revelle (2014) for additional demographic information and to Kirkegaard and Nordbjerg (2015) for information on its cross-cultural validity and applications.

In its validation study, the ICAR exhibited excellent internal consistency ($\alpha = 0.93$). Internal consistency scores for the four intended subdomains include: letter and number series (LN; $\alpha = 0.77$); matrix reasoning (MR; $\alpha = 0.68$); verbal reasoning (VR; $\alpha = 0.76$); and three-dimensional rotation (R3D; $\alpha = 0.93$). Results of initial exploratory factor analyses (EFAs) on the 60-item measure suggested three to five factor solutions based on visual scree plot inspection. However, the anticipated four-factor solution appeared to exhibit superior fit (Root Mean Square Error of Approximation [RMSEA] = 0.058, Root Mean Square Residual = 0.05) and acceptable reliability (Tucker-Lewis Index = 0.71) compared to three- and five-factor solutions. Subsequent EFA suggested excellent fit for a four-factor solution (RMSEA = 0.014, Tucker-Lewis Index = .99; Condon and Revelle 2014).

Stemming from these analyses, a shorter, 16-item version of the ICAR, named the ICAR Sample Test (ICAR-16), was created as a brief yet compendious measure and included four questions from each of the utilized cognitive domains (i.e., LN, MR, VR, and R3D). Internal consistency values for the ICAR-16 were also strong ($\alpha = 0.81$) and factor analyses suggested a four-factor solution which aligned with desired cognitive domains with no evidence of cross-loading (Condon and Revelle 2014; Young et al. 2019).

Additionally, Condon and Revelle (2014) sought to examine how the ICAR and ICAR-16 scores correlated with commercially available and well-known assessments of cognitive ability. Pearson's correlations with the combined score on the Scholastic Aptitude Test (SAT) were strong for both the ICAR ($p = 0.54$) and ICAR-16 ($p = 0.50$), as well as with the American College Test (ACT) standardized assessment (ICAR $p = 0.49$; ICAR-16 $p = 0.46$). Scores on the ICAR-16 were also strongly correlated ($p = 0.82$) with the Shipley-2, a commercially available assessment of general cognitive functioning (Condon and Revelle 2014; Shipley et al. 2009). Overall, while acknowledging that the ICAR and ICAR-16 are less well-known and exhaustive compared to comprehensive, commercially available cognitive assessments, they offer viable alternatives, especially for internet-based samples. Based on these findings, we elected to utilize the ICAR-16 within the current study.

Embedded Validity Checks To ensure non-random responding and provide additional credence to the internal validity of the obtained results, a total of four validity checks were included within the ICAR-16 items. These included basic questions (i.e., "Please select the matching shape" or "How many dots are pictured?"). These questions were designed to be exceptionally easy (e.g., counting five plainly described dots on the screen or matching a rectangle to a rectangle rather than a star or circle) to detect sub-optimal effort. Two validity checks were presented randomly within the first eight ICAR-16 items and two were presented randomly within the last eight ICAR-16 items. These items are available by request from the first and second author for use in future research.

Procedure

Data was collected following institutional review board (IRB) approval. Participants were provided with an electronic informed consent document outlining risks, benefits, and rationale for the current study. Voluntary completion of the survey served as provision of informed consent. ICAR-16 items were administered in two "sets" (i.e., screens in the mTurk interface), with each set containing 8 ICAR-16 items, two of each type (i.e., two LN, two MR, two VR, and two R3D). Two original validity check items created by the authors were embedded within each set in addition to other validity checks within the larger study. Items within each set (i.e., ICAR-16 and validity checks) were presented in randomized order, consistent with Condon and Revelle (2014). Participation was immediately ceased if one or more embedded validity checks were answered incorrectly. Only participants without evidence for sub-optimal effort who completed all 16 items of the ICAR-16, as well as four additional ICAR-16 validity checks, were included in the final dataset and subjected to statistical analysis. Participants completed a brief demographic questionnaire, which included items regarding age, gender, income, and self-estimated intelligence, at the end of the study consistent with methodological recommendations (Stoutenbourgh 2008).

Results

A total of 696 surveys were attempted. Overall, 434 of these attempts (62.4%) completed the full measure and passed all embedded effort checks. As such, reported data reflects completed and validated survey submissions ($N = 434$). ICAR-16 items demonstrated excellent internal consistency ($\alpha = .80$) and the four-factor solution revealed excellent fit ($\chi^2/df = 2.89$; Goodness of Fit Index = .99; Normed Fit Index = .97; Standardized Root Mean Square Residual = .03; Hu and Bentler 1999; Schreiber et al. 2006; Shevlin and Miles 1998; Tabachnik and Fidell 2013). Table 2 displays mean and standard deviation scores for the ICAR-16 and its subscales in the present sample. While skewness and kurtosis values for each variable did not indicate non-normality (Kim 2013), visual

Table 2 Descriptive data for ICAR-16 scores and results of one-sample t-tests in the present sample

Variable	Mean (SD) in present study	Skewness	Kurtosis	Mean per Condon and Revelle (2014)	Holm-Bonferroni corrected p
ICAR-16 Total	8.00 (3.64)	<.01	−.83	7.89	>.99
ICAR-16 LN	2.35 (1.32)	−.32	−1.04	2.25	.37
ICAR-16 VR	2.84 (1.24)	−.75	−.56	2.70	.11
ICAR-16 MR	2.00 (1.25)	.03	−.98	2.10	.45
ICAR-16 R3D	.82 (1.14)	1.40	1.06	.84	.66

SD = Standard Deviation; No ICAR16 variables were indicative of notable non-normality according to absolute comparison values for skewness (2) and kurtosis (7) provided by Kim (2013). Holm-Bonferroni corrected *p* represents significance between data obtained in the present sample and composite scores computed from normative item-level data in Condon and Revelle (2014) via one-sample *t*-tests

$N = 434$

inspection revealed non-normal patterns for LN, VR, and R3D scales. MR and the ICAR-16 Total scales revealed generally appropriate normal distributions via visual inspection. Statistical tests of normality were not conducted due to the likelihood of Type 1 error (i.e., incorrect assumption of non-normality) in large sample sizes (Field 2013; Oztuna et al. 2006). Visual depiction of ICAR-16 total score distribution is provided in Fig. 1.

To determine whether participant gender meaningfully related to ICAR-16 variables, a *t*-test (for the total score) and a multivariate analysis of variance (MANOVA) were performed utilizing ICAR-16 subtest scores (i.e., LN, MR, VR, and R3D) entered as dependent variables. Those reporting "other" gender identity ($n = 2$) were excluded from these analyses as group comparisons between these individuals and notably larger groups of males ($n = 155$) and females ($n = 277$) would be inappropriate. Results revealed a nonsignificant *t*-test, $t(430) = -.18$, $p = .858$, Cohen's $d = .02$, suggesting that ICAR-16 Total scores did not differ by gender. Additionally, results revealed a significant MANOVA, $F(4, 427) = 3.02$, $p = .018$, Wilks' $\Lambda = .973$, partial $\hat{\eta}^2 = .03$. Per Tabachnik and Fidell (2013, p. 272), Bonferroni-type adjustment for post hoc analyses of variance (ANOVAs) suggested a critical α of .01. Post hoc ANOVAs revealed that scores on LN trended toward, but did not reach, statistical significance, as females scored relatively higher ($M = 2.44$, $SD = 1.25$) than males ($M = 2.17$, $SD = 1.43$), $F(1, 430) = 4.05$, $p = .045$, partial $\hat{\eta}^2 = .01$. Additionally, scores on R3D also trended toward, but did not achieve statistical significance, as males scored relatively higher ($M = .95$, $SD = 1.24$) than females ($M = .73$, $SD = 1.08$), $F(1, 430) = 3.56$, $p = .06$, partial $\hat{\eta}^2 = .01$. However, scores on VR and MR (similar to the ICAR-16 Total score) did not significantly differ nor trend toward statistical significance between males and females ($ps > .70$). Taken together, these results suggested that participant gender did not appear to meaningfully relate to ICAR-16 Total or subtest scores.

Pearson's correlation coefficients were conducted between ICAR-16 Total scores and demographic data. ICAR-16 Total scores significantly, positively correlated with years of education ($r = .21$, $p < .01$), self-reported estimated annual income ($r = .12$, $p = .01$), and self-estimated IQ ($r = .41$, $p < .01$), but did not significantly correlate with participant age ($r = .06$, $p = .25$). In other words, as education, income, and estimated IQ increased, ICAR-16 Total scores also increased. At the subscale level, participant age did not correlate with VR, LN, and MR scores (*r*s ranged from −.04 to .07, *p*s ranged from .18 to .95); however, participant age *did* significantly correlate with R3D ($r = .14$, $p < .01$). Given the low magnitude of this correlation, it may be a type I error or a consequence of the fairly large sample size.

To determine if the ICAR-16 mean scores in the present sample were representative with normative data reported by the scale's authors, mean values for the ICAR-16 Total and its four subscales were computed according to item-level data provided by Condon and Revelle (2014, p. 55). Given the presence of multiple comparisons, a series of one-sample *t*-tests with Holm-Bonferroni step-down corrections were performed between these values and those of the present study (see Table 2). Results revealed that no ICAR-16 scales significantly differed between the present sample and values derived from Condon and Revelle's (2014) validation study ($ps > .11$) (Table 3).

Discussion

The present study investigated the general cognitive ability of a sample of participants recruited via Amazon Mechanical Turk (mTurk). It was hypothesized that the intelligence scores would not yield observable gender differences, be normally distributed across our obtained sample and commensurate with normative data reported for the measure of interest, and correlated with constructs known to covary with intelligence, thus demonstrating continued support for the external validity and generalizability of data obtained via this platform.

Results of the current study suggested intelligence of American mTurk participants may be broadly

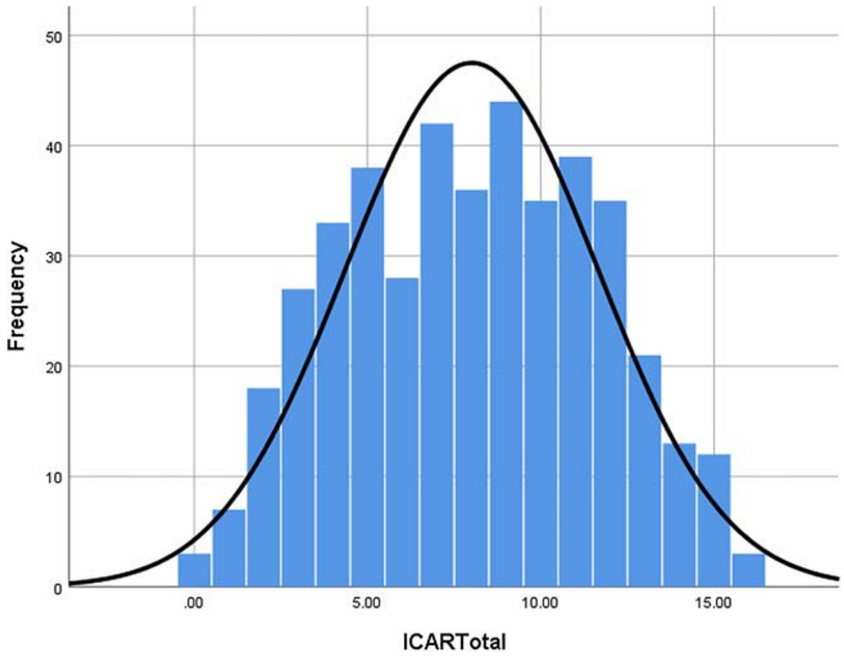

Fig. 1 Visual depiction of ICAR-16 total scores

representative of the population. Regarding the results, mTurk workers' intelligences scores tended to trend toward (though failed to meet statistical significance in support of) patterns of previously reported gender differences between males and females, in that females may perform relatively better on measures of verbal ability while males may perform better on measures of visuospatial ability (e.g., Feingold 1992; Hyde 2014; Wai et al. 2018; Weiss et al. 2003).

However, and more importantly, no overall gender differences were observed in the composite intelligence/IQ score (i.e., ICAR-16 Total), similarly consistent with previous literature (Halpern and LaMay 2000) and widely distributed commercial tests of cognitive ability that do not account for participant gender (e.g., Wechsler Adult Intelligence Scale, Fourth Edition; Wechsler 2008). Thus, it is likely that mTurk workers' general intellectual ability does not meaningfully relate to participant gender, in line with expectations. Additionally, mTurk workers' intelligence scores did not significantly differ from means calculated from available item-level data reported by the chosen measure's authors (Condon and Revelle 2014). As such, these results suggest that mTurk workers' intellectual abilities tend to mimic patterns and distributions observed in other samples and populations.

Finally, mTurk workers' intelligence tended to correlate with other demographic factors, also in line with previous research. For example, prior research has suggested a weak (*r*s between .30 and .32) correlation between IQ calculations and self-estimates (Borkenau and Liebler 1993; Paulhus et al. 1998) similar to what was exhibited within the current study ($r = .41$). Additionally, previously discussed relationships between IQ calculations and both years of education (*r*s between .2 and .3; Furnham and Cheng 2017) and annual income (*r*s between .1 and .2; Furnham and Cheng 2017; Strenze 2007) also resemble scores obtained within the current study (*r*s = .21 and .12 respectively).

Table 3 Item-level means and standard deviations in the present sample and reported by the ICAR-16's authors

ICAR-16 Item	Mean (SD)	Mean (SD) per Condon and Revelle (2014)
LN7	.63 (.48)	.62 (.49)
LN33	.59 (.49)	.59 (.49)
LN34	.68 (.47)	.62 (.48)
LN58	.44 (.50)	.42 (.49)
VR4	.78 (.41)	.67 (.47)
VR16	.67 (.47)	.69 (.46)
VR17	.77 (.42)	.73 (.44)
VR19	.62 (.49)	.61 (.49)
MX45	.51 (.50)	.52 (.50)
MX46	.55 (.50)	.60 (.49)
MX47	.55 (.50)	.62 (.48)
MX55	.39 (.49)	.36 (.48)
R3D3	.15 (.36)	.17 (.37)
R3D4	.19 (.39)	.21 (.41)
R3D6	.29 (.45)	.29 (.46)
R3D8	.18 (.38)	.17 (.37)

SD = Standard Deviation

N = 434

Interestingly, however, the relationship between mTurk workers' intelligence, as measured by ICAR-16 Total score, and age did not exhibit statistical significance. Regarding ICAR-16 subscales, only R3D (which is comprised of items assessing visuospatial and perceptual reasoning) correlated significantly, albeit weakly, with age. Of note, age is commonly associated with cognitive changes, especially declines in fluid/perceptual reasoning (Salthouse 2009). As such, one might expect to see a significant inverse relationship between ICAR-16 total score and age as three of the four ICAR domains (LN, MR, and R3D) arguably were designed to at least partially tap into fluid reasoning capabilities. In contrast, the *positive* correlation between R3D and age suggested that perceptual and fluid reasoning may *increase* with age. However, given the low magnitude of this correlation coefficient and the researchers not correcting for repeated analyses regarding correlation calculations, this finding may be related to range restrictions, be spurious in nature, and/or reflect type I error due to the fairly large sample size. Nonetheless, it is possible that the study design, namely the requirement to have Internet access, competently work a personal computer, and be able to appropriately navigate the mTurk interface, may have created a scenario by which information gathered from older adults may not be generalizable to that subsection of the population. That is, older adults who use mTurk may represent a unique subsample of this demographic and have "self-selected" for greater cognitive ability than may be expected for less computer-literate older adults. Nonetheless, this likely spurious finding aside, broad intellectual abilities of American mTurk workers may be comparable to the population.

Furthermore, additional limitations warrant mentioning. Demographically, there were fewer participants who self-identified as Black or African American and a larger number of those identifying as Asian or Asian American given recent census estimations, suggesting mild generalizability issues within the current sample (U.S. Census Bureau 2010). Additionally, and perhaps most concerning, is that nearly 38% of the initial sample was excluded due to suspect validity or participants electing to start but not complete the survey instrument. This represents a large percentage of participants and this ratio is not known to have been exhibited in previous mTurk based studies. While we feel that our results, given their likeness to ICAR-16 normative data and previously established demographic information, represent the strong validity of the subjects ultimately included within analyses, the size of this group strongly supports the need for multiple and increasingly sophisticated validity checks throughout survey-based instruments when using mTurk as a primary means of data collection. Nonetheless, participants who provided suspect validity or incomplete data were excluded from the final dataset and problems with validity did not meaningfully affect the results described herein. Of note, the original embedded validity items used in the present study are available upon request from the first and second authors.

Overall, results of the present study suggest the appropriateness of the mTurk marketplace and the representativeness of intellectual abilities for mTurk samples in social scientific research. MTurk samples appear to generate demographically rich, appropriately generalizable data across variables, as is described in previous research (e.g., Berinsky et al. 2012; Buhrmester et al. 2011; Harms and DeSimone 2015). This study was the first to suggest that mTurk workers' overall cognitive ability, as measured by the ICAR-16, is likely broadly commensurate with that of the general public through non-significant differences compared to normative data. As such, researchers may confidently assume that the distribution of intellectual and cognitive ability in mTurk workers is likely not a confounding effect in future studies. The authors encourage other researchers to use current results as normative comparison for other work using the ICAR-16 in mTurk samples. Additionally, the present study provides a strong foundation for future researchers who wish to expand upon current findings by correlating intelligence with other constructs of interest, creating and validating additional public domain measures of intelligence, and administering more comprehensive assessments across cognitive domains, especially ones addressing additional aspects of cognitive and neuropsychological functioning (e.g., attention/processing speed, language, and executive functions).

Compliance with Ethical Standards

Conflict of Interest The authors declare they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional research committee (Saint Louis University Institutional Review Board, Approved Protocol #29412) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

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

References

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*(3), 351–368. https://doi.org/10.1093/pan/mpr057.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*(3), 546–553. https://doi.org/10.1037/0022-3514.65.3.546.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980.

Burnham, M. J., Le, Y. K., & Piedmont, R. L. (2018). Who is Mturk? Personal characteristics and sample consistency of these online workers. *Mental Health, Religion & Culture*, 1–11. https://doi.org/10.1080/13674676.2018.1486394.

Casler, K., Bickel, L., & Hackett, E. (2013). Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior, 29*(6), 2156–2160. https://doi.org/10.1016/j.chb.2013.05.009.

Chandler, J., Mueller, P., & Paolacci, G. (2012). Non-naivety among experimental participants on Amazon Mechanical Turk. *Advances in Consumer Research, 40*, 112–116.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64. https://doi.org/10.1016/j.intell.2014.01.004.

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*(3), e57410. https://doi.org/10.1371/journal.pone.0057410.

Eriksson, K., & Simpson, B. (2010). Emotional reactions to losing explain gender differences in entering a risky lottery. *Judgment and Decision making, 5*(3), 159–163.

Feingold, A. (1992). Gender differences in mate selection preferences: A test of the parental investment model. *Psychological Bulletin, 112*(1), 125–139. https://doi.org/10.1037/0033-2909.112.1.125.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. California: Thousand Oaks.

Furnham, A., & Cheng, H. (2017). Socio-demographic indicators, intelligence, and locus of control as predictors of adult financial well-being. *Journal of Intelligence, 5*(2). https://doi.org/10.3390/jintelligence5020011.

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224. https://doi.org/10.1002/bdm.1753.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *The American Psychologist, 59*(2), 93–104. https://doi.org/10.1037/0003-066X.59.2.93.

Halpern, D. F., & Collaer, M. L. (2005). Sex differences in visuospatial abilities: More than meets the eye. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 170–212). New York: Cambridge University Press.

Halpern, D. F., & LaMay, M. L. (2000). The smarter sex: A critical review of sex differences in intelligence. *Educational Psychology Review, 12*(2), 229–246. https://doi.org/10.1023/A:1009027516424.

Harms, P. D., & DeSimone, J. A. (2015). Caution! MTurk workers ahead—Fines doubled. *Industrial and Organizational Psychology, 8*(2), 183–190. https://doi.org/10.1017/iop.2015.23.

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Most people are not WEIRD. *Nature, 466*(7302), 29–29. https://doi.org/10.1038/466029a.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399–425.

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118.

Hyde, J. S. (2014). Gender similarities and differences. *Annual Review of Psychology, 65*(1), 373–398. https://doi.org/10.1146/annurev-psych-010213-115057.

Ipeirotis, P. (2010). *Demographics of Mechanical Turk*. New York University Working Paper No. CEDER-10-01.

Johnson, D. R., & Borden, L. A. (2012). Participants at your fingertips: Using Amazon's Mechanical Turk to increase student-faculty collaborative research. *Teaching of Psychology, 39*(4), 245–251. https://doi.org/10.1177/0098628312456615.

Kees, J., Berry, C., Burton, S., & Sheehan, K. (2017). An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising, 46*(1), 141–155. https://doi.org/10.1080/00913367.2016.1269304.

Keith, M. G., & Harms, P. D. (2016). Is Mechanical Turk the answer to our sampling woes? *Industrial and Organizational Psychology, 9*(1), 162–167. https://doi.org/10.1017/iop.2015.130.

Kim, H.-Y. (2013). Statistical notes for clinical researchers: Assessing normal distribution (2) using skewness and kurtosis. *Restorative dentistry & endodontics, 38*(1), 52–54. https://doi.org/10.5395/rde.2013.38.1.52.

Kirkegaard, E. O. W., & Nordbjerg, O. (2015). Validating a Danish translation of the international cognitive ability resource sample test and cognitive reflection test in a student sample. *Open Differential Psychology*. https://doi.org/10.26775/odp.2015.07.31.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings SIGCHI Conference on Human Factors in Computing Cystems*, 453–456. https://doi.org/10.1145/1357054.1357127.

Landers, R. N., & Behrend, T. S. (2015). An inconvenient truth: Arbitrary distinctions between organizational, Mechanical Turk, and other convenience samples. *Industrial and Organizational Psychology, 8*(2), 142–164. https://doi.org/10.1017/iop.2015.13.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). *Using the Amazon Mechanical Turk for transcription of spoken language*. In 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, 5270–5273. https://doi.org/10.1109/ICASSP.2010.5494979.

Mason, W., & Suri, S. (2011). A guide to conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6.

Mason, W., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". *Proceedings of the ACM SIGKDD Workshop on Human Computation*, 77–85. https://doi.org/10.1145/1600150.1600175.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867–872.

Oztuna, D., Elhan, A. H., & Tuccar, E. (2006). Investigation of four different normality tests in terms of type 1 error rate and power under different distributions. *Turkish Journal of Medical Sciences, 36*(3), 171–176.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision making, 5*(5), 411–419.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science, 23*(3), 184–188.

Paulhus, D. L., Lysy, D. C., & Yik, M. S. M. (1998). Self-report measures of intelligence: Are they useful as proxy IQ tests? *Journal of*

Personality, 66(4), 525–554. https://doi.org/10.1111/1467-6494.00023.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y.

Reilly, D., Neumann, D. L., & Andrews, G. (2017). Gender differences in spatial ability: Implications for STEM education and approaches to reducing the gender gap for parents and educators. In *Visual-spatial ability in STEM education* (pp. 195–224). Cham: Springer.

Ren, X., Wang, T., Sun, S., Deng, M., & Schweizer, K. (2018). Speeded testing in the assessment of intelligence gives rise to a speed factor. *Intelligence, 66*, 64–71. https://doi.org/10.1016/j.intell.2017.11.004.

Ross, J., Zaldivar, A., Irani, L., & Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. Department of Informatics, University of California, Irvine, USA, Tech. Rep.

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging, 30*(4), 507–514. https://doi.org/10.1016/j.neurobiolaging.2008.09.023.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338.

Shevlin, M., & Miles, J. N. V. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences, 25*(1), 85–90. https://doi.org/10.1016/S0191-8869(98)00055-5.

Shipley, W., Gruber, C., Martin, T., & Klein, M. (2009). *Shipley institute of living scale-2*. Los Angeles: Western Psychological Services.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*(1), 155–167. https://doi.org/10.3758/s13428-010-0039-7.

Sternberg, R. J. (Ed.). (2019). *Human intelligence: An introduction*. Cambridge: New York.

Stoutenbourgh, J. W. (2008). Demographic measures. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (Vol. 1, pp. 185–186). Thousand Oaks: Sage.

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35*(5), 401–426. https://doi.org/10.1016/j.intell.2006.09.004.

Suri, S., & Watts, D. J. (2011). Cooperation and contagion in web-based, networked public goods experiments. *PLoS One, 6*(3), e16836. https://doi.org/10.1371/journal.pone.0016836.

Tabachnik, B. G., & Fidell, L. (2013). *Using multivariate statistics*. Boston: Allyn & Bacon.

U. S. Census Bureau. (2010). *Decennial census of population and housing*. Retrieved from https://www.census.gov/programs-surveys/decennial-census/decade.2010.html. Accessed 3/31/2020

Wai, J., Hodges, J., & Makel, M. C. (2018). Sex differences in ability tilt in the right tail of cognitive abilities: A 35-year examination. *Intelligence, 67*, 76–83. https://doi.org/10.1016/j.intell.2018.02.003.

Wechsler, D. (2008). *Wechsler adult intelligence scale–fourth edition (WAIS–IV)*. San Antonio: The Psychological Corporation.

Weiss, E. M., Kemmler, G., Deisenhammer, E. A., Fleischhacker, W. W., & Delazer, M. (2003). Sex differences in cognitive functions. *Personality and Individual Differences, 35*(4), 863–875. https://doi.org/10.1016/S0191-8869(02)00288-X.

Young, S. R., Keith, T. Z., & Bond, M. A. (2019). Age and sex invariance of the International Cognitive Ability Resource (ICAR). *Intelligence, 77*, 101399. https://doi.org/10.1016/j.intell.2019.101399.