Journal of Advertising

ISSN: 0091-3367 (Print) 1557-7805 (Online) Journal homepage: www.tandfonline.com/journals/ujoa20

Crowdsourcing the Implicit Association Test: Limitations and Best Practices

Scott Connors, Katie Spangenberg, Andrew W. Perkins & Mark Forehand

To cite this article: Scott Connors, Katie Spangenberg, Andrew W. Perkins & Mark Forehand (2020) Crowdsourcing the Implicit Association Test: Limitations and Best Practices, Journal of Advertising, 49:4, 495-503, DOI: 10.1080/00913367.2020.1806155

To link to this article: https://doi.org/10.1080/00913367.2020.1806155

Published online: 28 Aug 2020.

JOURNAL OF ADVERTISING
2020, VOL. 49, NO. 4, 495–503
https://doi.org/10.1080/00913367.2020.1806155

Routledge
Taylor & Francis Group

RESEARCH NOTE

Crowdsourcing the Implicit Association Test: Limitations and Best Practices

Scott Connors[a], Katie Spangenberg[b], Andrew W. Perkins[c], and Mark Forehand[b]

[a]Western University, London, Ontario, Canada; [b]University of Washington, Seattle, Washington, USA; [c]Washington State University, Pullman, Washington, USA

ABSTRACT

Although the use of crowdsourced online panels for behavioral data collection is commonplace in media and advertising research, only recently have software advancements made it possible for researchers to easily collect implicit measures online. Motivated by the recent decline in MTurk data quality and a dearth of literature examining the use of Implicit Association Tests with crowdsourced samples, we investigate cross-sectional data from eight IAT studies conducted using various samples (Mturk, online undergraduate students, and undergraduate behavioral labs). We document relative rates of participant inattention, non-naivety, and lack of motivation between crowdsourced and traditional samples and demonstrate the ramifications of these threats to the reliability and validity of IAT results. Finally, we build on these insights to outline best practices for crowdsourcing implicit measures in advertising and media research.

Since its inception, the Implicit Association Test (IAT; Greenwald, McGhee, and Schwartz 1998) has been widely used by advertising researchers to examine constructs and relationships not accurately captured through self-report or introspection, such as self-brand associations (Perkins and Forehand 2012) or ad-based affect transfer (Gibson 2008). Although latency-based implicit measures are traditionally conducted in a controlled laboratory setting, recent technological advancements have enabled online data collection, shifting a large number of IAT studies from the lab to online crowdsourced samples (e.g., MTurk, Prolific). For example, the *Inquisit* platform (*Millisecond.com*) provides researchers editable scripts for conducting implicit measures online and IATGEN (Carpenter et al. 2019) allows researchers to easily conduct IATs directly within the Qualtrics online survey platform, making them more suitable for crowdsourced panels. However, despite these methodological developments, little research has examined the efficacy of administering the IAT in an online crowdsourced environment, with the exception of one broad scale replication project, which included two Implicit Association Test effects amongst its battery of psychological tests (Klein et al. 2014).

Given the numerous applications implicit measures hold for advertising research, it is necessary to consider the limitations of conducting them using crowdsourced samples. We suggest there are unique properties of crowdsourced samples that are highly detrimental to the quality of IAT data if not taken into consideration. Motivated by research examining the characteristics of MTurk participants (e.g., Goodman, Cryder, and Cheema 2013; Hauser, Paolacci, and Chandler 2019) and a recent decline in MTurk data quality (Chmielewski and Kucker 2020), the current research reconciles cross-sectional IAT data from eight studies with existing research on the characteristics of crowdsourced samples. We identify three specific participant characteristics—participant motivation, inattention, and non-naivety—and document the unique ramifications each poses for IAT data quality sourced from MTurk, online undergraduate students, and undergraduate behavioral lab populations. Finally, we build on these insights to outline a series of best practices for collecting, cleaning, and analyzing implicit measures in an online setting.

CONTACT Scott Connors sconnor4@uwo.ca Western University, DAN Management, 1151 Richmond Street, Social Science Center, London, ON N6A 5C2, Canada.
Scott Connors (Ph.D., Washington State University) is an Assistant Professor, DAN Department of Management and Organizational Studies, Western University.
Katie Spangenberg (Ph.D. Candidate, University of Washington) is a Ph.D. Marketing Candidate, Foster School of Business, University of Washington.
Andrew W. Perkins (Ph.D., University of Washington) is an Associate Professor, Carson College of Business, Washington State University.
Mark Forehand (Ph.D., Stanford University) is a Professor of Marketing, Foster School of Business, University of Washington.

The Prevalence of the IAT in Advertising and Consumer Research

The IAT is a latency-based categorization task designed to assess the relative strength of association between two pairs of concepts in memory. The task operates under the assumption that behavioral responses to target stimuli objects (i.e., pressing the appropriate key on a keyboard) should be easier (and therefore faster) when the underlying associations between the objects are more strongly held in memory. Thus, faster response latencies represent stronger associations between concepts. The IAT can effectively predict attitudes and behavior when consumers may be unaware of or unable to accurately report their "true" feelings or thoughts, or when self-presentation or impression management desires motivate consumers to misrepresent their own attitudes or beliefs when responding to traditional explicit (self-report) measures (Greenwald, McGhee, and Schwartz 1998).

Consumer researchers have used the IAT to examine associative relationships and processes related to the formation, change, and consequences of consumer attitudes, the self-concept, decision making, and behavior. Most common is the attitude IAT, which captures the relative association of positive and negative attributes with two target categories (e.g., *Coke* and *Pepsi*). Relatedly, self-brand associations are captured using a self-attribute IAT wherein positive and negative terms are replaced with *Self* and *Other*.

Within the advertising literature, the IAT has been used in myriad ways. For example, a study examining the cognitive and attitudinal effects of ads featuring same sex couples found that implicit attitudes affected processing and evaluation in ways that explicit measures were unable to account for (Read, van Driel, and Potter 2018). In another study, the ethnicity of the model in an advertisement was varied, and researchers found a dissociation between implicit and explicit attitudes toward the advertisements (Brunel, Tietje, and Greenwald 2004). Perkins and Forehand (2012) presented participants the names of fictitious brands as interstitial social media advertisements to demonstrate that consumers automatically self-associate with advertised brands and that self-association predicted choice behavior. Additional research has used the IAT to examine advertising content in video games (Waiguny, Nelson, and Marko 2013), ad-based evaluative conditioning (Gibson 2008), susceptibility to false advertising (LaTour and LaTour 2009), political advertising (Arendt, Marquart, and Matthes 2015), e-cigarette advertising (Pokhrel et al. 2016), and the effects of ad exposure on experiential memory (Braun-LaTour et al. 2004). Relatedly, the IAT is premised on accessibility, a concept which has been studied widely in media contexts. For example, research has shown that media figures influence self-body perceptions when the figures are more accessible (Chandler, Konrath, and Schwarz 2009) and that self-concept accessibility predicts behavioral responses to anti-drug advertisement exposure (Comello 2013).

Conducting a Valid and Reliable Implicit Association Test

The IAT effect is captured by the *D*-score, the most common method of scoring the IAT (Greenwald, Nosek, and Banaji 2003). This scoring algorithm divides the mean latency difference (target-congruent trials minus target-incongruent trials) for each participant by the pooled standard deviation of both the target-congruent and target-incongruent trials. The *D*-score shows higher internal consistency and greater resistance to order effects or individual response differences than conventional scoring methods (Greenwald, Nosek, and Banaji 2003).

As with all latency measures, the resulting data must be properly cleaned to ensure reliable and valid responses (Fazio 1990). Although it is important to remove subjects who are disinterested, unengaged, and not sufficiently motivated (i.e., paying no attention to accuracy), standard data cleaning practice is to remove as few research participants as possible. Traditionally, participant removal is based on the observation of excessively fast response latencies. Specifically, participants who respond in less than 300 ms to more than 10% of the IAT trials are removed (10%/300 ms; Greenwald, Nosek, and Banaji 2003). Excessive speed indicates a participant is likely responding to the stimuli quickly, yet inattentively. Research has shown that removing a small number of participants (8.9%) based on excessively fast latencies increases data reliability (Greenwald, Nosek, and Banaji 2003). The 300 ms cutoff represents the lower bound at which humans can comprehend stimuli and therefore has been commonly deemed best practice in response latency research (Fazio 1990; Greenwald, Nosek, and Banaji 2003; Nosek et al. 2014).

In keeping with best data practices to not frivolously remove data, error trials are retained, as IAT results do not significantly improve when error trials are removed (Nosek et al. 2014). However, excessively high error rates often accompany data with fast latencies. Extant research has shown that participants who displayed excessively fast response times for greater

than 10% of trials also exhibited an error rate of 35.7% (as compared to 8.7% for the rest of the participants; Greenwald, Nosek, and Banaji 2003). A high error rate suggests that a participant ignored the instructions or made no attempt to classify the data correctly.

Finally, participant attentiveness is critical to ensure that response latency times are reliable (Luce 1986). It is important that participants read and understand the detailed instructions of the IAT to complete it correctly. Further, the repeated trials require focused attention for the duration of the IAT (approximately five minutes). As a result, traditional practice is to reduce distractions. Specifically, IAT data is cleaner and more reliable when administered in a behavioral lab where distractions are limited and external stimuli can be controlled by the experimenter. Furthermore, it is beneficial to ensure that participants will be unfamiliar with the test. Research suggests that participants can adapt to IATs after having taken the test multiple times (Greenwald, Nosek, and Banaji 2003), leading to decreased response extremity and a reduction in the validity of IAT scores for repeat participants (Nosek, Banaji, and Greenwald 2002).

Issues Crowdsourcing the Implicit Association Test

Over the past decade, the use of crowdsourced online panels for behavioral data collection has become commonplace within advertising research. We conducted an examination of articles in the Journal of Advertising from the past three years (2017–2020) and found that 50% of all studies were crowdsourced, and that MTurk alone was used in 34% of all studies – a finding on par with that of consumer research (Goodman and Paolacci 2017). The reliance on MTurk subjects stems from the unparalleled efficiency that the platform affords. Crowdsourcing is simple to conduct, offers access to a diverse sample, costs significantly less than other alternatives, can be conducted in a timely manner, and offers substantial flexibility (Hauser, Paolacci, and Chandler 2019).

In terms of data quality, MTurk samples have shown mixed results. On the positive side, research has observed that MTurk data performs comparably to professional panels and student samples in reliability (Behrend et al. 2011; Buhrmester, Kwang, and Gosling 2011), replicability (Casler, Bickel, and Hackett 2013), response to general attention checks (Hauser and Schwarz 2016) and overall data quality (Kees et al. 2017). However, other research has questioned MTurk data on these dimensions, finding evidence for reduced performance on reliability (Rouse 2015), replicability (DeVoe and House 2016; Casey et al. 2017) and overall quality (Ford 2017). Some of these inconsistencies may be temporal, as a recent four-year natural experiment observed substantial decreases in MTurk data reliability, validity and replicability beginning in the summer of 2018 (Chmielewski and Kucker 2020). To combat this issue, MTurk offers access to "Master" workers who maintain a set standard for quality, consistency, and variety of task completion.

Given the reported recent declines in crowdsourced data quality and the unique requirements for IAT data collection, it is essential that researchers monitor the appropriateness of crowdsourcing in this domain. Although research has demonstrated crowdsourced replicability of response latency-based effects in cognitive psychology (Zwaan et al. 2018) and behavioral research (Crump, McDonnell, and Gureckis 2013), data examining IAT reliability remains scarce. One large scale replication project included two Implicit Association Test effects amongst its battery of psychological tests (Klein et al. 2014). However, the results indicated notable differences in effect sizes between samples and only one of the nine "online" datasets was a crowdsourced panel (Mturk). To assess the implications of crowdsourcing on the IAT, we juxtapose data from eight IAT studies—conducted across four sample types—with a thorough review of extant research on the characteristics of crowdsourced samples. To our knowledge, this is the first broad analysis of the effects of crowdsourced data collection on IAT responses.

Methodology

We conducted eight studies over a period of 18 months (January 2019 to June 2020) that included the same IAT ($N = 2378$, 44% female, $M_{age} = 30.1$ years). Four were collected on MTurk (one using Master workers), two were collected online using undergraduate samples, and two were collected using an undergraduate behavioral lab setting. All data were collected in Qualtrics using IATGEN (Carpenter et al. 2019) and were restricted from being completed on mobile devices. Participants began each study by completing a self-attribute Implicit Association Test which involves sorting stimuli into categories that relate to the self (i.e., *Self* vs. *Other*). For a full overview of IAT methodology, see Greenwald, McGhee, and Schwartz (1998). Participants in each study also completed the same general demographics

Table 1. Cross-sectional data from the implicit association test across MTurk, online, and in-lab samples.

Data Source	1	2	3	4	5	6	7	8
Sample	MTurk	MTurk	MTurk	Online undergraduate	Online undergraduate	Behavioral Lab	Behavioral Lab	MTurk Masters
Full Dataset								
N_{FULL}	601	409	333	285	239	97	314	100
Age	38.6	37.4	37.7	19.2	18.5	19.1	18.9	43.5
Gender (% female)	39.3	38.7	36.8	44.1	46.4	53.2	59.5	44.0
Avg. Latency (ms)	668	727	768	850	881	873	893	829
D-Score	0.23	0.24	0.31	0.42	0.42	0.41	0.47	0.34
D-Score SD	0.44	0.40	0.44	0.48	0.48	0.41	0.40	0.33
Error Rate	0.23	0.21	0.17	0.13	0.12	0.10	0.09	0.10
Reliability	0.86	0.80	0.86	0.89	0.89	0.80	0.86	0.75
Captcha included[2]	No	Yes	Yes	No	No	No	No	No
Attention Check[3]	82.5	87.2	91.7	79.7	85.2	76.6	91.5	95.5
Burners								
$N_{BURNERS}$[1]	223	128	78	26	20	2	8	5
$\%_{BURNERS}$	37%	31%	23%	9%	8%	2%	3%	5%
Avg. Latency (ms)	256	239	315	310	497	258	437	312
Error Rate	0.50	0.48	0.47	0.47	0.33	0.48	0.39	0.51
Clean Dataset (Non-Burners)								
N_{CLEAN}	378	281	255	259	219	95	306	95
Avg. Latency (ms)	911	949	906	905	916	886	905	857
D-Score	0.38	0.36	0.41	0.48	0.46	0.43	0.48	0.34
D-Score SD	0.38	0.40	0.36	0.39	0.38	0.41	0.38	0.32
Error Rate	0.08	0.09	0.08	0.09	0.10	0.09	0.08	0.08
Reliability	0.81	0.78	0.72	0.87	0.85	0.78	0.85	0.74

[1]Reflects total number of participants with greater than 10% of latencies $<$ 300 ms.
[2]Used to screen out non-human bots.
[3]Percentage of sample correctly responding to the question "Reliability is important, please select option two" embedded within a series of 7-pt Likert scale questions.

questionnaire and attention check measure. Two studies (Samples 1 and 2) began with a *Captcha* question to prevent bots from completing the study. For each study, the *D* measure scoring algorithm was used to rescale the IAT effects (Greenwald, Nosek, and Banaji 2003). In Table 1, we report *D*-scores, sample statistics, and common indicators of IAT data quality: 1) reliability; 2) number of participants with greater than 10% of trials with an average latency of less than 300 ms; and 3) mean error rates.

Results and Discussion

Many MTurkers are Speeding Through Implicit Tasks

Although some MTurkers may complete studies due to an intrinsic motivation to be a part of the research process, most are driven by financial incentives (Smith et al. 2016). Further, MTurkers are incentivized to prioritize speed over attentiveness and maximize earnings by completing as many studies as possible, as quickly as possible. As Ford notes, "MTurk respondents do not make very much, which further pressures participants to take as many surveys as they can to make sufficient money for their efforts" (2017, p. 156). As a result, research shows that compared to traditional panels, MTurkers spend less time reading questions and complete surveys faster (Smith et al. 2016; Kees et al. 2017). This inattentive "burning" through response items has been associated with inflated inter-item correlations, and thus higher Type I error rates (Wood et al. 2017). Herein we use the term "burner" to refer to respondents who speed through studies as quickly as possible.

Our data suggest that IAT burners are much more prevalent on MTurk compared to other sample populations. Burning through an IAT occurs when the participant indiscriminately presses the two response keys as quickly as possible for the duration of the trials with no regard for the stimuli presented. Burners can be identified within the data by observing the speed of participants' response latencies on each trial. To identify burners, we examined the IAT latency data across the eight studies and identified the proportion of respondents who exhibited response latencies below 300 ms for greater than 10% of trials (Greenwald, Nosek, and Banaji 2003). We found that this proportion was higher on standard MTurk (23–37% of participants) compared to MTurk Masters (5%), our online student panel (8–9%), or in-lab undergraduate participants (2–3%). Furthermore, the proportion of burner participants on MTurk well exceeds the levels found in traditional IAT research (8.9%; Greenwald, Nosek, and Banaji 2003). Thus, excluding IAT burner participants from standard MTurk studies results in removal rates substantially higher than those in our other sample populations and, notably, are well above the generally accepted maximum removal rate of 15% in behavioral research (Chandler, Mueller, and

Paolacci 2014). Although these high rates of removal are warranted by the underlying response patterns, this may cause researchers difficulty in the review process as reviewers are often apprehensive toward studies that exceed accepted data removal rate norms.

Next, to further assess response speed we examined the effect of sample type (MTurk, Masters, online undergraduate, in-lab undergraduate) on average response latencies. Results from a one-way ANOVA indicated a significant effect of sample type, ($F(3, 2374) = 44.40$, $p < .001$). Planned contrasts confirmed that mean response latencies were faster for standard MTurk participants ($M = 711$ ms) as compared to Masters ($M = 829$ ms, $p < .001$) online undergraduate ($M = 864$ ms, $p < .001$) and in-lab undergraduate participants ($M = 888$ ms, $p < .001$), while response latencies did not differ between Masters, online, and in-lab undergraduates (all $ps > .11$). Although these findings imply that IAT data are strongly affected by burner subjects on MTurk, it is possible that the differences in response latencies are due to idiosyncratic differences between the sample populations. To investigate this further, we classified respondents as either "burners" or "non-burners" based on the 10%/300 ms criteria and collapsed across all samples based upon this distinction. A one-way ANOVA indicated a significant effect of burners on mean response latency, ($F(1, 2376) = 2828.11$, $p < .001$), such that response latencies were significantly faster for burner participants ($M = 277$ ms) than for non-burner participants ($M = 911$ ms). It is important to note that not only did burner participants respond significantly faster, their *average* response latency ($M = 277$ ms) was significantly below the 300 ms threshold ($t(489) = -2.14$, $p < .05$) that is commonly agreed upon as the minimum time that a human needs to recognize stimuli (Fazio 1990; Greenwald, Nosek, and Banaji 2003; Nosek et al. 2014).

Finally, we examined the role of burner subjects in suppressing the IAT effect. Results from a one-way ANOVA indicated a significant effect of sample type on D scores across all sample types, ($F(3, 2374) = 32.04$, $p < .001$). Planned contrasts confirmed that D scores were smaller for standard MTurk participants ($M = .26$) as compared to Masters ($M = .34$, $p = .07$), online undergraduate ($M = .42$, $p < .001$) and in-lab undergraduate participants ($M = .46$, $p < .001$) while D scores did not differ between online and in-lab undergraduates ($p = .20$). To rule out that this suppression is due to idiosyncratic differences between the samples, we examined the effect of burners on D scores. Results indicated a significant effect, ($F(1, 2376) = 509.36$, $p < .001$), such that D scores were substantially smaller for burner participants ($M = -.03$) than for non-burner participants ($M = .43$). It is important to note that the mean D score for burners—which reflects the average latency difference between congruent and incongruent trials—did not differ from zero, ($t(489) = 1.71$, $p = .09$), which indicates burners' tendencies to indiscriminately speed through all trials.

Many MTurkers Are Not Paying Attention to Implicit Tasks

The high level of inattention displayed by MTurkers is well-documented and common in most crowdsourced samples. Whereas a laboratory setting eliminates or controls external stimuli and distractions, MTurk participants are not observed during the study. As a result, MTurk participants are less likely than participants from student or community samples to attend to experimental stimuli and often look to the Internet to find sample answers to survey questions (Goodman, Cryder, and Cheema 2013). Self-report evidence suggests that MTurkers often complete surveys while simultaneously engaged in other activities such as watching TV, listening to music, or messaging (Chandler, Mueller, and Paolacci 2014). In experimental and survey research, low-effort or inattentive responding can increase random error, systematic error, or spurious correlations, inflate internal reliability, and reduce discriminant validity of independent measures (Hauser, Paolacci, and Chandler 2019).

This tendency toward distracted or inattentive responding is particularly troubling for implicit tasks such as the IAT for a number of reasons. First, the IAT requires understanding detailed instructions to inform participants of the task procedure. Inattentive participants are unlikely to sufficiently attend to this information, leaving them inadequately informed when the task begins. Second, the IAT requires focus and concentration throughout a lengthy course of trials. It relies on repeated split-second, reactionary responses in which the participant must quickly identify the stimuli and place it in one of two sets of categories. It is unlikely that inattentive respondents will remain focused throughout the duration of the task or be able to effectively discern between stimuli. Finally, IAT stimuli appear on the screen only briefly, meaning participants distracted by external stimuli may be unaware of the stimuli they were shown on a given trial.

Although we do not make a distinction regarding whether MTurker inattention stems from distraction

or a desire to burn through a particular study, we address the impact of inattention by first examining error rates between sample types. Recall that each IAT trial requires correctly categorizing stimuli into one of two categories. In each trial, there is a correct response. Error rates reflect the total number of incorrect trials as a function of the total number of target trials. Results from a one-way ANOVA indicated a significant effect of sample type on error rate across all participants, ($F(3, 2374) = 78.28$, $p < .001$). Planned contrasts show that error rates were higher on standard MTurk ($M = .21$) as compared to Masters ($M = .10$, $p < .001$) online undergraduates ($M = .12$, $p < .001$), and in-lab undergraduate participants ($M = .09$, $p < .001$).

Next, given a prevalence of burners on MTurk, we examined the impact of burner participants to better understand the error rate disparity between the sample types. Results from a one-way ANOVA indicated a significant effect of burners on error rates, ($F(1, 2376) = 11859.20$, $p < .001$), such that error rates were higher for burner participants ($M = .47$) than for non-burner participants ($M = .09$). Importantly, not only were mean error rates substantially higher for burner participants, but they approached 50%, which is the error rate that would be expected due to chance (such that when responding to the IAT, participants place the item randomly into one of two categories each trial). This finding provides further evidence that the concerningly fast response latencies displayed by burners cannot be attributed to any idiosyncratic IAT skill that MTurk participants may possess. Burner respondents are not quickly and accurately completing the IAT task but are instead randomly pressing response keys quickly and indiscriminately. The observed increased error rates vividly capture this careless responding.

MTurkers indiscriminately speeding through the IAT undermines the validity of the instrument, but it also arbitrarily inflates the reliability of the instrument. To examine this inflated reliability, for each sample we calculated split-half reliability estimates with a Spearman-Brown correction. Our findings indicated that on average our standard MTurk samples showed a 9% increase in split-half reliability estimates with burner participants included in the sample as compared to a sample with burners removed. As burner participants respond equally quickly to all trials, the uniformity of their responses falsely increases the tendency of these trials to "move together." As a result, the prevalence of burner participants on MTurk arbitrarily inflates the split-half reliability estimates for the instrument. Importantly, we find no such increases in reliability for the Masters, undergraduate online, or undergraduate lab samples, reflecting the relative absence of burners in these samples.

Many MTurkers Are Already Familiar with Implicit Tasks

Although crowdsourcing ostensibly offers researchers near instantaneous access to a diverse sampling population, the actual MTurk population is much smaller than most researchers realize (Stewart et al. 2015). Many professional MTurkers have engaged with the platform for years and account for a substantial portion of total tasks completed. In a comprehensive cross-section of MTurkers, Chandler, Mueller, and Paolacci (2014) found that 10% of MTurkers were responsible for 41% of all tasks completed on the platform. Furthermore, while MTurk Masters exhibited signs of improved data quality over standard MTurkers, maintaining their Master designation requires consistent high levels of activity on the platform, exacerbating issues of non-naivety. Most MTurkers have seen common manipulations and measurement tools many times – a problem compounded by the fact that the MTurk pool is shared by researchers worldwide (Stewart et al. 2015). Although non-naivety can likewise be a problem for traditional student samples, undergraduate subject pools experience more frequent turnover (Chandler et al. 2015) and do not complete anywhere near the quantity of studies as MTurkers. Student participants complete approximately three to six studies per academic semester, and even traditional survey panelists complete substantially fewer surveys on average compared to MTurk participants (3.2 versus 10 to 17 surveys per week for MTurk participants; Chandler et al. 2015). Although screening procedures can be employed to address non-naivety, many MTurk participants are dishonest when asked about whether they have completed a similar study before (Chandler and Paolacci 2017) and do so in order to participate in multiple related experiments (Chandler, Mueller, and Paolacci 2014).

Mturker non-naivety has numerous identifiable implications for IAT data quality. First, Mturkers' repeated participation in traditional experiments has been shown to substantially diminish effect sizes (e.g., Chandler et al. 2015). On the IAT, this further supresses D scores and sample variability, reducing the predictive validity of the IAT as participant scores are less extreme for participants who have taken more than one IAT (Nosek, Banaji, and Greenwald 2002). Indeed, we also find evidence of suppression for

Masters, as their *D* scores were smaller than in-lab ($p < .05$) and online undergraduates ($p = .09$), suggesting this pattern persists across both subsections of the MTurk population. Second, MTurkers' repeated exposure to the same methodologies can lead to practice effects (Holden, Dennie, and Hicks 2013), such that completing an IAT on any topic (not necessarily an identical one) may influence future IAT performance. Furthermore, participants who practice the IAT can actually "fake" an IAT score (Röhner, Schröder-Abé, and Schütz 2011). Finally, additional research has shown that well-established effects may not replicate with more experienced MTurkers (e.g., DeVoe and House 2016).

General Discussion

Given the many applications of the IAT for advertising research and the growing prevalence of crowdsourced advertising studies, recent advancements that bring the IAT online are a boon for researchers. However, such research must be conducted with caution. Hauser, Paolacci, and Chandler (2019) advise researchers to adopt a cautious approach to using MTurk participants, noting that while it is possible to collect quality data on MTurk, it is far from guaranteed. We take this recommendation one step further and propose that quality IAT data collection from panels such as MTurk requires that researchers directly address problematic respondents. Due to the unique characteristics and financial incentives of MTurk, data will likely include a substantially higher proportion of burner participants resulting in faster average response latencies, supressed IAT effects, heightened error rates, and inflated reliability estimates compared to traditional samples. To alleviate the issues inherent in crowdsourcing implicit measures, we offer recommendations for researchers conducting and reviewers assessing IAT-based research conducted using MTurk samples.

First, we recommend that researchers remove all respondents identified as burners (i.e., those with greater than 10% of trials faster than 300 ms) from their datasets even if the resulting participant removal rate exceeds the 8.9% rate reported in foundational IAT research (Greenwald, Nosek, and Banaji 2003). Although, this tradeoff has traditionally been viewed as a substantial increase in data quality in exchange for the removal of a trivial number of participants, our findings demonstrate that for crowdsourced IATs this is no longer a clear decision as the proportion of participants exceeding the 10%/300 ms criteria is no longer trivial. Although there is concern regarding the extent to which researchers already overclean crowdsourced data (e.g., Babin, Griffin, and Hair 2016), our results suggest removing these participants is warranted. Although removing participants with response latencies below 300 ms in at least 10% of trials is the benchmark, our data suggests that MTurk participant performance is far worse than this criterion. Specifically, MTurk burners' *average* trial latencies (across *all* trials) are below the 300 ms cutoff ($M = 277$ ms). Further, these abnormally fast average latencies resulted in an average error rate (47%) that approaches chance. Supporting our recommendation that all burner subjects be removed, we observe comparable average latencies, *D* scores, and error rates across all samples when burners subjects are removed (see Table 1).

Second, if researchers wish to continue using crowdsourced samples for IAT-based projects, both the researchers and reviewers alike should be comfortable with removing a much higher percentage of participants ($\sim$ 20% to 40%) than is traditional in behavioral research (Chandler, Mueller, and Paolacci 2014). It is here we put out a call to reviewers, at least in the specific instance of IAT-based research, to relax the traditional acceptance of the 15% participant removal rule (Chandler, Mueller, and Paolacci 2014). To help alleviate these reviewer concerns, researchers should provide documentation of their data cleaning procedures (Goodman, Cryder, and Cheema 2013).

Finally, although addressing burner participants in IAT-based research necessitates the removal of a substantial portion of an MTurk sample, researchers can employ strategies to reduce removals and mitigate some of the wasted resources that result. For example, including a *Captcha* question, in which participants have to answer a question using visual stimuli, reduces concerns that a portion of burner participants may be attributable to non-human bots (Chmielewski and Kucker 2020). Indeed, the data from our MTurk sample suggests that including a *Captcha* question at the beginning of the study decreased burner rates from 37% of the sample to 23–31%. Additionally, opting for Master workers over a standard MTurk sample seems to reflect a marked difference in data quality across the IAT metrics examined. Finally, while our findings highlight important concerns in conducting IATs using MTurk samples, future research is needed to explore the prevalence of these issues across other crowdsourced panels (e.g., *Qualtrics, Dynata, Prolific*).

With these processes and safeguards in place, we believe that IAT-based research can be conducted

effectively with crowdsourced samples. However, diligent monitoring of response patterns and reviewer acceptance of pre-established participant removal metrics is paramount.

ORCID

Scott Connors http://orcid.org/0000-0002-9076-9221
Katie Spangenberg http://orcid.org/0000-0003-3443-9225
Andrew W. Perkins http://orcid.org/0000-0003-3020-6491
Mark Forehand http://orcid.org/0000-0002-1973-6400

References

Arendt, F., F. Marquart, and J. Matthes. 2015. Effects of right-wing populist political advertising on implicit and explicit stereotypes. *Journal of Media Psychology* 27 (4): 178–89. doi:10.1027/1864-1105/a000139

Babin, B. J., M. Griffin, and J. F. Hair. 2016. Heresies and sacred cows in scholarly marketing publications. *Journal of Business Research* 69 (8):3133–38. doi:10.1016/j.jbusres.2015.12.001

Behrend, T. S., D. J. Sharek, A. W. Meade, and E. N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behavior Research Methods* 43 (3):800–13. doi:10.3758/s13428-011-0081-0

Braun-LaTour, K. A., M. S. LaTour, J. E. Pickrell, and E. F. Loftus. 2004. How and when advertising can influence memory for consumer experience. *Journal of Advertising* 33 (4):7–25. doi:10.1080/00913367.2004.10639171

Brunel, F. F., B. C. Tietje, and A. G. Greenwald. 2004. Is the implicit association test a valid and valuable measure of implicit consumer social cognition? *Journal of Consumer Psychology* 14 (4):385–404. doi:10.1207/s15327663jcp1404_8

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

Carpenter, T. P., R. Pogacar, C. Pullig, M. Kouril, S. Aguilar, J. LaBouff, N. Isenberg, and A. Chakroff. 2019. Survey-software implicit association tests: A methodological and empirical analysis. *Behavior Research Methods* 51 (5): 2194–2208. doi:10.3758/s13428-019-01293-3

Casey, L. S., J. Chandler, A. S. Levine, A. Proctor, and D. Z. Strolovitch. 2017. Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *SAGE Open* 7 (2): 1–15. doi:10.1177/2158244017712774

Casler, K., L. Bickel, and E. Hackett. 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): 2156–60. doi:10.1016/j.chb.2013.05.009

Chandler, J., S. Konrath, and N. Schwarz. 2009. Online and on my mind: Temporary and chronic accessibility moderate the influence of media figures. *Media Psychology* 12 (2):210–26. doi:10.1080/15213260902849935

Chandler, J., P. Mueller, and G. Paolacci. 2014. Nonnaïveté among Amazon mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods* 46 (1):112–30. doi:10.3758/s13428-013-0365-7

Chandler, J., and G. Paolacci. 2017. Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science* 8 (5):500–08. doi:10.1177/1948550617698203

Chandler, J., G. Paolacci, E. Peer, P. Mueller, and K. A. Ratliff. 2015. Using nonnaive participants can reduce effect sizes. *Psychological Science* 26 (7):1131–39. doi:10.1177/0956797615585115

Chmielewski, M., and S. C. Kucker. 2020. An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science* 11 (4): 464–73. doi:10.1177/1948550619875149

Comello, M. L. G. 2013. Activated self-concept as a mechanism underlying prevention message effects. *Media Psychology* 16 (2):177–98. doi:10.1080/15213269.2012.742359

Crump, M. J. C., J. V. McDonnell, and T. M. Gureckis. 2013. Evaluating Amazon's mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8 (3):e57410. doi:10.1371/journal.pone.0057410

DeVoe, S. E., and J. House. 2016. Replications with MTurkers who are naive versus experienced with academic studies: A comment on Connors, Khamitov, Moroz, Campbell, and Henderson. *Journal of Experimental Social Psychology* 67:65–67. doi:10.1016/j.jesp.2015.11.004

Fazio, R. H. 1990. A practical guide to the use of response latency in social psychological research. *Research Methods in Personality and Social Psychology* 11:74–97.

Ford, J. B. 2017. Amazon's mechanical Turk: A comment. *Journal of Advertising* 46 (1):156–58. doi:10.1080/00913367.2016.1277380

Gibson, B. 2008. Can evaluative conditioning change attitudes toward mature brands? New evidence from the Implicit Association Test. *Journal of Consumer Research* 35 (1):178–88. doi:10.1086/527341

Goodman, J. K., C. E. Cryder, and A. Cheema. 2013. Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making* 26 (3):213–24. doi:10.1002/bdm.1753

Goodman, J. K., and G. Paolacci. 2017. Crowdsourcing consumer research. *Journal of Consumer Research* 44 (1): 196–210. doi:10.1093/jcr/ucx047

Greenwald, A. G., D. E. McGhee, and J. L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology* 74 (6):1464–80. doi:10.1037/0022-3514.74.6.1464

Greenwald, A. G., B. A. Nosek, and M. R. Banaji. 2003. Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology* 85 (2):197–216. doi:10.1037/0022-3514.85.2.197

Hauser, D. J., G. Paolacci, and J. Chandler. 2019. Common concerns with MTurk as a participant pool: Evidence and solutions. In *Handbook in research methods in consumer psychology*, eds. Frank R. Kardes, Paul M. Herr,

and Norbert Schwarz, 1st ed., 319–37. New York: Routledge.

Hauser, D. J., and N. Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods* 48 (1):400–07. doi:10.3758/s13428-015-0578-z

Holden, C. J., T. Dennie, and A. D. Hicks. 2013. Assessing the reliability of the M5-120 on Amazon's mechanical Turk. *Computers in Human Behavior* 29 (4):1749–54. doi:10.1016/j.chb.2013.02.020

Kees, J., C. Berry, S. Burton, and K. Sheehan. 2017. An analysis of data quality: Professional panels, student subject pools, and Amazon's mechanical Turk. *Journal of Advertising* 46 (1):141–55. doi:10.1080/00913367.2016.1269304

Klein, R. A., K. A. Ratliff, M. Vianello, R. B. Adams, Š. Bahník, M. J. Bernstein, K. Bocian, M. J. Brandt, B. Brooks, C. C. Brumbaugh, et al. 2014. Investigating variation in replicability. *Social Psychology* 45 (3):142–52. doi:10.1027/1864-9335/a000178

LaTour, K. A., and M. S. LaTour. 2009. Positive mood and susceptibility to false advertising. *Journal of Advertising* 38 (3):127–42. doi:10.2753/JOA0091-3367380309

Luce, R. D. 1986. *Response times: Their role in inferring elementary mental organization*. Oxford: Oxford University Press on Demand.

Nosek, B. A., M. R. Banaji, and A. G. Greenwald. 2002. Harvesting intergroup implicit attitudes and beliefs from a demonstration website. *Group Dynamics* 6 (1):101–15. doi:10.1037/1089-2699.6.1.101

Nosek, B. A., Y. Bar-Anan, N. Sriram, J. Axt, and A. G. Greenwald. 2014. Understanding and using the brief implicit association test: Recommended scoring procedures. *PLoS One* 9 (12):e110938. doi:10.1371/journal.pone.0110938

Perkins, A. W., and M. R. Forehand. 2012. Implicit self-referencing: The effect of nonvolitional self-association on brand and product attitude. *Journal of Consumer Research* 39 (1):142–56. doi:10.1086/662069

Pokhrel, P., P. Fagan, T. A. Herzog, Q. Chen, N. Muranaka, L. Kehl, and J. B. Unger. 2016. E-cigarette advertising exposure and implicit attitudes among young adult non-smokers. *Drug Alcohol Depend* 163:134–40. doi:10.1016/j.drugalcdep.2016.04.008

Read, G. L., I. I. van Driel, and R. F. Potter. 2018. Same-sex couples in advertisements: An investigation of the role of implicit attitudes on cognitive processing and evaluation. *Journal of Advertising* 47 (2):182–97. doi:10.1080/00913367.2018.1452653

Röhner, J., M. Schröder-Abé, and A. Schütz. 2011. Exaggeration is harder than understatement, but practice makes perfect! *Experimental Psychology* 58 (6):464–72. doi:10.1027/1618-3169/a000114

Rouse, S. V. 2015. A reliability analysis of mechanical Turk data. *Computers in Human Behavior* 43:304–07. doi:10.1016/j.chb.2014.11.004

Smith, S. M., C. A. Roster, L. L. Golden, and G. S. Albaum. 2016. A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples. *Journal of Business Research* 69 (8):3139–48. doi:10.1016/j.jbusres.2015.12.002

Stewart, N., C. Ungemach, A. J. L. Harris, D. M. Bartels, B. R. Newell, G. Paolacci, and J. Chandler. 2015. The average laboratory samples a population of 7,300 Amazon mechanical Turk workers. *Judgment and Decision Making* 10 (5):479–91.

Waiguny, M. K. J., M. R. Nelson, and B. Marko. 2013. How advergame content influences explicit and implicit brand attitudes: When violence spills over. *Journal of Advertising* 42 (2–3):155–69. doi:10.1080/00913367.2013.774590

Wood, D., P. D. Harms, G. H. Lowman, and J. A. DeSimone. 2017. Response speed and response consistency as mutually validating indicators of data quality in online samples. *Social Psychological and Personality Science* 8 (4):454–64. doi:10.1177/1948550617703168

Zwaan, R. A., D. Pecher, G. Paolacci, S. Bouwmeester, P. Verkoeijen, K. Dijkstra, and R. Zeelenberg. 2018. Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review* 25 (5):1968–72. doi:10.3758/s13423-017-1348-y