Computers in Human Behavior 29 (2013) 1749–1754

Contents lists available at SciVerse ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Assessing the reliability of the M5-120 on Amazon's mechanical Turk

Christopher J. Holden [*,1], Trevor Dennie [1], Adam D. Hicks [1]

Department of Psychology, Western Carolina University, 91 Killian Building Lane, Room 302B, Cullowhee, NC 28723, United States

ARTICLE INFO

Article history:

Keywords:
Mechanical Turk
MTurk
Personality
Online research
M5-120
Replication

ABSTRACT

Amazon's online service, Mechanical Turk (MTurk) has become a popular option for data collection among social scientists. Early work (Buhrmester, Kwang, & Gosling, 2011) indicated that data collection through MTurk was faster and less expensive than traditional collection methods (undergraduate human subject pool), as well as being reliable when administered at different dates. Building on their work, we sought to extend this investigation of reliability to a larger measure. For the current research we chose a 120-item measure of personality. After collecting data through MTurk, it was determined that our MTurk sample had strong test–retest reliability, indicating that they did not significantly change between administration dates.

1. Introduction

There have been many theories behind the composition of human personality throughout history. Many older theories, such as Freudian theory, were constructed around the presence of aberrant behavior in individuals (Feist & Feist, 2006). However, much of social psychology is not concerned with abnormal behavior; instead, the focus lies on what many term "normal" personality, or specific components of personality that are common throughout the human population (Costa & McCrae, 1992a). The theories that quickly filled this niche are often referred to as trait and factor theories, because they are based on personality factors created through the statistical process of factor analysis (Cattell, 1947; Costa & McCrae, 1995; Feist & Feist, 2006). These factor theories have become very popular as they are well suited for research because their measures are quantifiable and often rely on self-report methods.

Of these factor theories, the five-factor model (FFM) has become one of the more favored. (Digman, 1990). Two of the most prolific and well-known pioneers in the field of FFM research are Costa and McCrae; and their instrument, the Revised NEO Personality Inventory (NEO-PI-R; 1992b), is one of the most widely utilized in trait personality research. The NEO-PI-R breaks the factors of personality into five specific domains with each domain being composed of six underlying facets. For a full list of the domains and facets, see Fig. 1. One drawback of the NEO-PI-R is the fact that it is copyrighted, which inhibits any type of customization and adds additional monetary constraints to researchers.

In response to these concerns, Goldberg (1999) developed the International Personality Item Pool (IPIP) in 1996. The entire IPIP is not copyrighted and is in the public domain, which allows researchers to create unique personality inventories based on the topic of their study. In fact, several proxy instruments, based on well-known instruments such as: the California Psychological Inventory (CPI; Cloninger, 1994); the NEO-PI-R (Costa & McCrae, 1992b); the Sixteen Personality Factor Scale (16PF; Conn & Rieke, 1994); and the Hogan Personality Inventory (HPI; Hogan and Hogan (1992)), have been created using the IPIP.

One such proxy instrument, created to resemble Costa and McCrae's NEO-PI-R (1992b), is referred to as the M5 Questionnaire (Johnson, 2001). Because of its versatility, the M5 Questionnaire comes in several different forms based on survey length, but the most popular form is the M5-120 because it is relatively short (120 items) and is able to provide reliable measures of all five domains and 30 underlying facets (Johnson, 2001). The M5-120 has been compared to the NEO-PI-R and has been shown to be a reliable substitute. For a full list of the corrected correlations between the IPIP and the NEO-PI-R, see Table 1.

1.1. Overview of MTurk

Many researchers are using online data collection platforms (Reimers, 2007), and Amazon's Mechanical Turk (MTurk) has become a popular option. Amazon offers this service free-of-charge, with two types of accounts. The first is the worker account; when a worker logs in, they can choose from a variety of Human Intelligence Tasks (HITs), each of which offer monetary rewards. These HITs consist of tasks that are too complex to be computerized, yet not sophisticated enough that they require specialization on the part of the workers. Typically, HITs are brief and pay only a

* Corresponding author. Address: 492 Timberlea Drive Apartment 101, Rochester Hills, MI 48309, United States. Tel.: +1 919 880 2233.

E-mail addresses: cjholden2@catamount.wcu.edu (C.J. Holden), tmdennie1@catamount.wcu.edu (T. Dennie), adhicks@wcu.edu (A.D. Hicks).

[1] Tel.: +1 828 227 7361; fax: +1 828 227 7005.

0747-5632/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.chb.2013.02.020

- Neuroticism
    - Anxiety
    - Angry Hostility (*Anger*)
    - Depression
    - Self-Consciousness
    - Impulsiveness (*Immoderation*)
    - Vulnerability
- Extraversion
    - Warmth (*Friendliness*)
    - Gregariousness
    - Assertiveness
    - Activity
    - Excitement-Seeking
    - Positive Emotions (*Cheerfulness*)
- Openness to Experience
    - Fantasy (*Imagination*)
    - Aesthetics (*Artistic Interest*)
    - Feelings (*Emotionality*)
    - Actions (*Adventurousness*)
    - Ideas (*Intellect*)
    - Values (*Liberalism*)
- Agreeableness
    - Trust
    - Straightforwardness (*Morality*)
    - Altruism
    - Compliance (*Cooperation*)
    - Modesty
    - Tender-Mindedness
- Conscientiousness
    - Competence (*Self-Efficacy*)
    - Order
    - Dutifulness
    - Achievement Striving
    - Self-Discipline
    - Deliberation (*Cautiousness*)

Fig. 1. The five broad domains and 30 underlying facets. The wording of the facets presented in this figure represents the original wording by Costa and McCrae's Revised NEO-PI-R (1992b). Those words appearing italicized and in parentheses denote facet name changes within the M5 Questionnaire.

Table 1
Corrected correlations between the IPIP scale (M5 Questionnaire) and the NEO-PI-R.

Domains Facets	Corrected correlations
Neuroticism	.93
Anxiety	.90
Angry hostility (*Anger*)	.91
Depression	.92
Self-consciousness	.94
Impulsiveness (*Immoderation*)	.98
Vulnerability	.96
Extraversion	.88
Warmth (*Friendliness*)	.91
Gregariousness	.98
Assertiveness	.99
Activity	.98
Excitement-seeking	.95
Positive emotions (*Cheerfulness*)	.95
Openness to Experience	.92
Fantasy (*Imagination*)	.90
Aesthetics (*Artistic Interest*)	.95
Feelings (*Emotionality*)	.90
Actions (*Adventurousness*)	.99
Ideas (*Intellect*)	.95
Values (*Liberalism*)	.86
Agreeableness	.90
Trust	.95
Straightforwardness (*Morality*)	.86
Altruism	.90
Compliance (*Cooperation*)	.97
Modesty	.95
Tender-Mindedness	.90
Conscientiousness	.88
Competence (*Self-Efficacy*)	.89
Order	.99
Dutifulness	.87
Achievement striving	.97
Self-discipline	.92
Deliberation (*Cautiousness*)	.95

Note: facet names italicized and in parentheses denote name changes in the M5 Questionnaire.

few cents. MTurk was originally developed for commercial use, but a growing number of HITs are dedicated to academic research.

The second option is the requester account. Requesters provide the HITs for the workers to complete, and this is the account used by researchers. This account provides the researcher access to all of the built-in survey tools and to the entire population of workers. Requesters can post multiple HITs at once, select the number of workers desired and collect data simultaneously. Payment for each HIT is determined by the requester, and is then multiplied by the number of workers desired to produce the total cost to the requester (Amazon adds a 10% commission to this cost for their service of paying the workers individually). Thus, the data collection process is streamlined, allowing the researcher to focus on survey design and data analysis.

1.2. Previous personality work using MTurk

In their 2011 article, Buhrmester, Kwang and Gosling (hereafter BKG) take a detailed look at MTurk. After providing a brief overview of MTurk, BKG demographically compare the MTurk samples to other Internet samples, and to samples derived from traditional data collection methods. It was found that workers on MTurk come from over 50 countries, making the sample much more diverse than traditional college samples, and more diverse than other Internet samples (Buhrmester et al., 2011). Keeping in mind that workers must be paid and requesters' preference for inexpensive data collection, BKG manipulated length of task and compensation amount to test whether these two factors contribute to the speed at which data is collected and to the quality of data obtained. While participants were recruited faster for shorter tasks, as well as tasks with higher compensation, data quality was acceptable (as measured by reliability alpha) even in the lowest of payment conditions (Buhrmester et al., 2011).

Most importantly, BKG used the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) to compare the personality data gathered via MTurk to personality data from traditional samples. This comparison was made using reliability alphas, all of which were found to be acceptable (Buhrmester et al., 2011). Furthermore, test–retest reliabilities were used to assess data quality. Personality measures were distributed to participants who consented to both waves of the research, with a three-week period between the two distributions. Test–retest reliabilities ranged from $r = .86$ to $r = .94$ for the five-factors, each of which exceeded the test–retest reliability found in previous literature for the BFI (Gosling, Rentfrow, & Swann, 2003). With such sound psychometrics, BKG considered data collection via MTurk to be promising.

This article by BKG has inspired many researchers to use MTurk for data collection (353 citations in Google Scholar as of February 27th, 2013). As this use of MTurk continues to develop, researchers will begin to use new measures on this online platform, making it necessary to have a full understanding of the reliability of popular measures in their applications on MTurk. However, the BKG paper utilized the BFI which consists of 44 items. Compared to other personality scales such as the NEO-PI-R, which is made up of 240 items, the BFI is a relatively short measure of personality.

Therefore, the researchers sought to replicate the findings of the BKG paper, with an extension to a personality measure that is more nuanced, and is open to the psychology community. Openness of reporting, and of measures used in research is becoming a central issue in psychology today, and efforts in this area have been spearheaded by those such as Brian A. Nosek and Jeffrey Spies (Open Science Framework, www.opensciencefram ework.org). As these efforts continue, it will be important to establish the reliability and validity of openly available measures. This paper provides an assessment of the reliability of one such openly available measure, the M5-120. Furthermore, replication has become a central issue in psychology today, and efforts are being made by many, including those at the Open Science Framework, to replicate existing research (The Reproducibility Project). This paper provides a replication of a well-cited article discussing an area that is receiving much interest at this point in time – i.e. online research methodology.

2. Materials and methods

2.1. Participants

There were 281 participants recruited, 67 of whom agreed to participate in the post-test portion of this study. However, only 46 participants were able to be paired with their data set from the pre-test portion of this study. Twenty-six identified as being male and 20 as female. About 48% of all the participants identified as being Caucasian, 43.5% as Asian, and all other individual ethnic groups (Native American, Hispanic, Pacific Islander, mixed race, and other) made up less than 9% of the total participant population. Participants ranged in age from 21 to 63, with a mean age of 32.87 and a standard deviation of 10.62. The majority of participants, 60.9%, were between the ages of 21 and 31; and 41.3% of participants were from India. All of the MTurk participants were compensated $0.10 for their completion of the pre-test and were then paid $0.15 for their completion of the post-test.

2.2. Measures

M5-120. The M5-120 (Johnson, 2001) is a 120-item measure of normal personality designed to measure the FFM of personality traits: Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience. These five personality traits are assessed using a series of statements rated by a participant on a five-point Likert-type scale, such as: "Make friends easily" (Extraversion), "Love to help others" (Agreeableness), "Like to tidy up" (Conscientiousness), "Panic easily" (Neuroticism), and "Prefer variety to routine" (Openness). This instrument also provides measures for the six underlying facets within each domain. For Extraversion, these six facets are: friendliness, gregariousness, assertiveness, activity level, excitement-seeking, and cheerfulness. The six facets under Agreeableness are: trust, morality, altruism, cooperation, modesty, and sympathy. The six Conscientiousness facets are comprised of: self-efficacy, orderliness, dutifulness, achievement-striving, self-discipline, and cautiousness. For Neuroticism, the six facets are: anxiety, anger, depression, self-consciousness, impulsiveness and vulnerability. Openness to Experience constitutes the last trait, and the facets associated with it are: imagination, artistic interests, emotionality, adventurousness, intellect, and liberalism. The M5-120 is designed to measure all traits and facets using 5-point Likert-type items.

2.3. Procedure

A HIT was placed on MTurk with information related to the survey material. More specifically, the HIT had a title, description and keywords associated with the listing. Personality, psychology, survey, and short were used as keywords for this HIT. Aside from the keywords, no other unique or identifying characteristics were associated with the HIT and it was broadcast to the entire MTurk community. MTurk workers could find the HIT associated with this study by simply scrolling through the listing of other HITs, or they could find the HIT be searching the specific keywords provided. Therefore, workers self-selected to complete the HIT associated with this study.

Once MTurk participants decided to complete our HIT, they were directed to the Qualtrics survey website, which contained the survey materials. First, participants were directed to the informed consent form, where they provided electronic consent by clicking "yes" before continuing to the next page. Participants who did not want to provide consent were asked to exit the survey at that point. Electronic signatures were not requested in order to increase confidentiality. After providing consent, participants completed the M5-120. Next, participants completed a demographics form containing basic information about their age, gender, race/ethnicity, country of origin, and country of residence. Finally, the participants were asked if they would like to participate in a follow-up study in 3 weeks (the post-test). Participants indicated their interest in completing follow-up measures by providing an email address at which they could be reached with a hyperlink to the follow-up measures, which were also completed in Qualtrics.

Once the participants received the follow-up email containing the hyperlink, they were directed to a similar consent form. Once again, participants provided consent by clicking "yes" and continuing to the next page. Upon providing consent, participants were then directed to the M5-120, which they completed in full before providing the same demographic information (age, gender, race/ethnicity, country of origin, country of residence). Once participants completed and submitted their post-test survey, the authors cross-checked their reported demographic data with the information they submitted at pre-test. Those MTurk participants who met adequate levels of consistent and complete responses were compensated within 24 h; and those who did not were still compensated but their information was excluded from the data analysis.

3. Theory

While BKG found that their MTurk personality data sample showed good test–retest reliability, it should be noted that the BFI has fewer items when compared to other broad domain personality instruments. For example, the BFI consists of 44 items, while the NEO-PI-R contains 240 items. Admittedly, short and quick instruments often appeal to researchers targeting online populations because of their ease in completion, which often equates to a higher return rate. However, with the increased use of MTurk as a primary data collection pool by professors and students in university systems, it is likely that many future researchers will attempt to gather data using longer surveys which may not yield as reliable data as shorter surveys. This is not a suggestion that longer tests are less reliable than shorter tests, on the contrary the measurement estimates of internal consistency (a type of reliability) increase with the addition of more items (Pedhazur and Schmelkin (1991)). Instead, we are suggesting that the use of longer instruments may lead to an increase in random measurement error (e.g. miss marking an item, the MTurk worker adopting a random response pattern, etc.) related to the situation of online test taking.

This increase in MTurk use also poses another problem as the MTurk community is a growing and changing body of participants made up of people from many different countries, with their own "real life" culture, "online" culture, and socialization. Since many "Turkers," as they often call themselves, are from foreign countries (BKG, 2011; Holden, 2011) researchers must consider the cultural appropriateness and ramifications when utilizing MTurk. Many of the current instruments in psychology have not been normed across cultures and could be confounded by such cross-cultural use. Also, many Turkers participate in online message boards and other communication activities, such as singling-out researchers that do not regulate the type of responses they accept (e.g. they pay everyone regardless of effort) and posting warnings about certain HITs to avoid (Holden, 2011). Consequentially, such activity may increase the number of invalid participants, or it could even lead to a "black-ball" of the researcher's HIT altogether. While these issues may not affect the results of studies with participant pools numbering in the thousands, they may have a more noticeable effect on the results of studies with smaller samples.

The current study attempted to address the above concerns by completing a conceptual replication of the BKG article. To address the first concern, participants completed the M5-120 instrument instead of the BFI. The M5-120 has 76 more items than the BFI and has a relatively large normed sample (Johnson, 2001). Second, the participant population was capped at less-than 300 in order to study reliability when using smaller samples. Finally, there were no restraints utilized in collecting MTurk participants, thus allowing anyone from any country to participate, and for all the participants to be paid regardless of the effort they put forth. The results will not only indicate the reliability of MTurk populations in small sample research studies; they will also indicate possible cultural confounds when using data collected from MTurk participants.

4. Results and discussion

4.1. Results

A paired-samples *t*-test was conducted to examine differences in Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience; comparing the pre-test scores of the participants with their post-test scores. There were no statistically significant differences. This suggests that the scores for those who completed the follow-up measures did not significantly differ between survey administration periods. Means, standard deviations, test values, and significance for these five factors are presented in Table 2.

Pearson's *r* correlations were used to determine test–retest reliability between each of the five factors for participants' pre-test and post-test scores. These correlations ranged in value from $r = .79$ to $r = .91$ and were all significant at the $p < .01$ level. This suggests that the scores for those who completed the follow-up measures were significantly correlated between survey administration periods. Each of the correlations and their corresponding significance levels are provided in Table 3.

Taken together, these results suggest that personality traits can be reliably measured through the use of FFM scales on MTurk. Furthermore, these results are directly in line with the results of the BKG paper. Although a different scale was used, these findings provide a conceptual replication of their work. Finally, these results support the reliability of the MTurk population in the context of psychological research – though researchers should be aware of several limitations that may make data collection more difficult than recruiting a college sample (see discussion).

Table 3
Test–retest reliability.

Factor	Correlation
Extraversion	$r = .90^*$
Agreeableness	$r = .86^*$
Conscientiousness	$r = .82^*$
Neuroticism	$r = .79^*$
Openness	$r = .91^*$

* $p < .001$.

4.2. Conclusions

The BKG article has become a primary work on conducting research via MTurk. Since its publication, the article has spurred a great deal of research and has been cited 353 times on Google Scholar (As of February 27th, 2013). There is also a growing body of literature on MTurk and its applicability to social sciences (Horton, Rand, & Zeckhauser, 2011; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010, & Rand, 2012). The results of the current study add credence to the possibility of carrying out studies on personality using MTurk. Through the use of paired-samples *t*-tests, it was determined that participants' pre and post-test responses on the M5-120 did not differ significantly. Furthermore, Pearson's *r* correlations indicate good test–retest reliabilities for all personality domains on the M5-120 (see Table 3). These results are comparable to those found by BKG using the BFI (*r* ranges from .86 to .94).

There were some notable differences between the BKG study and the current study. The BKG study was carried out at a large Midwestern university, and may have had more financial resources available than smaller schools. The current study was conducted at a small university in the Southeast with minimal financial resources available. Out of nearly 300 participants paid for completing the pre-test, and then 67 of those getting paid for the post-test as well, only 46 provided usable data. This roughly equates to a $40.00 investment, or 10% of the departments MTurk budget, equating to nearly $1.00 per subject. Thus, schools with restricted budgets may run into limitations when conducting research via MTurk. Also, this study utilized a different personality instrument (M5-120) than the one used by BKG (BFI), with over twice as many items. Since the M5-120 proved to be as reliable as the BFI, it indicates that MTurk workers provide reliable data even when completing longer surveys. However, since MTurk is still a relatively new data collection tool (in comparison to the traditional subject pool), measures should be investigated before being trusted as reliable on MTurk. Researchers may also be faced with other obstacles that can influence the outcomes of the HITs they post on MTurk.

Table 2
Paired-sample *t*-tests: means, standard deviations, *t* values, and significance level.

Factor	Pre-test	Post-test	*t* value	Significance
Extraversion	$M = 78.30$, $SD = 15.40$	$M = 78.20$, $SD = 15.74$	$t(45) = .104$	$p = .918$
Agreeableness	$M = 87.96$, $SD = 11.71$	$M = 88.33$, $SD = 11.66$	$t(45) = .–400$	$p = .691$
Conscientiousness	$M = 91.43$, $SD = 11.57$	$M = 90.00$, $SD = 13.15$	$t(45) = 1.277$	$p = .208$
Neuroticism	$M = 66.20$, $SD = 13.36$	$M = 65.30$, $SD = 13.36$	$t(45) = .946$	$p = .349$
Openness	$M = 81.33$, $SD = 11.83$	$M = 81.39$, $SD = 12.84$	$t(45) = .–083$	$p = .934$

4.3. Discussion

Findings from this study, as well as those from the BKG paper show that MTurk can be used to obtain valid and reliable results. More specifically, with the corroborating findings between different personality scales, researchers can confidently use MTurk to gather personality data. These findings are promising and offer researchers with new opportunities and a new avenue for data collection. However, researchers must be aware of potential limitations before conducting research on MTurk, and further research should be conducted to assess the feasibility and validity of using this data collection platform.

The first limitation is the increased anonymity provided to workers on MTurk. Unless the researcher asks for personal information, the only identification provided by Amazon is the MTurk worker ID, which is a string of letters and numbers, typically 13 characters in length. This ID is all that is needed to pay the worker. Occasionally, the researcher may ask the worker to provide a random number, which is generated in their survey, unique to that worker, for another means of matching the participant to their data. Although this level of anonymity expedites the IRB process, it can make the work of the researcher more hectic and stressful.

If a researcher desires to contact their subjects at a later date to complete a second wave of the survey, as was done in this project, the researcher must ask that the workers provide an email, or if the researcher is familiar with computer programming, a Python script created by Mueller and Chandler (2012) can be used. Some workers may be hesitant of providing their emails, as there are scam HITs occasionally posted on MTurk, and researchers should note that they will likely have fewer respondents in their follow-up. A second outcome of this heightened anonymity is a lack of control over repeat responders. It is possible, on the part of the researcher, to deny work from the same worker ID and avoid repeat responders by that means, but in the case where researchers do not have any additional personal information, they cannot be certain whether workers have multiple accounts. If a worker were to have multiple accounts, they could complete a survey multiple times unbeknownst to the researcher. It is likely that Amazon regulates the number of accounts one person has (Rand, 2012), reducing the likelihood of multiple accounts, however, this possibility exists. In addition to this, the researcher can investigate any suspicious data by comparing demographics between subjects and excluding any matches or subsequent responses. This screening technique does provide additional security, but a savvy worker may be aware of this and could possibly alter their demographics on subsequent responses to ensure payment for all of their responses.

Another limitation of MTurk is the reliance on self-report measures, and other similar designs. While studies using self-report measures are prevalent in psychology, this is the only option available when conducting research using MTurk. As mentioned by BKG, subjects cannot come into the lab for observation on MTurk, excluding social scientists who want to go beyond questionnaires. If an external survey site is used, like Qualtrics, photos and videos can be uploaded, opening some possibilities, but not all of those afforded by in-person lab research.

Finally, MTurk is open to an international population of workers. While this will make samples demographically diverse, cautions should be taken when using materials that are culturally sensitive. For example, the Minnesota Multiphasic Personality Inventory-2 (Butcher et al., 2001), which utilized a norming sample based on an American population, needed to be modified before use in other cultures in order to address specific cultural concerns beyond simple translation issues (see Cheung, 2009). It should be noted that if a researcher was interested in using the MMPI-2 (or the MMPI-2-Restructured From; Ben-Porath & Tellegen, 2008) online, that they would have to seek permission from the University of Minnesota Press prior to doing so. MTurk does allow the researcher to limit their HITs to workers in the United States, or any country of their preference, but the decision to do this should be made before collecting any data.

Having considered the limitations, the current study still provides new insights on the use of MTurk in conducting social science research. First, the researcher must take preliminary steps before conducting full scale research projects on MTurk. Some of these preliminary steps include pilot testing their instruments if no reliability testing has been conducted on the MTurk population. Past studies have shown this to be less necessary if the researcher limits their sample to the United States (Hicks, Dennie, Taurasi, & McCord, 2011); however, caution is still advised, especially with instruments that have little psychometric information available.

Another consideration that should be made by MTurk researchers is learning as much as possible about MTurk before conducting a study using that population. This is even more paramount if the researcher is not familiar with computers or online community websites. This goes beyond the basics of learning how to deposit money in the requester account and creating/publishing a HIT. Researchers need to consider the use of key-words for their HIT, how they will keep track of the workers taking their HIT to insure no repeat survey takers, how long to leave their HIT active once they begin to collect data, and the specific payment schedule for workers who complete their HIT. It would take far too much time to discuss the ins-and-outs of MTurk use; however, a separate article is being prepared discussing some of the authors' experiences working with the MTurk population, including tips on how best to use key words and how to keep track of workers through the aid of a random number generator.

Besides these implications for individual researchers, this study also suggests several implications to the broader arena of psychological research. For instance, several ethical issues arise when using any online community as a participant population. MTurk complicates matters because it introduces the specific inducement of monetary rewards, which can create problematic issues when presenting research proposals to the IRB. The current authors had to concede to pay all participants, regardless of their effort and ability to pass the validity checks within the survey. While this may seem to be nothing more than a streak of frugality on the authors part, workers may be targeting Requesters who blindly pay (i.e. do not filter for valid responses) all participants. This is based on the number of repeat responders that were deleted from this sample (76 out of 242), as well as the number of repeat responders the authors have had in other studies (14 out of 211).

One final broad ethical issue is the consideration of responses from minors. There are certain natural check-points in the MTurk system that keeps many minors from being able to complete HITs. The most salient safe-guard is that workers must have an active Amazon account, which requires an active credit card. This is by no means fool-proof, as it is possible for minors (those under 18) to obtain a credit card. There are more direct means that are available to the researcher, such as specifically asking the participant to disclose their age, or endorse an agreement that states that they are at least 18 years-old; however, if the worker is told they cannot participate in the research if they are not 18, then there is always the chance of a subject lying to receive payment. One possible solution would be to use deception by not telling a participant that they will be excluded from the research if they are under the age of 18; but this also causes more ethical dilemmas. For instance, would the underage worker still be paid and their data removed from analysis, or would they not be paid and given an explanation as to why? Currently, there is no precedent to follow, which means that different university IRB standards become the basis for such decisions. If MTurk research continues to be a popular source of participants, overarching APA standards need to be considered so that researchers have clear, ethical guidelines to follow.

In conclusion, MTurk is a popular, fast, and comparably cheap means of gathering psychological research participants. Though many studies are limited to self-report and survey research, it has been shown to provide reliable data, comparable to the normal, college undergraduate samples utilized by most other research. However, researchers must have an understanding of the MTurk environment before they can best utilize this resource, and fully consider variables that may adversely affect their ability to collect usable data. Therefore, researchers should continue to investigate the use of MTurk as a means for data collection. For example researchers could work to determine whether certain populations can be reached through MTurk, such as clinical samples, the LGBT community, and special interest groups. In addition, research should be conducted to establish the reliability and validity of measures outside the realm of clinical and personality psychology. That is, future research should be conducted to establish the reliability and validity of other widely-used scales such as the Rosenberg Self-Esteem Scale (Rosenberg, 1965) and specialized scales such as the Contingent Self-Esteem Scale (Paradise & Kernis, 1999) and the Mate Retention Inventory (Buss, 1988; Buss, Shackelford, & McKibbin, 2008).

Continued research on the general demographics of MTurk is also necessary. Although this has been established across various papers and studies, researchers should maintain their understanding of these demographics, as shifts may occur with the growth of MTurk and these shifts may influence responses obtained via MTurk. Additionally, much of the research on MTurk has relied on self-report measures. Further research is needed for different types of techniques and designs. Although the MTurk platform is a bit limited, with an understanding of computer programming the capabilities increase. Furthermore, many researchers only use MTurk to broadcast their survey, and direct participants to an external website such as Qualtrics or SurveyMonkey. These survey design packages allow for the use of more sophisticated techniques such as the presentation of images and video. Research should be conducted to determine whether these more sophisticated techniques can be successful on MTurk. Researchers have already demonstrated that cooperation experiments are possible on MTurk (Rand, 2012), and new techniques should be researched as they are developed. Finally, if MTurk remains a sought-after research tool, then APA standards should be made to better accommodate researchers, as well as provide precedents for IRBs to better adjudicate research proposals.

References

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2-RF Minnesota multiphasic personality inventory – 2-restructured form: Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press by NCS Pearson, Inc..

Buhrmester, M. D., 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.

Buss, D. M. (1988). From vigilance to violence: Tactics of mate retention in American undergraduates. *Ethology and Sociobiology, 9*, 291–317.

Buss, D. M., Shackelford, T. K., & McKibbin, W. F. (2008). The mate retention inventory-short form (MRI-SF). *Personality and Individual Differences, 44*, 322–334.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *The Minnesota multiphasic personality inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Cattell, R. (1947). Confirmation and clarification of the primary personality factors. *Psychometrika, 12*, 197–220. http://dx.doi.org/10.1007/BF02289253.

Cheung, F. M. (2009). The cultural perspective in personality assessment. In J. N. Butcher (Ed.), *Oxford handbook of personality assessment* (pp. 44–58). New York: Oxford University Press.

Cloninger, C. R. (1994). *The temperament and character inventory (TCI): A guide to its development and use*. St. Louis. MO: Washington University, Center for the Psychobiology of Personality.

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

Costa, P. T., & McCrae, R. R. (1992a). Four ways five factors are basic. *Personality and Individual Differences, 13*(6), 653–665. http://dx.doi.org/10.1016/0191-8869(92)90236-I.

Costa, P. T., & McCrae, R. R. (1992b). *Revised neopersonality inventory (NEO-PI-R) and neo five-factor inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1995). Primary traits of Eysenck's P–E–N system: Three- and five-factor solutions. *Journal of Personality and Social Psychology, 69*(2), 308–317. http://dx.doi.org/10.1037/0022-3514.69.2.308.

Digman, J. (1990). Personality structure: Emergence of the five factor model. *Annual Review of Psychology, 41*, 417–440. http://dx.doi.org/10.1146/annurev.ps.41.020190.002221.

Feist, J., & Feist, G. J. (2006). *Theories of personality* (6th ed.). New York, NY: McGraw Hill.

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. Jr., (2003). A very brief measure of the big-five personality domains. *Journal of Research in Personality, 37*, 504–528. http://dx.doi.org/10.1016/S0092-6566(03)00046-1.

Hicks, A. D., Dennie, T., Taurasi, J., & McCord, D. M. (2011, April). The Association Between Anxiety, Depression, and Personality. Poster session at the North Carolina Psychological Association Spring Conference, Chapel Hill, NC.

Hogan, R., & Hogan, J. (1992). *Hogan personality inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Holden, Christopher J. (2011, March). Amazon's MTurk: A flawed source of participants in psychology studies? Paper session presented at Western Carolina University's Nineteenth Annual Graduate Research Symposium, Cullowhee, NC.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*, 399–425. http://dx.doi.org/10.1007/s10683-011-9273-9.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory–versions 4a and 54*. Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

Johnson, J. A. (2001, May). Screening massively large data sets for non-responsiveness in web- based personality inventories. Invited talk to the joint Bielefeld-Groningen Personality Research Group, University of Groningen, The Netherlands.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's mechanical Turk. *Behavioral Research Methods, 44*(1), 1–23. http://dx.doi.org/10.3758/s13428-011-0124-6.

Mueller, P. & Chandler, J. (2012) *Emailing workers using python*. Available at SSRN: <http://ssrn.com/abstract=2100601> or <http://dx.doi.org/10.2139/ssrn.2100601>.

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

Paradise, A. W., & Kernis, M. H. (1999). *Development of the contingent self-esteem scale*. University of Georgia: Unpublished data.

Pedhazur, E. J., & Schmelkin, L. P. (1991). Reliability. In E. J. Pedhazur & L. P. Schmelkin (Eds.), *Measurement, design and analysis: An integrated approach* (pp. 81–117). Hillsdale, NJ: Lawrence Erlbaum and Associates.

Rand, D. G. (2012). The promise of mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172–179. http://dx.doi.org/10.1016/j.jtbi.2011.03.004.

Reimers, S. (2007). The BBC internet study: General methodology. *Archives of Sexual Behavior, 36*(2), 147–161. http://dx.doi.org/10.1007/s10508-006-9143-2.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.