A comparative digital survey investigation of the construct validity of the Trait Anxiety Inventory within a UChicago community sample and an MTurk sample

Nora Nickels

6/6/2018

Abstract

In psychological and biological research, quantitative scales are used to measure individual differences in stable traits across populations. In our research, our lab seeks to measure trait anxiety as a potentual stable trait that may predict how individuals respond to acute anxiety. However, we are often concerned over both the constrcut validity of our measures, as well as the ability to consider external validity of our findings within a WEIRD population of a University community. In this study, we ask the following questions: 1) Do distributions of trait anxiety scores differ in samples acquired from a University of Chicago community vs. the Amazon Mechanial Turk community? 2) How strong is the construct validity of the Trait Anxiety Inventory in a sample drawn from these two populations; specifically, do setting and mood relate to trait anxiety responses of UChicago or MTurk community members when the T.A.I. is completed outside of a controlled laboratory setting? We found that individuals from a University of Chicago community sample and individuals from an MTurk sample do not significantly differ in trait anxiety scores. Further, we found that positive and negative acute mood are the strongest predictors of trait anxiety scores amongst various demographic and extraneous measures that have the potential to affect trait anxiety scores when measured digitally. These results are an important explorative first step in the study of how digital measurement of persaonlity traits has to potential to introduce extraneous variables when measuring what should theoretical be stable inherent traits.

Introduction

In psychological and biological research, emotions are measured both in terms of acute states of arousal and in terms of individual differences in the propensity to experience that emotion. Scientists define the difference between these two elements of emotional measurement as state and trait measures of emotions. For example, anxiety as an emotion can be generally defined as heightened feelings of tension, apprehension, and worry, in combination with an aroused physiological state (Charles D Spielberger 2010). It is particularly important to distinguish between state and trait for anxiety, as higher trait anxiety, or higher individual proneness to experience anxiety, could affect the way an individual reacts behaviorally in both acute and long term situations. Trait anxiety can be defined as an individual measure of intensity and frequency of experienced anxiety, which involves these feelings of apprehension and heightened response of the autonomic nervous system (Charles D Spielberger 1966). Importantly, trait anxiety is seen as a relatively *stable* trait, and individuals who have higher trait anxiety tend to perceive situations as more dangerous or stressful over time (Charles D Spielberger 1966).

The State-Trait Anxiety Inventory, or STAI, is a long-standing measure that uses two scales to report these two measures (state anxiety and trait anxiety)(Charles Donald Spielberger 1989). The STAI is designed as a self-report measure, with items that map specifically onto the two factors of anxiety. The STAI trait scale consists of twenty statements that have individuals rate, on a four-point Likert scale, different statements about how they feel generally (e.g., "I feel nervous and restless."). Both the state and trait scales of the STAI are long-standing, frequently used scales in psychology, and theoretically, the inventory has been shown to measure response to experimental manipulation in meaningful ways (Chapman and Cox 1977). Further, the two subscales have been shown to correlate with other measures of anxiety that is consistent the content of measure (Bieling, Antony, and Swinson 1998).

Importantly, retest correlations of the iventory have shown strong reliability, and re-test coefficients for the trait scale have shown to be even higher for those items that measure the trait scale (Charles D Spielberger 2010; Barnes, Harp, and Jung 2002). The STAI is reported to have high validity, with concurrent validity with other anxiety questionnaires reported as ranging from 0.73 - 0.85 (Bieling, Antony, and Swinson 1998). However, some researchers argue that a general, yet incorrect, implication that is attached to re-test reliability is that which assumes that once an instrument is found to be reliable, its reliability does or cannot change (Barnes, Harp, and Jung 2002). If reliability is simply a property of scores from a specific sample of survey-takers, as opposed to being a property of the test itself, then reliability of a measure can be affected by any source of variability that also affects the scores (e.g., demographics in a particular sample, such as gender, age, motivation, mood, etc.)(Barnes, Harp, and Jung 2002). Therefore, although re-test reliability and concurrent measures of validity are incredibly important, considering the specific sample involved in one's study is crucial in discussion of the interpretation of one's results.

In my work, I share equal concern in that my specific sample is taken from a community whose specific demographics may affect the distribution of anxiety scores. Like most psychology study populations, our work frequently involves participant samples drawn from a university setting. Specifically, the University of Chicago ranks as one of the top undergraduate research institutions in the U.S., and is often viewed as a competitive and stressful environment. Beyond the concern that many research institutions have about their willing research participants coming from a primarily Western, educated, industrialized, rich, and democractic (WEIRD) population (Jones 2010), our lab also deals with the concern of recruiting willing participants from a sample that may not only have higher than usual scores of trait anxiety, but also have rapidly fluctuating rates of both state and trait anxiety throughout their academic experience.

To control some of these concerns of validity and reliability, researchers often use a controlled, laboratory setting, to remove extraneous effects of the environment. For example, our laboratory has research participants spend about twenty-five minutes in the laboratory before first saliva samples are taken, to reduce the potential for effects outside of the lab to result in hormonal concentrations that do not represent true baseline. In this way, we also administer many psychological surveys in the lab as well. Howevers, due to both time and monetary restraints, we occasionally adminster *trait* based questionnaires digitally in advance of the lab session, as trait based questionaires theoretically measure relatively stable personality measures.

Digital surveys and digital ethnographic methods are seen as new technologies for social research that allow scientists to avoid more costly research methods, to easily alter questionnaires to access different cultural groups, to access hard-to-reach populations, to collect higher response rates, and to consolidate data more quickly and efficiently (Murthy 2008). In the case of administering our surveys digitally outside of the lab, we save both temporal and monetary costs, yet run the risk of extraneous factors of the environment interacting with demographics of our sample and therefore affecting the trait anxiety scores of our participants. If certain factors environmentally outside of a controlled laboratory could affect trait anxiety scores, then we experience a trade off of losing validity when our survey is administered digitally.

Obviously, our sample taken from the University of Chicago community is not the only sample from which digital survey data is drawn. Digitally web-based data collection is a relatively new method that contains the primary elements needed to conduct social scientific research, while benefitting from the same aspects discussed above. In fact, despite the concerns of losing the control of a laboratory environment, some researchers have argued that survey data that is digitally collected are in fact preferred to data that has been collected in-person. For example, Castler et al. compared data that had been collected in the lab and also online, and found that the test results themselves resulted in equivalent, high-quality data for both groups, and that the data collected digitally was in fact more socioeconomically and ethnically diverse (2013). Further, Hauser and Schwarz found that data digitally collected using Amazon's Mechanical Turk (MTurk) showed higher rates of participant attentiveness (measured using attentiveness an instructional manipulation check) when compared to data collected from college students (2016).

On the other hand, other research that compares samples collected from digital populations with in-person samples have found differences that may be less beneficial. Further, even if both samples are collected digitally but come from differing populations, the samples themselves may compare and contrast in interesting ways, based on the population from which the digital survey sample is taken. For example, Goodman and colleagues compared MTurk participants with student samples on multiple measures, including attentiveness, personality, and certain decision-making biases (2013). The authors found that MTurk participants were actually less attentive and had different personality profiles (e.g., less extroverted, less emotionally stable) when compared to a student population, but were similar in terms of how they value money and time and in terms of their risk aversion (Goodman, Cryder, and Cheema 2013).

Clearly then, the results of this line of research have been mixed thus far. What we can confirm is that much of the literature focusing on the strengths and weaknesses of digital data has focused specifically on globally digital as opposed to local populations, where data can be crowdsourced or collected in a completely digital way. Conducting psychology research using crowdsourced data has largely revolved around the Amazon Mechanical Turk (MTurk) platform, based on its popularity and ease of access. MTurk provides a platform to outsource small tasks (referred to as HITS, or human intelligence tasks) to a workforce collected globally that is made up of "workers" (Behrend et al. 2011). The MTurk platform has been investigated to confirm that it provides an efficient and reliable alternative from the university participant population (Behrend et al. 2011; Rand 2012). Further, MTurk has been used to successfully replicate experimental work, showing its viability in terms of experimental design and validity flexibility (Berinsky, Huber, and Lenz 2012).

In particular, a solid amount of work has been done investigating the specific MTurk population, focusing on the demographics, responsivity, and motivation of the community of MTurk workers. Many studies show that the demographics of MTurk workers fluctuate, and that depending on the research questions being asked, researchers must use caution when selecting participants by filtering targeted study pools on MTurk [Huff and Tingley (2015); Ross et al. (2010); Casey et al. (2017);]. Others suggest that the pros and cons of using the MTurk pool are based on both controllable and uncontrollable factors, and that often the benefits, such as accessing hard-to-reach populations, exceed the downsides of the use of in-person populations and lab studies (Paolacci and Chandler 2014; Smith et al. 2015). Fields of psychology, political science, and industrial / organization psychology in particular pay particular attention to the personality characterisics and ideology of the MTurk pool, as those factors are incredibly important when considering the external validity of individual characteristics of one's research participants (Bates and Lanza 2013; Clifford, Jewell, and Waggoner 2015; Woo, Keith, and Thornton 2015). Overall, there has been much discussion regarding the methodology of MTurk sampling and the MTurk population, as its promise of accessing high quality, inexpensive data is ground breaking to many lines of research (Buhrmester, Kwang, and Gosling 2011).

Based on this literature and common restrictions of both money and time, our lab continues to have standing concerns based on the comparison between our sample population, drawn from the UChicago community and containing many undergraduate college students, and a sample population coming from a wider population, such as the the MTurk commmunity. Past research has discussed the pros and cons of data collection from in-person vs. digital methodologies, and it is critical to know the specific descriptive statistics of a specific sampling frame, and how these descriptions differ from other sample populations, such as a wider and argulably more externally valid, global community. In particular, our use of psychology research is invested in stable personality, emotional, and psychological traits that map on to biological and behavioral responses. Therefore, we are focused on the distribution of stable traits in our population, how this distribution differs from other samples, and how the scores that lead to this distribution are impacted by extraneous factors. This study seeks to answer, specifically, how the TAI scores of a sample from a UChicago community compare to this collected from a digitally crowdsourced sample from Amazon Mechanical Turk. These comparisons in scores could be tied to differences in specific anxiety traits between the two samples, or differences in diversity amongst the groups. Further, this study will look into the construct validity of both the UChicago sample and the MTurk sample, by focusing on how extraneous factors, such as setting and mood, affect the responses of the TAI for both a UChicago based sample and an MTurk collected sample. The focus on these factors will add to the literature surrounding how in-person vs. digitally collected data compare.

Theoretical Model

Our lab's research explores stable personality, emotional, and psychological traits that map on to biological and behavioral responses. In this study, we are focused on the distribution of a stable trait (trait anxiety) in our population, how this distribution differs from other samples, and how the scores that lead to this distribution are impacted by extraneous factors. This study seeks to answer, specifically, how the TAI scores of a sample from a UChicago community compare to this collected from a digitally crowdsourced sample from Amazon Mechanical Turk. This study will look into the construct validity of both the UChicago sample and the MTurk sample, by focusing on how extraneous factors, such as setting and mood, affect the responses of the TAI for both a UChicago based sample and an MTurk collected sample.

To do so, we will use a multiple linear regression model that will identify key variables that predict trait anxiety scores in our digital survey participants. The survey explores our main predictor variable of community or group (whether you were recruited from a UChicago population or an MTurk population), along with a variety of demographic factors that seek to control for what else may predict TAI scores. These potential predictors of our model will include age, gender, income, research participation experience, where the survey was taken (physical setting), and positive and negative acute mood.

Current Aims and Hypotheses

Methods

Variables

To measures trait anxiey, we used the trait scale of the State Trait Anxiety Inventory. The STAI trait scale consists of twenty statements that have individuals rate, on a four-point Likert scale, different statements about how they feel generally (e.g., "I feel nervous and restless."). The items were randomized for each participant. To measure acute positive and negative mood of participants, we will use the Positive and Negative Affect Schedule (PANAS; Watson et al., 1988). The PANAS is a psychometric scale that measures positive and negative affect using words that describe feelings and emotions on a Likert scale of 1 (not at all) to 5 (very much). A multitude of studies have been used to confirm the reliability and validity of the PANAS. We will use the PANAS scales to measure how acute positive or negative moods affect the trait anxiety scores of the participants. PANAS measures will be collected after the TAI is administered, so as not to prime participants' with the current mood and thus affect the TAI scores. The PANAS items were also randomized for each participant.

To measure the setting of where the survey was taken, we included items asking about type of setting, the amount of people present at the time, and whether the participant was interacting wtih anyone else at the time. To measure demographics variables, we included items asking questions about age, gender, occupation, income level, relationship status, gender, and whether or not the participant has taken part in a research study prior to this one.

Data

Data collection and Participants

Data from this study is collected from two different populations. One sample was collected from a population of MTurk workers, and one sample was collected from the UChicago community. Both groups completed the same survey that was administered via Qualtrics. The Qualtrics survey was anonymous, and included measures in the following order: Trait Anxiety Inventory, PANAS, questions regarding setting, and finally demographic variables.

We surveyed 104 MTurk workers on Thursday, May 3rd of 2018. The survey was advertised as taking about 5 minutes and we paid paid respondents 15 cents each. Because we wish to benchmark MTurk against a sample of University of Chicago community members, we resitrcited the survey to individuals classified as 18

Table 1: Demographics [a]

Group	N	% Female	Mean Age	Mode Income Level	Mode Setting	% of Prior Research Partici-pants	% Single	% Married
Mturk	104	54.8%	37.3	45k - 60k	In Home / Apartment	81.2%	51.9%	34.6%
U of C	96	68.8%	22.8	15k - 30k	In Home / Apartment	60.6%	63.5%	5.2%
Overall	200	61.5%	30.4	15k - 30k	In Home / Apartment	70.5%	57.5%	19.0%

[a] Prior research experience defined as having answered yes to participanting in a research study previously.

or older and living in the United States. Further, we excluded individuals with approval ratings below 90% on previous MTurk tasks.

We surveyed 96 University of Chicago community members on Thursday, May 3rd of 2018. The survey was advertised through several platforms that are exclusively accessbily to individuals with a University of Chicago certified email address. These platforms include: UChicago Marketplace; UChicago private Facebook groups; UChicago private listservs; and UChicago current student class email lists (approved accessed by individual course instructors).

The data is accessible within our github repository: https://github.com/nnickels/MACS30200proj.

Summary Statistics

Table 1 presents key summary statistics of demographics amongst both the UChicago and MTurk groups, as well as overall. Both samples included more females than males, with the UChicago sample having 69% female participants. The mean age of MTurk workers (37.3 years) was significantly greater than that of the UChicago participants (22.8 years), and MTurkers were more likely to be married than UChicago students (34.6% married vs. 5.2% married). Both UChicago students and MTurkers most frequent response in terms of setting where the survey was completed was at home or in their apartment. Table 2 presents key summary statistics of TAI and PANAS scores among both the UChicago and MTurk groups, as well as overall.

Data Analysis

To compare mean TAI scores between UChicago and MTurk participants, we use a t-test of mean TAI scores between the groups. To analyze the effects of group (UChicago vs. MTurk) on TAI scores while controlling

Table 2: Descriptive Statistics of TAI and PANAS Inventories [a]

Inventory	Mturk	U of C	Overall
Mean TAI Score	45.41	45.68	45.54
St. Dev. TAI Score	12.64	11.48	12.07
Min TAI Score	21.00	25.00	21.00
Max TAI Score	75.00	70.00	75.00
Mean Positive Mood Score	30.31	28.80	29.58
St. Dev. Positive Mood Score	8.76	8.43	8.61
Min Positive Mood Score	11.00	10.00	10.00
Max Positive Mood Score	50.00	48.00	50.00
Mean Negative Mood Score	17.21	18.23	17.70
St. Dev. Negative Mood Score	8.05	6.78	7.46
Min Negative Mood Score	10.00	10.00	10.00
Max Negative Mood Score	47.00	41.00	47.00

[a] Positive and Negative mood scores assessed with the positive and negative scales of the PANAS. TAI Total scores calculated from the Trait scale of the STAI.

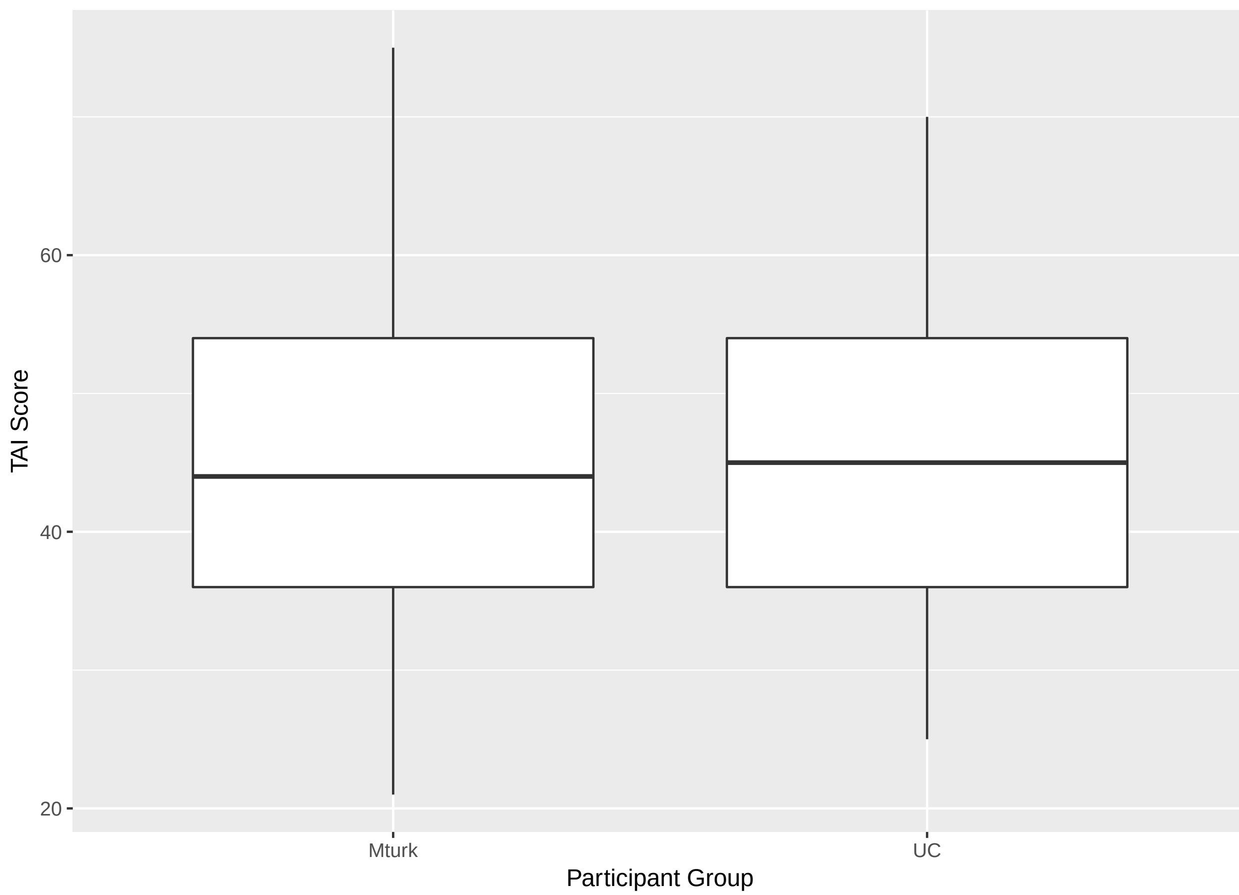
for demographics and other potential extraneous factors, we use a multiple linear regression model with TAI score as our dependent variable. Alpha was set at 0.05

Results

Our results seek to answer the following research questions: 1) Do distributions of trait anxiety scores differ in samples acquired from a University of Chicago community vs. the Amazon Mechanial Turk community? 2) How strong is the construct validity of the Trait Anxiety Inventory in a sample drawn from these two populations; specifically, do setting and mood relate to trait anxiety responses of UChicago or MTurk community members when the T.A.I. is completed outside of a controlled laboratory setting?

In comparing average Trait Anxiety Scores between UChicago (M = 45.68) and MTurk samples (M = 45.41), we found no difference in TAI Scores between the two groups (t = -0.155, p-value = 0.877). Figure 1 shows the average TAI score between groups.

Fig. 1: UChicago & MTurk Participants Do Not Differ in TAI Score

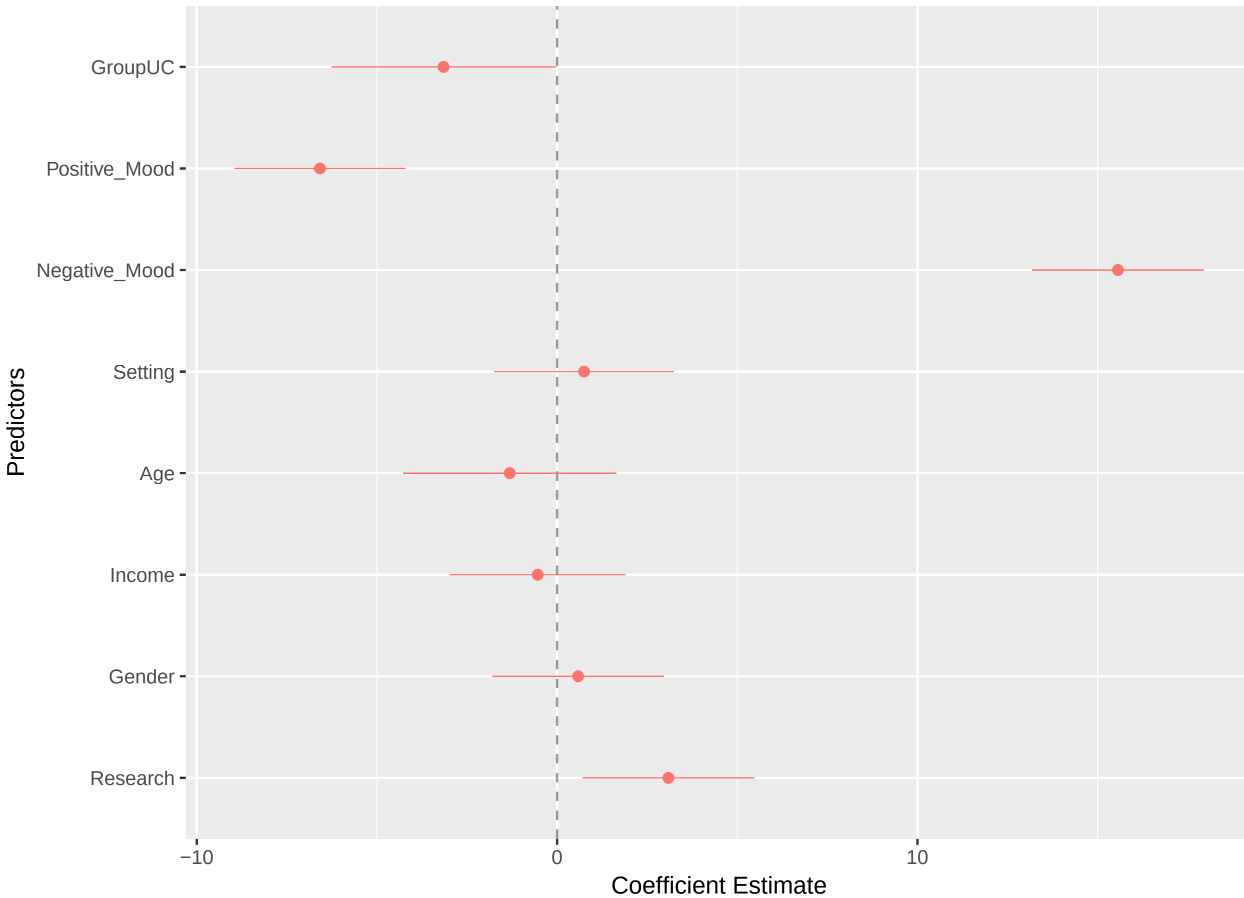


A multiple linear regression was calculated to predict TAI scores based on Group, with further predictors of positive acute mood, negative acute mood, setting, age, income level, gender, and previous research participant experience. Table 3 presents the regression model results. A signification regression equation was found, such that positive acute mood, negative acute mood, and past research participation were all significant predictor of TAI scores, while controlling for setting, age, income, and gender. Figure 2 presents the regression results as the significance and 95% confidence intervals of the coefficient estimates for each of the predictors. Positive acute mood is a negative predictor of TAI score. Negative acute mood is a positive predicor of TAI score. Past research participant experience is a slight positive predictor of TAI score. Group (U of C vs. MTurk) is a slight negative predictor of TAI score, such that, controlling for all other model predictors, being a U of C student increases TAI score by 3.15.

Table 3: Regression Model Results

term	estimate	std.error	statistic	p.value
(Intercept)	34.7391055	4.7123210	7.3719734	0.0000000
GroupUC	-3.1505441	1.5772899	-1.9974414	0.0472270
Positive_Mood	-0.3821156	0.0699989	-5.4588789	0.0000002
Negative_Mood	1.0376737	0.0807471	12.8509041	0.0000000
Setting	0.3723475	0.6295075	0.5914902	0.5549063
Age	-0.0531344	0.0606556	-0.8760030	0.3821522
Income	-0.0703395	0.1621634	-0.4337572	0.6649647
Gender	0.5304448	1.0975846	0.4832837	0.6294597
Research	3.4098659	1.3371631	2.5500748	0.0115716

Fig. 2: Predictors of Trait Anxiety Scores



Conclusion

In this study, we report that individuals from a University of Chicago community sample and individuals from an MTurk sample do not significantly differ in trait anxiety scores. We also report that, when controlling for demographic variables, sample group only marginally predicts trait anxiety scores. Instead, acute mood predicted trait anxiety scores, such that higher acute negative mood predicted higher trait anxiety scores and higher acute positive mood predicted lower trait anxiety scores. Past research experience also predicted trait anxiety scores, such that individuals who had past experience participating in research studies were more likely to have higher trait anxiety scores.

This study is a first explorative step in terms of determining how separate populations differ in individual differences and how extraneous variables have the ability to affect trait measures as quantified by scales that can be administered digitally. The fact that we did not find differences in trait anxiety between sampling frames can be interpreted as a sign of the potentual for external validity of the Trait Anxiety Inventory. For example, if we had found that University of Chicago community members had significantly lower trait anxiety levels, it would be difficult for our lab's research to be considered external valid outside of the University of Chicago community. Instead, this finding present hopefuly usability of digital communities, such as Amazon MTurk. Although the two sampling frames may differ in terms of

References

Barnes, Laura LB, Diane Harp, and Woo Sik Jung. 2002. "Reliability Generalization of Scores on the Spielberger State-Trait Anxiety Inventory." *Educational and Psychological Measurement* 62 (4). Sage Publications Sage CA: Thousand Oaks, CA: 603–18.

Bates, John A, and Brian A Lanza. 2013. "Conducting Psychology Student Research via the Mechanical Turk Crowdsourcing Service." *North American Journal of Psychology* 15 (2). North American Journal of Psychology: 385.

Behrend, Tara S, David J Sharek, Adam W Meade, and Eric N Wiebe. 2011. "The Viability of Crowdsourcing for Survey Research." *Behavior Research Methods* 43 (3). Springer: 800.

Berinsky, Adam J, Gregory A Huber, and Gabriel S Lenz. 2012. "Evaluating Online Labor Markets for Experimental Research: Amazon. Com's Mechanical Turk." *Political Analysis* 20 (3). Oxford University Press: 351–68.

Bieling, Peter J, Martin M Antony, and Richard P Swinson. 1998. "The State–Trait Anxiety Inventory, Trait Version: Structure and Content Re-Examined." *Behaviour Research and Therapy* 36 (7-8). Elsevier: 777–88.

Buhrmester, Michael, Tracy Kwang, and Samuel D Gosling. 2011. "Amazon's Mechanical Turk: A New Source of Inexpensive, yet High-Quality, Data?" *Perspectives on Psychological Science* 6 (1). Sage Publications Sage CA: Los Angeles, CA: 3–5.

Casey, Logan S, Jesse Chandler, Adam Seth Levine, Andrew Proctor, and Dara Z Strolovitch. 2017. "Intertemporal Differences Among Mturk Workers: Time-Based Sample Variations and Implications for Online Data Collection." *SAGE Open* 7 (2). SAGE Publications Sage CA: Los Angeles, CA: 2158244017712774.

Casler, Krista, Lydia Bickel, and Elizabeth 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). Elsevier: 2156–60.

Chapman, C Richard, and Gary B Cox. 1977. "Determinants of Anxiety in Elective Surgery Patients." *Stress and Anxiety* 4. John Wiley & Sons New York: 269–90.

Clifford, Scott, Ryan M Jewell, and Philip D Waggoner. 2015. "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics* 2 (4). SAGE Publications Sage UK: London, England: 2053168015622072.

Goodman, Joseph K, Cynthia E Cryder, and Amar Cheema. 2013. "Data Collection in a Flat World: The Strengths and Weaknesses of Mechanical Turk Samples." *Journal of Behavioral Decision Making* 26 (3). Wiley Online Library: 213–24.

Hauser, David J, and Norbert Schwarz. 2016. "Attentive Turkers: MTurk Participants Perform Better on Online Attention Checks Than Do Subject Pool Participants." *Behavior Research Methods* 48 (1). Springer: 400–407.

Huff, Connor, and Dustin Tingley. 2015. "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of Mturk Survey Respondents." *Research & Politics* 2 (3). SAGE Publications Sage UK: London, England: 2053168015604648.

Jones, Dan. 2010. "A Weird View of Human Nature Skews Psychologists' Studies." American Association for the Advancement of Science.

Murthy, Dhiraj. 2008. "Digital Ethnography: An Examination of the Use of New Technologies for Social Research." *Sociology* 42 (5). Sage Publications Sage UK: London, England: 837–55.

Paolacci, Gabriele, and Jesse Chandler. 2014. "Inside the Turk: Understanding Mechanical Turk as a Participant Pool." *Current Directions in Psychological Science* 23 (3). Sage Publications Sage CA: Los

Angeles, CA: 184–88.

Rand, David G. 2012. "The Promise of Mechanical Turk: How Online Labor Markets Can Help Theorists Run Behavioral Experiments." *Journal of Theoretical Biology* 299. Elsevier: 172–79.

Ross, Joel, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. "Who Are the Crowdworkers?: Shifting Demographics in Mechanical Turk." In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 2863–72. ACM.

Smith, Nicholas A, Isaac E Sabat, Larry R Martinez, Kayla Weaver, and Shi Xu. 2015. "A Convenient Solution: Using Mturk to Sample from Hard-to-Reach Populations." *Industrial and Organizational Psychology* 8 (2). Cambridge University Press: 220–28.

Spielberger, Charles D. 1966. "Theory and Research on Anxiety." *Anxiety and Behavior* 1 (3).

———. 2010. *State-Trait Anxiety Inventory*. Wiley Online Library.

Spielberger, Charles Donald. 1989. *State-Trait Anxiety Inventory: A Comprehensive Bibliography*. Consulting Psychologists Press.

Woo, Sang Eun, Melissa Keith, and Meghan A Thornton. 2015. "Amazon Mechanical Turk for Industrial and Organizational Psychology: Advantages, Challenges, and Practical Recommendations." *Industrial and Organizational Psychology* 8 (2). Cambridge University Press: 171–79.