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

4/23/2018

Research Questions:

- Do distributions of trait anxiety scores differ in samples acquired from a University of Chicago community vs. the Amazon Mechanial Turk community?
- 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?

Literature Review

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, train 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 to 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-stnading, 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 state 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 of which states 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. 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 to 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, trait personality measures.

Digital surveys and digital ethnography methods are seen as new technologies for social research that allow scientests to avoid more costly research methods, to easily alter questionnaires to access different cultural groups, to access more hard to reach populations, to collect higher response dates, 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 to interact with demographics of our sample and therefore affects the trait anxiety scores of our participants. If certain factors environmentally outside of a controlled laboratory could affect trait anxiety score, then we experience a trade off of 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 research, while benefitting from the same aspects discussed above. In fact, despite the concerns above, some researchers have argued that survey data that is digital 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 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 digital populations with in-person samples have found differences that may be less beneficial, and that even if both samples are collected digitally, 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). Although digital population

Clearly then, the results of this line of research has 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 digital populations, where data can be crowdsourced or collected in a completely digital way. Conducting psychology research using crowdsourced data has mainly revolved around the Amazon Mechanical Turk (MTurk) platform, based on its popularity and ease of access. MTurk provides a platform to oursource 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 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 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 incredible crucial 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 dicussion regarding the methodology of MTurk sampling and the MTurk population, as its promise of accessing high quality, inexpensive data is incredibly important to many lines of research (Buhrmester, Kwang, and Gosling 2011).

Based on this literature and our equivalent 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.

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.