5-2014

An evaluation of alternative methods for online survey recruitment: the equivalency of data collected through amazon mechanical turk and second life to traditional undergraduate student samples

Elizabeth M. Briones
University of Texas-Pan American

AN EVALUATION OF ALTERNATIVE METHODS FOR ONLINE SURVEY RECRUITMENT: THE EQUIVALENCY OF DATA COLLECTED THROUGH AMAZON MECHANICAL TURK AND SECOND LIFE TO TRADITIONAL UNDERGRADUATE STUDENT SAMPLES

A Thesis

by

ELIZABETH M. BRIONES

Submitted to the Graduate School of
The University of Texas - Pan American
In partial fulfillment of the requirements for the degree of

MASTER OF ARTS

May 2014

Major Subject: Experimental Psychology

AN EVALUATION OF ALTERNATIVE METHODS FOR ONLINE SURVEY RECRUITMENT: THE EQUIVALENCY OF DATA COLLECTED THROUGH AMAZON MECHANICAL TURK AND SECOND LIFE TO TRADITIONAL UNDERGRADUATE STUDENT SAMPLES

A Thesis
by
ELIZABETH M. BRIONES

COMMITTEE MEMBERS

Dr. Grant Benham
Chair of Committee

Dr. Darrin Rogers
Committee Member

Dr. Mark Winkel
Committee Member

May 2014

ABSTRACT

Briones, Elizabeth M., An Evaluation of Alternative Methods for Online Survey Recruitment: The Equivalency of Data Collected Through Amazon Mechanical Turk and Second Life to Traditional Undergraduate Student Samples. Master of Arts (MA), May, 2014, 57 pp., 2 tables, 10 figures, 75 references, 46 titles.

A major proportion of psychological research uses subjects from undergraduate populations. Increasingly, researchers are exploiting the Internet to reach beyond the traditional undergraduate sample. The current study sought to compare data obtained from a conventional undergraduate student sample to data collected via two online survey recruitment platforms: Amazon's Mechanical Turk (MTurk) crowdsourcing site and the virtual environment of Second Life (SL). Data obtained via these online recruitment platforms was statistically equivalent to the data obtained from the college sample, based on standardized measures of psychological stress and sleep quality. Additionally, correlations between the sleep and stress measures were not statistically different between the groups. These results, along with practical considerations in the use of these platforms, are discussed.

DEDICATION

For my parents, for believing in me and providing me with unconditional love and support.

For my husband Fred, you have always provided me with undying love, endless encouragement, continuous assistance, and helped me keep my sanity though this trying time. You are my rock and I love you.

ACKNOWLEDGEMENTS

I would like to personally acknowledge the following people for their personal contributions towards the completion of my thesis and endowing me with motivation to delve into the study of psychology. I thank Dr. Grant Benham for his unquestionable position in assisting me in such unfathomable lengths. He provided me with continuous support, guidance, and always replied to my emails. Together we forged ahead into uncharted territories of learning and came out more knowledgeable and for that I am forever grateful. I thank Dr. Darrin Rogers for providing a keen eye of concern with meticulous detail and a critical assessment of my work. I thank Dr. Winkel for his essential critique and invaluable feedback.

I would like to further thank these other people who contributed in some way to my fruitful pursuit in completing my thesis. Mark Bell at Indiana University Bloomington for giving me his time and acquainting me on Second Life scripting code. Jason Beckstead, without your statistical tool my analyses might not have been possible. For the Center for Online Learning, Teaching and Technology, you have been my home away from home for the past six years. And lastly, my husband Fred, for policing my grammar and never refusing my requests to proofread.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Whether qualitative or quantitative, survey methods have an established track record in the history of psychological research. Over time, the available mechanisms for conducting survey research have expanded from traditional paper-and-pencil and face-to-face interview approaches to telephone surveys, computer-administered surveys and, more recently, online (web-based) surveys. The relatively high cost of survey administration, including incentives for participation, has resulted in a high reliance on samples restricted to undergraduate psychology students (Peterson, 2010). Questions about the generalizability of the obtained data, coupled with technological advances and the explosion of the World Wide Web in the mid-1990s, have increasingly led researchers to explore the more efficient and cost effective option of online survey administration; a method that provides opportunities for access to more heterogeneous populations. In concert with these new methods of survey administration, novel approaches to recruitment of these online populations have also been developed.

Deviation from common practices requires evidence of validity. Early research was criticized, in part, because of justifiable concerns of sample bias (Buchanan & Smith, 1999; Kraut et al., 2004): only individuals with a computer and internet access could readily participate, a situation that skewed samples toward subjects with higher socioeconomic status. Though this potential bias may still exist, it is much reduced. Decreased hardware costs, increased connection speeds, and changing norms, have dramatically expanded the number of

individuals with online access. In 2012, there were over two billion Internet users worldwide ("Internet Users," 2012) and online surveys have gained a foothold as a valid and reliable method for survey data collection (Gosling, Vazire, Srivastava, & John, 2004).

Beyond questions of validity, researchers have argued that these online methods offer distinct advantages over traditional methods (Horton, Rand, & Zeckhauser, 2011; Lewis, Watson, & White, 2009; Ollesch, Heineken, & Schulte, 2006). The cost and time associated with the recruitment of large samples, removal of observer's presence phenomenon (Kraut et al., 2004), the ability to access unique groups of people that could otherwise be inaccessible, as well as a gateway to the ever increasing internet population (Wright, 2005) are just some of the purported benefits of online surveys over conventional methods. Consequently, the next step in the progression of Internet research is the analysis of sampling techniques used to recruit such populations for online investigations.

In addition to specialized online survey services, such as Qualtrics (Qualtrics, Provo, UT, 2013) and SurveyMonkey (SurveyMonkey, Palo Alto, CA, 2013), novel alternatives have been proposed for survey administration as well as sample recruitment or access to diverse populations at low costs. In 2005, Amazon established a "crowdsourcing" site named Mechanical Turk (MTurk; "Amazon Mechanical Turk," 2013) that allows users to solicit for, or participate in, a variety of "*Human Intelligence Tasks*" ("*HITs*"). These paid activities can be as diverse as identifying key words on a web page for a business to performing specified duties outlined by a researcher. Academic faculty have begun to take notice; a number of researchers are turning to this cost-effective method of survey recruitment and administration to mass audiences.

In 2003 Linden Lab established a "virtual world" called Second Life (SL; "Second Life," 2013). This online platform, populated with buildings, objects, and virtual representations of users (avatars), provided yet another enticing route through which researchers could access potential research participants. Second Life's membership has grown to a current level of approximately 50,000 daily users. Though some research has been conducted through SL, the studies have been restricted to social presence and engagement, educational utility, and business incorporation; limited research to date has investigated SL as a viable option for online sample recruitment.

Given the potential of both SL and MTurk to serve as low-cost survey recruitment modes that connect with diverse populations, it is important to establish whether the data collected from these platforms is comparable to results collected through more common practices (online surveys administered to undergraduate psychology students).

CHAPTER II

REVIEW OF THE LITERATURE

Online Survey Research

Collecting data from online respondents might offer some advantages over the use of traditional undergraduate psychology subject pools. For those who have limited access to large numbers of undergraduate students, e.g., faculty and student researchers at small institutions, access to a large pool of diverse participants may only be attainable through mail, telephone, or online survey approaches. Additionally, some researchers have critiqued the frequent use of college student samples in social-behavioral research, arguing that these samples do not accurately represent the broader population; that reliance on the "narrow data base" (Sears, 1986, p. 515) that is American college students (Henry, 2008) produces a skewed view of the world.

One relatively inexpensive solution to this dilemma is the use of the Internet to reach broader populations (Gosling, Sandy, John, & Potter, 2010). Although limitations exist with any method, internet survey research may provide a mechanism through which the generalizability of studies can be increased (Gosling et al., 2004). A recent study found that of 564,502 participants that completed an online personality questionnaire, 19% were not from advanced economies; 20% were from non-Western societies; 35% of the Western-society sample were not from the United States; and 66% of the U. S. sample were not in the 18–22 (college) age group (Grohol, 2010). Taken as a whole, data collected via the online survey method does not differ

significantly from values obtained through traditional survey administration methods. For example, a direct comparison of online survey data tied to health-related messages was shown to be statistically equivalent to data obtained from paper-and-pencil surveys (Lewis et al., 2009). Similarly, no differences were found between online and paper-and-pencil administration of social anxiety and social phobia measures based on data from an undergraduate college population (Hirai, Vernon, Clum, & Skidmore, 2011).

Broadening one's reach can also be achieved through telephone survey methods. However, online methods appear to provide comparable results at low cost. In a study examining mistaken beliefs about memory, data collected using Amazon Mechanical Turk (MTurk) was compared to data obtained from telephone surveys (Simons & Chabris, 2012). Having weighted each sample to match the United States 2010 Census demographics, inaccurate memory beliefs reported by the MTurk sample closely matched the telephone sample. Beyond this basic equivalency, online recruitment and survey administration methods have also demonstrated some distinct advantages, such as addressing the limitation of overrepresented elderly and underrepresented younger populations commonly found in telephone surveys (Simons & Chabris, 2012).

Online survey approaches are not without limitations. Although online surveys are *faster* than the postal method, mail surveys have generally been shown to generate higher response rates (Kwak & Radler, 2002). A finer-grained analysis of this response rate issue shows some subtleties, however. In an analysis of thirty-nine published studies that directly compared online and mail surveys, considerable variations in response rates were found between the two survey administration modes, with population type being the largest contributing factor of variability. Though online survey response rates were lower than mail survey response rates amongst

professionals, they were higher in studies based on college student populations (Shih & Fan, 2008). Variances in response rates have also been noted in recruitment mode differences. While online survey panels typically have more experience and regularity taking online surveys, one-time respondents generally show a decrease in response rates (Manfreda, Berzelak, Vehovar, Bosnjak, & Haas, 2008). Given the rapid evolution of technological approaches to online recruitment and survey administration, these findings will need to be frequently reexamined and recommendations reevaluated.

Beyond issues of equivalency and validity, researchers must also weigh issues of practicality when deciding which survey methods to employ. A mixed-mode approach could offer a balanced method and remove certain limitations that come with choosing one method over another, but time concerns and expense must also be considered (Kaplowitz, Hadlock, & Levine, 2004; Potoglou, Kanaroglou, & Robinson, 2012). The population of interest and the extent that one needs to generalize findings will likely determine the mode best suited for a study. The time and manpower required for manual data entry for mail or paper surveys are a distinct disadvantage in relation to the full automation that comes with online data collection, not to mention the associated costs of printing and mailing (Medlin, Roy, & Chai, 1999). Online surveys, accessible to anyone with an internet connection and appropriate hardware (e.g., smartphone, tablet, computer), open the door to a broader range of recruitment methods. Literature examining the impact of alternative recruitment methods is severely limited, however.

The Internet provides several mechanisms through which to recruit participants for psychological research, allowing scientists to connect with populations that might not otherwise be readily accessible. Some have explored the use of social media sites such as Facebook and search engine advertising (Google, Bing, Yahoo!) as a means of Web-based recruitment (Fenner

et al., 2012; Morgan, Jorm, & Mackinnon, 2013; Samuels & Zucco, 2013), but one of the most popular emerging methods for online research recruitment has been the breakthrough in *crowdsourcing* sites such as crowdSPRING, Innocentive, oDesk, and Amazon Mechanical Turk. An overview of one of these crowdsourcing sites is provided in the next section, along with an introduction to the realm of Linden Lab's virtual world: *Second Life*.

Emerging Online Research Methods

Amazon's Mechanical Turk

Crowdsourcing is the ability to outsource services or tasks to large groups of people, specifically online communities (Crowdsourcing, n.d.). Amazon's Mechanical Turk (*MTurk*) is one of the leading crowdsourcing services, introduced in 2005 as a way to get people to accurately identify duplicate pages on its website for very low costs (Pontin, 2007): computer algorithms and automated software were simply not advanced enough to perform these tasks alone. Almost a decade later, MTurk is still in its beta version. However, businesses and individuals alike now use the site to contract out duties that might otherwise be too costly to be freelanced though traditional employment. The MTurk's requester site allows requesters to custom design tasks that must be completed by MTurk workers – tasks referred to as *human intelligence tasks* (*HITs*). Beyond simply setting up tasks, a requester can select specific worker qualifications that must be met in order to accept the listed task. This can range from the percentage of approval rates, specific geographic locations, or customized tests a worker must pass before accepting a HIT.

Perhaps fueled by the high demand for online survey data collection, MTurk has taken this a step further: including a turnkey survey tool option for creating HITs. However, if investigators are to feel confident decreasing their dependence on traditional recruitment

methods, then empirical analysis of the validity of these evolving technologies must be undertaken. A small number of recent studies have supported the utility of MTurk as a mechanism for online research studies (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Crump, McDonnell, & Gureckis, 2013; Goodman, Cryder, & Cheema, 2012; Paolacci, Chandler, & Ipeirotis, 2010; Simons & Chabris, 2012). Additionally, MTurk's workers have been found to be significantly more diverse than undergraduate samples (Behrend, Sharek, Meade, & Wiebe, 2011; Buhrmester et al., 2011). In 2010, roughly half of MTurk workers were from the United States. Although 54% were found to be between the ages of 21-35 years old, these numbers widen the age gap by almost ten years when compared to the traditional college aged student that is between the ages of 18-24 years old (Ipeirotis, 2010). It is not uncommon for MTurk workers to be paid $0.10 USD for completing small tasks and research suggests that these small payment incentives do not affect data quality (Buhrmester et al., 2011; Marge, Banerjee, & Rudnicky, 2010; Mason & Watts, 2009).

Linden Lab's Second Life

Second Life (SL) has been endorsed as one of the leading applications for immersive technology. Launched in 2003 as an MMORPG or massively multiplayer online role-playing game, millions of people around the world now use SL as a platform for various activities: from business to leisure to education. SL's residents create digital representations of themselves (avatars), which can be individually personalized and custom designed or purchased from the consumer marketplace. Once a user is 'in-world' he or she can engage in a number of different activities that range from social gatherings, art exhibits, religious organizations, or corporate collaborations. Some users, or residents, choose to cultivate businesses, which exist on SL's own system of e-commerce based on the Linden dollar (L$). Others purchase virtual real estate, even

private islands, which afford a stage for seemingly endless architectural possibilities. Universities have invested staff and budgets toward creating and supporting online virtual classrooms and learning environments in SL in which instructors and students can harness the tools of the virtual environment for pedagogical activities.

The potential for research using virtual environments is not a new concept. In 1997, the National Science Foundation supported NetLab, a workshop to explore new mediums for research, specifically social interactions using large-scale Internet experiments (Bainbridge, 2007). Technological advancements in e-learning have enhanced the way in which learning has been made available in recent years. A review of educational research using SL found the potential for promoting role-playing activities and simulation of games when SL was used as an educational tool. While arduous learning curves and possible distractions were found as potential problems, SL was also found to provide an environment to build communication, social participation, and collective interaction (Inman, Wright, & Hartman, 2010). Various opportunities have been found for research using virtual worlds in a wide variety of disciplines such as business, political science, and neurobiology (Bray & Konsynski, 2007).

Bell, Castronova, and Wagner (2008) have made major contributions and developments in virtual world survey data collection methods (Bell, Castronova, & Wagner, 2009). One such advancement introduced a new way to collect survey data in SL without having to leave the virtual user interface. Before their development of the Virtual Data Collection Interface (VDCI), an individual would need to leave the SL interface and open an internet browser to take part in a given survey (Bell et al., 2008). This interrupted the user's virtual experience and possibly altered attitudes and perceptions that were trying to be assessed (Bell et al., 2009). The VDCI was the first of its kind to keep the user fully immersed in the virtual environment as well as

provide a way to collect survey data. However, as with most advancing technology, SL has found a way to incorporate web-based media content without having to leave the virtual space. It is now possible to display a web-based survey using online survey software directly to an individual without virtual interruption. This method was used to examine a new platform for qualitative research and sample recruitment of special populations (Dipko, Billington, & Brick, 2012). While further research was recommended, the authors found little evidence to suggest effects of mode between a survey taken via SL and alternative methods, though same sample respondents were used for all methods. Although there is no established standard, the average amount paid to Second Life participants for survey research has generally been around L$250 Linden dollars, which is approximately equivalent to $1.00 USD.

Several studies have explored methodological considerations of research performed in SL, but few have examined the utility of SL as a platform for survey data collection in relation to established approaches (Anstadt, Bradley, & Burnette, 2013; Bell, 2008; Bell et al., 2009; Dean, Cook, Murphy, & Keating, 2012; Martey & Shiflett, 2012). Though the virtual world initially began as a tool for online social interaction and communication, the focus has shifted to the creative possibilities SL offers. This transformation provides a new online audience for alternatives in method survey administration and offers another avenue for sample recruitment for Web-based surveys. Given the limited amount of studies on SL as a research tool, further examination is needed to support it as an alternative method for online survey recruitment.

Research Questions and Hypotheses

Researchers are increasingly turning to non-traditional recruitment methods for social behavioral research, including the MTurk and SL platforms. However, there is limited research

validating these methods. The purpose of this study was to compare a relatively novel online survey recruitment method (SL) and a more established method (MTurk) with a traditional college undergraduate approach. Specifically, we aimed to assess the influence of recruitment methods on the data obtained, using established measures of self-perceived psychological stress (Perceived Stress Scale [PSS]; Cohen, Kamarck, & Mermelstein, 1983) and self-reported sleep quality (Pittsburgh Quality Sleep Index [PSQI]; Buysse, Reynolds, Monk, Berman, & Kupfer, 1989).

RQ 1: Are perceived stress scores (PSS) obtained via SL equivalent to those obtained from college students?

RQ 2: Are perceived stress scores (PSS) obtained via MTurk equivalent to those obtained from college students?

RQ 3: Are sleep quality scores (PSQI) obtained via SL equivalent to those obtained from college students?

RQ 4: Are sleep quality scores (PSQI) obtained via MTurk equivalent to those obtained from college students?

RQ 5: Does the strength of the relationship between sleep and stress differ significantly between SL participants and college students?

RQ 6: Does the strength of the relationship between sleep and stress differ significantly between MTurk participants and college students?

Hypothesis 1: Perceived stress scores of SL participants will be statistically equivalent to the college students' stress scores.

Hypothesis 2: Perceived stress scores of MTurk participants will be statistically equivalent to the college students' stress scores.

Hypothesis 3: Sleep quality scores of SL participants will be statistically equivalent to the college students' sleep quality scores.

Hypothesis 4: Sleep quality scores of MTurk participants will be statistically equivalent to the college students' sleep quality scores.

Hypothesis 5: The strength of the association between sleep quality and stress in SL participants will not differ significantly from that observed in college students.

Hypothesis 6: The strength of the association between sleep quality and stress in MTurk participants will not differ significantly from that observed in college student.

CHAPTER III

METHODOLOGY

Participants

Participants included the following three groups:

Three hundred and sixty three college students at the University of Texas – Pan American who had completed an online survey in 2013 that contained measures of stress and sleep quality. College student participants ranged in age from 18 to 50 years old (M = 22.9, SD = 4.9), 79% were female, and 93% described themselves as Hispanic.

Two hundred individuals recruited through Amazon Mechanical Turk. These participants ranged in age from 18 to 72 (M = 36.4, SD = 13.7), 62% were female, and 10% identified themselves as Hispanic. Of the 213 surveys started, 200 respondents completed the MTurk survey producing a response rate of 94%.

Sixty-seven respondents[1] recruited through Second Life. These participants ranged in age from 18 to 61 (M = 29.1, SD = 11.2), were 55 % female, and 30% Hispanic. The response rate for the SL survey was 87%.

[1] Though 68 respondents completed the survey, one extreme outlier, reported age 94, was omitted from analyses.

Measures

Perceived Stress Scale (PSS)

The PSS is used in measuring the perception of one's stress. The PSS is a 10-item Likert-type scale that asks respondents 'In the last week, how often have you . . .' and includes items such as 'felt nervous and stressed?', 'felt that you were unable to control the important things in your life?' The 10-item version of the scale is a revision of the originally published 14-item version, has been shown to provide a slight gain in psychometric quality over the longer version, and is recommended over the 14-item version by the scale's authors (Cohen & Williamson, 1988). The PSS has been reported as a better predictor of psychological symptoms, physical symptoms, and health service utilization than life-event scales (Cohen et al., 1983). Possible scores range from 0 to 40 and were calculated by summing up the 10-item ratings (after reverse scoring specific items). Higher PSS scores represent more stress. Good internal reliability has been reported for the PSS, Cronbach's $\alpha = .91$.

Pittsburgh Sleep Quality Index (PSQI)

The PSQI consists of 19 questions and provides a global measure of sleep quality. The global PSQI score is based on seven components: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction over the last month, each of which is weighted equally on a 0–3 scale. Scores on the PSQI range from 0 to 21 with higher scores indicating worse sleep quality. A global score greater than 5 provides a sensitive and specific measure that distinguishes poor from good sleepers (Buysse et al., 1989) and has thus been established as a standardized cutoff score for the PSQI. The PSQI has been reported as having good internal consistency with a reliability coefficient of .83 (Cronbach's alpha) with its seven components.

Demographics

A number of demographic variables were collected including age, gender, ethnicity, and education level. Though we were requiring MTurk participants to be located in the United States, there is no reliable way to set location restrictions for the Second Life users and thus a location question was asked of the SL participants.

Procedures

College Student Recruitment Procedure

College student data was extracted from an existing dataset, based on an online survey study conducted in 2013. Students had been recruited on a voluntary basis through in-class announcements and recruitment flyers posted on professors' learning management systems (BlackBoard, Washington, DC) in exchange for extra credit.

MTurk Recruitment Procedure

MTurk participants were recruited using the website's default survey method, available only to those whom were eligible to view the listing. Eligibility to view the HIT or task listing consisted of a worker having a 95% approval rating and located in the United States; those who did not meet the requirements were not aware of its publication. For the eligible MTurk participants, a listing titled "5-minute Academic Survey" appeared in their list of available HITs. Figure 1 presents a graphic of the task listings or HITs a user would see when they log into MTurk. When they clicked on the link for the current study, a page with a short description of the survey, the survey link, and a verification code box was shown (see Figure 2). If the worker chose to accept the HIT, the user was redirected to an external webpage for the survey.

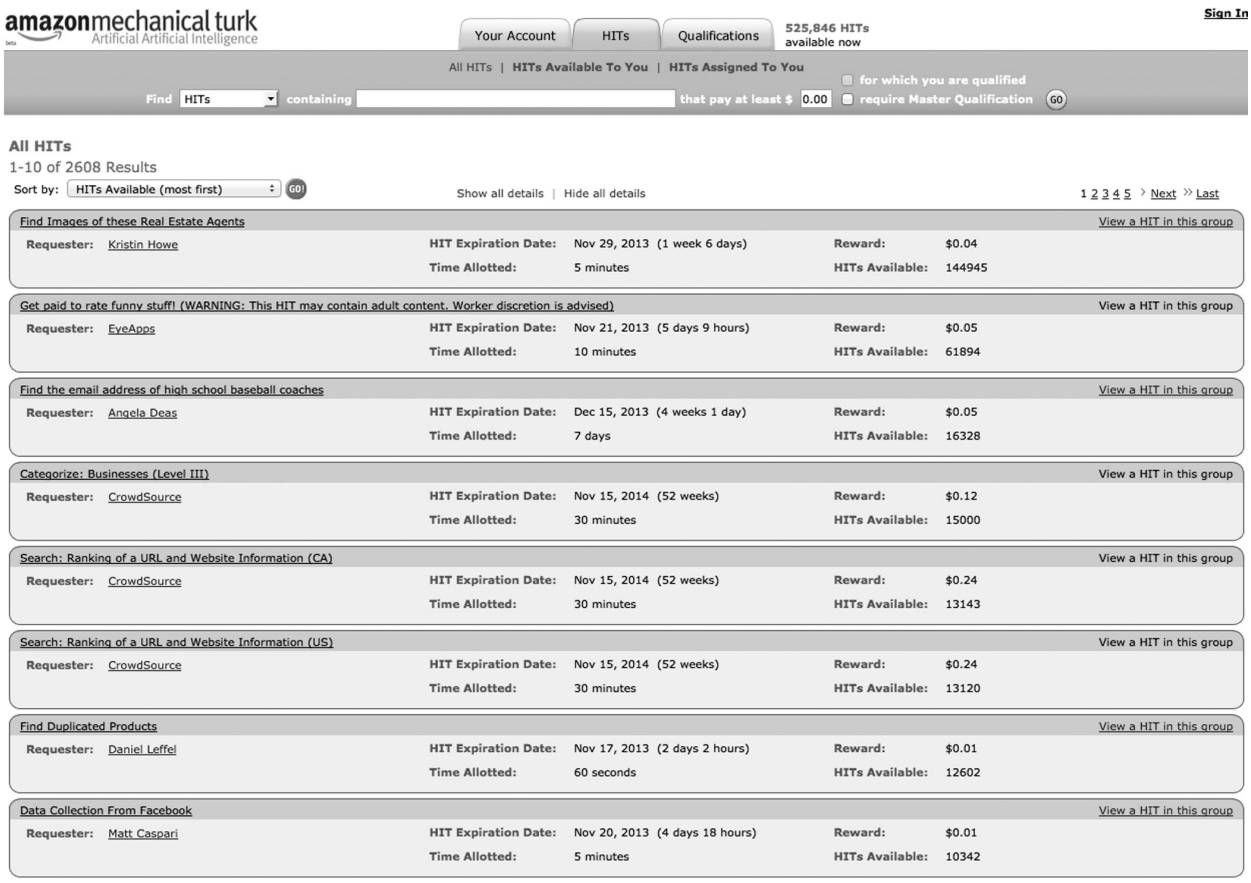

Figure 1. MTurk Task Listings (HITs)

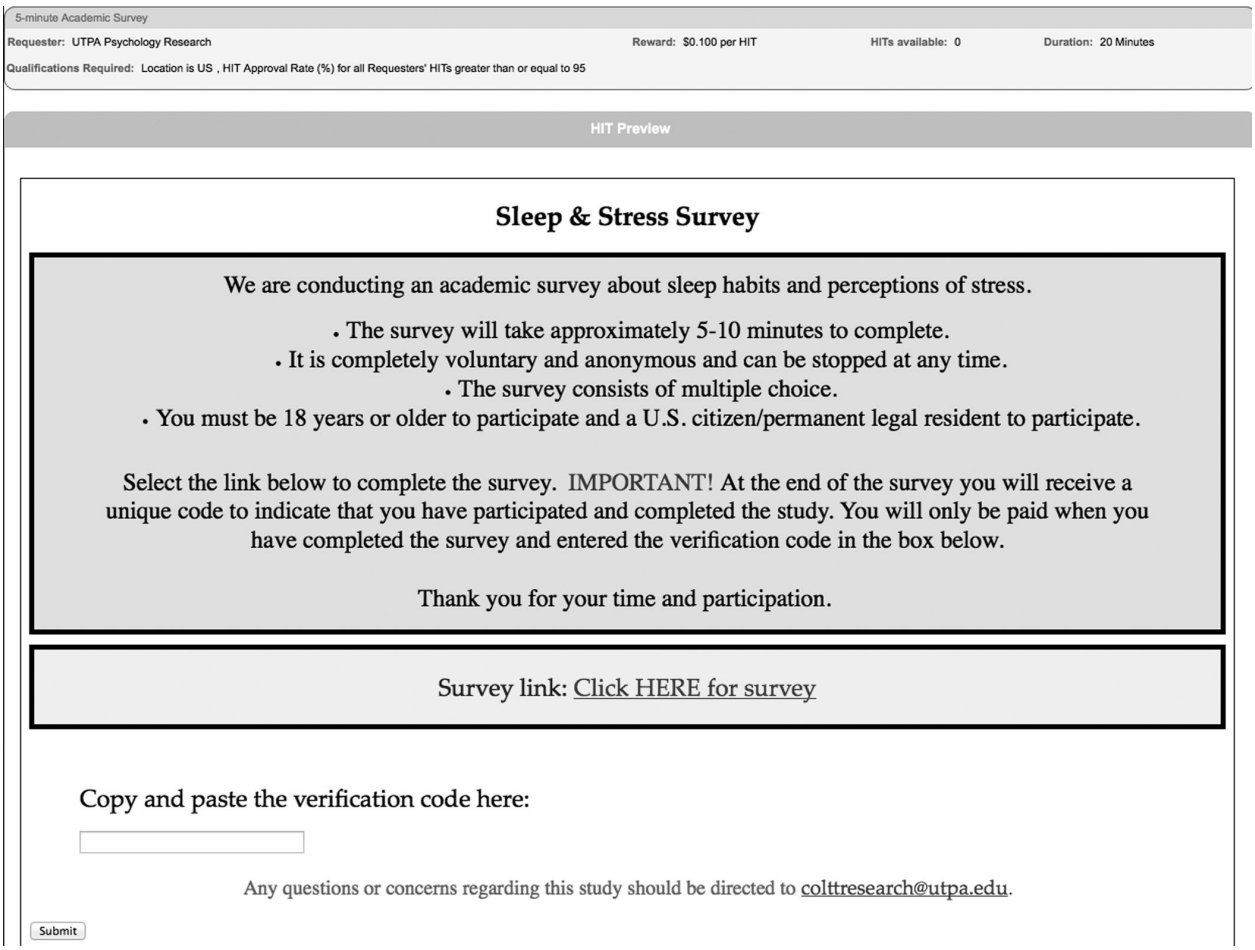

Figure 2. MTurk Survey HIT Page

SL Recruitment Procedure

Second Life participants were recruited through an advertisement published on Second Life's classified advertisement platform. This service is available to all SL users and provides information for resources and services from other residents. Promotion of research opportunities through the classified advertisements have been shown to be most effective when compared to other modes of recruitment in Second Life (Bell et al., 2009; Dean et al., 2012). This can be accessed through the game's user interface or through Second Life's website. The classified contained information and a description of the survey, similar to the MTurk's description. It also contained a direct link or SLurl to the location where the survey could be obtained 'in-world'.

Participation in the survey included the user's avatar "teleporting" to the University of Texas-Pan American SL's island location where a kiosk was set up. A folder with instructions, a copy of the informed consent, along with the *heads up display* (*HUD*) was automatically given to the user when a box labeled "CLICK ME for survey" was touched. The HUD was worn on the user's avatar and occupied a large portion of the user's screen. This HUD presented the survey as it would have been seen in an external web browser. This method provided added privacy; once the user attached it, it was invisible to other residents (see Figure 3).

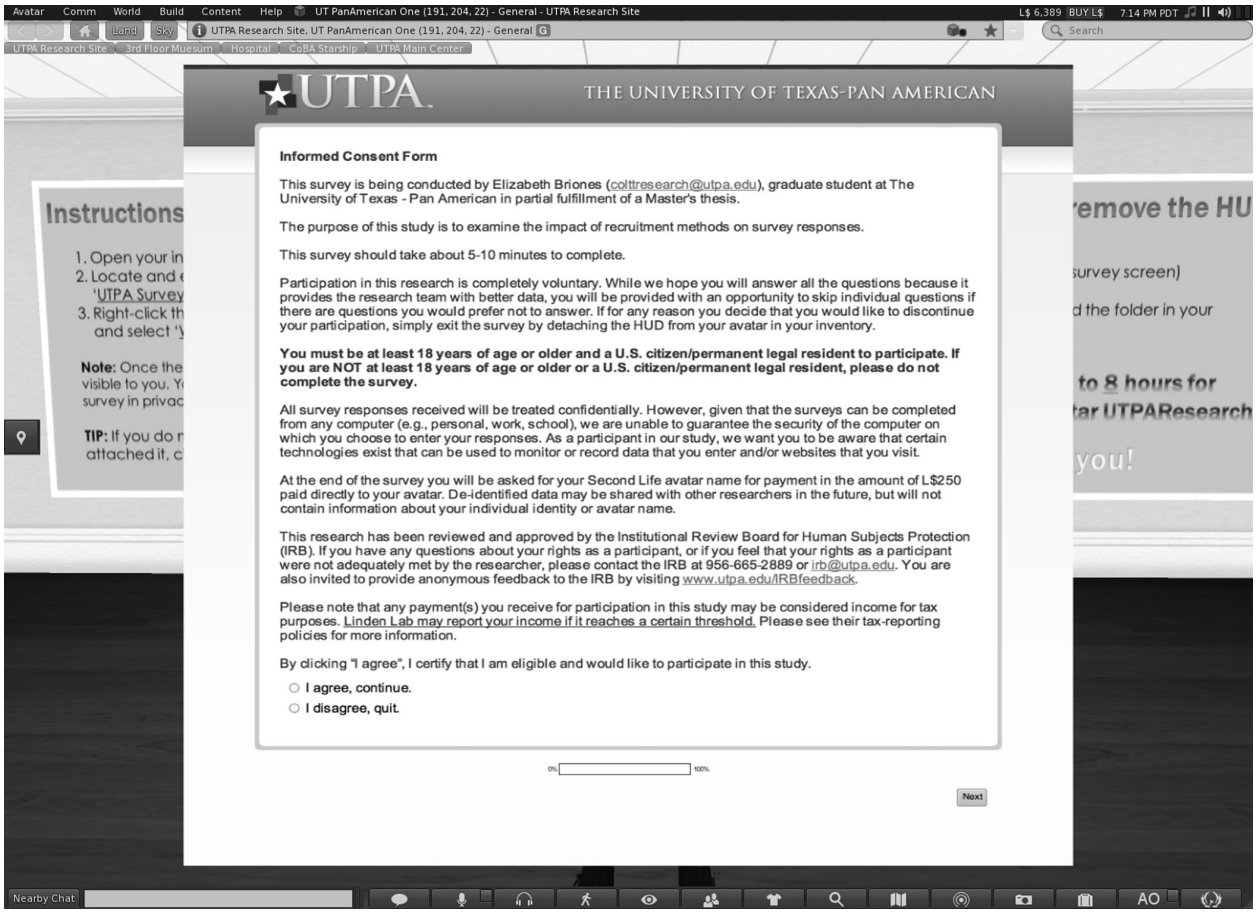

Figure 3. Second Life participant's view of online survey via attached HUD.

Survey Administration

All participants were directed to complete the online survey, created and hosted through Qualtrics (Qualtrics, Provo, UT). The survey included demographic questions, an Internet usage

question, the Perceived Stress Scale, and the Pittsburgh Sleep Quality Index (Buysse et al., 1989; Cohen et al., 1983). Respondent requirements consisted of a minimum age of 18 years and be a US citizen or legal resident for participation.

The MTurk participant was presented with a consent form, which the user agreed to before proceeding with the survey. Once the participant completed the survey, a verification code and instructions were offered for entry into the verification code box at the MTurk HIT page for credit. The MTurk participants were credited $0.10 USD for completion of the survey.

For SL participants, the HUD provided the survey just as the MTurk user viewed the survey with the exception that participation was performed exclusively in Second Life's interface. Both MTurk and SL surveys were identical with the exception of two questions added to the SL survey version: (i) respondents were asked if they resided either inside or outside the United States and (ii) the last survey question asked for the user's SL avatar name for payment purposes. Instructions were also provided at the end of the survey for detachment of the HUD from the avatar. To maintain consistency with MTurk, each SL participant was paid L$22 (approximately $0.10 USD) for completion of the survey. After two weeks, due to low response rate, an announcement was place on the Second Life's Facebook page advertising the survey and the rate was increased to L$250 (approximately $1.00 USD).

CHAPTER IV

RESULTS

Analyses Overview

The aim of this study was to compare the data obtained from a conventional undergraduate college student sample to that obtained via two online recruitment platforms[2]. The default approach to group comparisons in social science research relies on null hypothesis significance testing (NHST). This approach assesses whether groups *differ* to a statistically significant degree, with accepted levels of risk for inaccurate rejection of a null (no difference) hypothesis based on an a-priori established alpha value (e.g., .05). However, the failure to demonstrate a significant difference between groups does not directly demonstrate *equivalence* between groups, though it is frequently misinterpreted and misrepresented as such (Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers & Howard, 1993; Tryon, 2001; Walker & Nowacki, 2010). Failure to reject the null hypothesis, is not the same as accepting the null hypothesis. Traditional NHST is based on assuming that the null (no difference) is true. Equivalence testing takes a different approach and assumes that the groups differ by a predetermined amount. Thus, the purpose of our study was not to demonstrate that obtained means were smaller or larger than

[2] Though an argument might be made for the creation of matched samples, based on demographic criteria such as age and sex, this may be more of an issue for those wishing to demonstrate significant differences between groups. The purpose of our study was to evaluate whether the obtained data, in their unadulterated form, were statistically equivalent. Thus, our decision not to match provided a more conservative test of this equivalency with greater external validity.

those obtained from a traditional college sample, but to examine whether we obtained similar (equivalent) results when using these non-traditional methods of recruitment.

To test whether two groups are equivalent (null hypothesis that the groups differ), social science researchers have begun to borrow from the field of biostatistics. Frequently, for example, medical researchers must demonstrate that a new procedure or drug is as good as (equivalent to) the standard care, a question best answered through equivalence testing (Epstein, Klinkenberg, Wiley, & McKinley, 2001; Lewis et al., 2009; Weigold, Weigold & Russell, 2013). A number of methods have been developed to test for statistical equivalence, including the two one-sided test procedure (TOST; Schuirmann, 1987). For our study, we elected to base our determination of equivalence on *inferential* confidence intervals (ICIs; Tryon, 2001; Tryon & Lewis, 2008) because it provides a visual representation of the equivalency of groups whilst simultaneously evaluating statistical significance. Analyses were conducted using an Excel spreadsheet developed by Jason Beckstead (2008) that incorporates Tryon's ICI test of equivalence (Tryon, 2001; Tryon & Lewis, 2008). Sample output from this equivalence test is shown in Figure 4 and explained below.

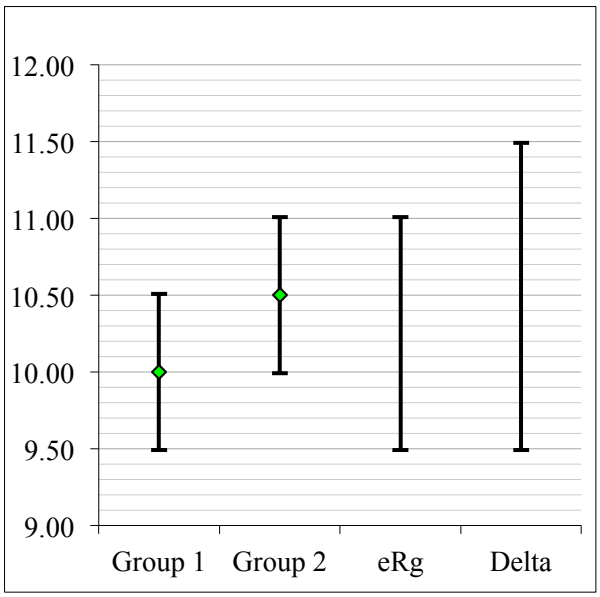

Figure 4. Sample Equivalence Test

Rather than relying on standard descriptive confidence intervals, Tryon's procedure shortens these confidence intervals such that nonoverlapping inferential confidence intervals (ICIs) are algebraically equivalent to a null hypothesis significance test (NHST) between two means. In Figure 4, the 95% ICIs of Group 1 and Group 2 are presented. The first step in equivalency testing is to establish the Delta (Δ) that will be used: an a-priori criterion for how far the two groups can differ while still being considered equivalent. Given the lack of prior research on which to base this Delta value, we elected to follow the criterion used by a number of other researchers, setting Delta to +/- 20% (Cribbie et al., 2004; Epstein et al., 2001; Lewis et al., 2009; Rusticus & Lovato, 2011; Rogers & Howard, 1993; Steele, Mummery, & Dwyer, 2009; Weigold et al., 2013). This Delta is used to create an equivalence interval around the reference group (e.g., college students). In the example represented in Figure 4, the mean of the reference group is 10.0. Delta is set at 2 (20%) and therefore extends from the lower bound of Group 1's inferential CI (9.5) to a value of 11.5 (9.5 + 2 = 11.5). A confidence interval (eRg) is then

created from the range of the two group's inferential confidence intervals: extending from the lower CI limit of the lesser mean (Group 1) to the upper CI of the greater mean (Group 2).

Statistical difference is said to exist between the two means if the ICIs of Group 1 and Group 2 do not overlap. *Statistical equivalence* is said to exist when the equivalence range (eRg) provided by the ICIs is less than the minimum inconsequential difference (Delta), i.e., when eRg is contained within Delta. *Statistical indeterminacy* is said to exist when the means are neither statistically different nor equivalent (Beckstead, 2008; Tryon, 2001; Tryon & Lewis, 2008). In Figure 4, the ICIs of Group 1 and Group 2 overlap, indicating no statistical difference (traditional NHST approach). Additionally, the confidence interval range (eRg) fits within the 20% Delta interval chosen, indicating that the two groups are statistically equivalent. This test of equivalence was applied to our measures of stress (PSS) and sleep quality (PSQI) to examine whether the means obtained from our MTurk and SL samples were equivalent to the means obtained from traditional college student samples. Given that our Delta values were based on a rule-of-thumb criterion, we also determined the Delta value for which equivalence would no longer hold (i.e., the Delta value that would result in a lack of equivalency between the two groups).

In addition to comparisons of measurement means, we wanted to examine whether the correlation between stress and sleep quality differed between samples. Because of a lack of developed tests of equivalence testing for correlational comparisons, we were constrained to traditional approaches. We thus used Fisher's z transformations of each group's stress-sleep correlation to assess whether these associations were statistically significantly different from one another.

Given the small number of a-priori tests being conducted, and the debate of the appropriateness of such methods (Feise, 2002; Nakagawa, 2004; Rothman, 1990; Perneger, 1998), we elected not to correct for multiple comparisons. An alpha level of .05 was used for all statistical analyses.

Demographics

Table 1 provides a summary of sample characteristics for each group according to age, gender, ethnicity, and education. The MTurk group was generally more diverse in age than the college and SL groups. Additionally, the SL group showed a more evenly distributed ratio of males to females than the college group. The college sample was recruited from a university in a predominantly Hispanic region and was consequently overrepresented by that ethnicity.

Table 1

Demographic Sample Percentages for Characteristics by Group

		College group (N=363)	MTurk group (N=200)	SL Group (N=67)
		%	%	%
Age				
	18 - 25	82.4	23.5	52.2
	26 - 35	13.8	37	23.9
	36 - 45	3.3	14	10.4
	46 - 55	0.6	12	10.4
	Above 55	0	13.5	3
	No response	0	0	0
Gender				
	Male	21.5	38	44.8
	Female	78.5	62	55.2
	No response	0	0	0
Ethnicity				
	Caucasian	1.9	69.5	46.3
	Hispanic	92.3	10	29.9
	African American	1.4	10	6
	Asian	2.2	6	1.5
	Other	2	4.5	16.4
	No response	0.3	0	0
Education				
	Less than high school		0	4.5
	High school/GED		7.5	34.3
	Some college		31.5	29.9
	2-year degree		10	13.4
	4-year degree		36.5	13.4
	Master's degree		12	3
	Doctoral degree		0.5	1.5
	Professional degree		2	0
	No response		0	0

Note: Education was not assessed for the college group.

Equivalence Tests

Tests of statistical equivalence and statistical difference were conducted to examine whether MTurk and SL participants' stress and sleep quality scores were the same as those obtained from college students. Inferential confidence interval (ICI) statistics, equivalence determinations, the critical Delta value for equivalence no longer holds, and statistical difference determinations are summarized in Table 2 for each comparison.

Table 2
ICIs and Equivalence Testing Results

		95% ICIs	Statistical equivalence at Δ = 20%	Critical Delta	Statistical difference
PSS					
	CO	[19.60, 20.52]	YES	11%	NO
	MT	[18.37, 19.86]			
	CO	[19.60, 20.52]	YES	9%	NO
	SL	[18.64, 20.22]			
PSQI					
	CO	[6.68, 7.19]	YES	9%	NO
	MT	[6.62, 7.36]			
	CO	[6.66, 7.21]	YES	18%	NO
	SL	[6.39, 7.94]			

Note: CO = college group, MT = MTurk group, SL = Second Life group;
95% ICIs = inferential confidence intervals [lower bound, upper bound];
Critical Delta = values of Delta for which equivalence no longer holds.

RQ 1: Are perceived stress scores (PSS) obtained via SL equivalent to those obtained from college students? ICIs were compared for the mean PSS score of the college group (N = 363), *M* = 20.06, *SD* = 6.16, 95% CI[3] [19.42, 20.69], and the mean PSS score for the SL group (N = 67), *M* = 19.43, *SD* = 4.46, 95% CI [18.35, 20.52]. Using a Δ = 4.01 (20.06 * .2), we found statistical equivalence and no statistical difference between the groups. Equivalence testing results are graphically depicted in Figure 5.

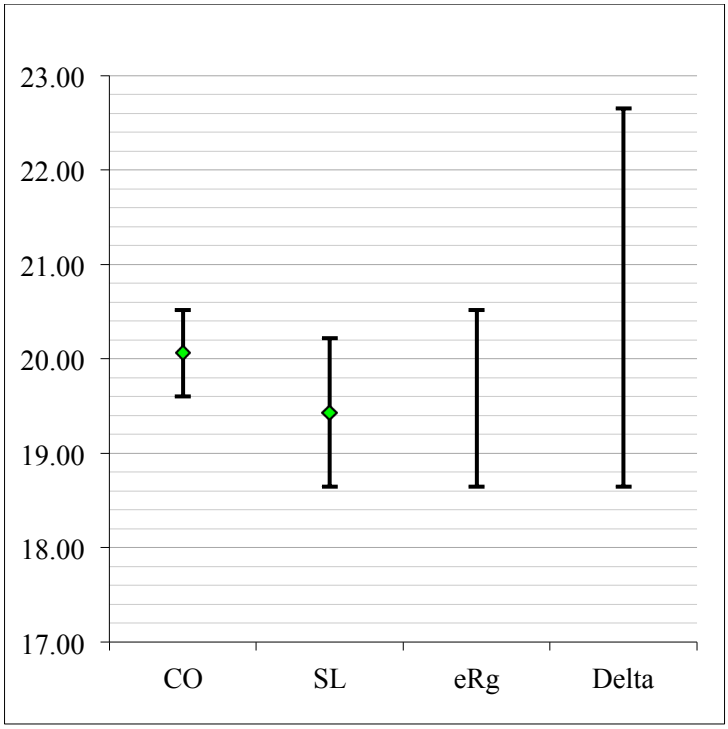

Figure 5. College and SL PSS Comparison
95% ICIs for college and SL groups, eRg = 1.88, Δ = 4.01

RQ 2: Are perceived stress scores (PSS) obtained via MTurk equivalent to those obtained from college students? For our second research question, we compared ICIs of the mean PSS score of the college group, with the mean PSS score of the MTurk group (N = 200), *M* = 19.11,

[3] Descriptive confidence intervals should not be confused with inferential confidence intervals (ICIs) that were used in statistical equivalence analyses.

SD = 7.35, 95% CI [18.09, 20.13]. As shown in Figure 6, with a Delta set at 4.01, we found statistical equivalence and no statistical difference between the groups.

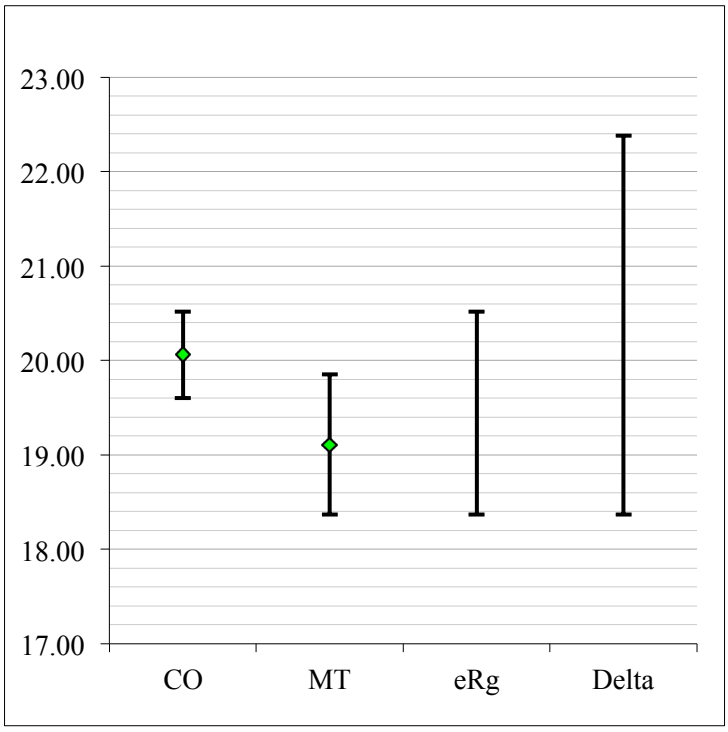

Figure 6. College and MTurk PSS Comparison
95% ICIs for college and MTurk groups, eRg = 2.15, Δ = 4.01

RQ 3: Are sleep quality scores (PSQI) obtained via SL equivalent to those obtained from college students? The third research question attempted to assess equivalence for the sleep quality scores between the college and SL samples. The Delta was computed at +/-20% of the college group's mean PSQI score, Δ = 1.39. Results indicate the college group (N = 361), *M* = 6.94, *SD* = 3.48, 95% CI [6.58, 7.30] was statistically equivalent and not statistically different from the SL group (N = 64), *M* = 7.17, *SD* = 4.01, 95% CI [6.17, 8.17]. Figure 7 presents the ICIs comparison.

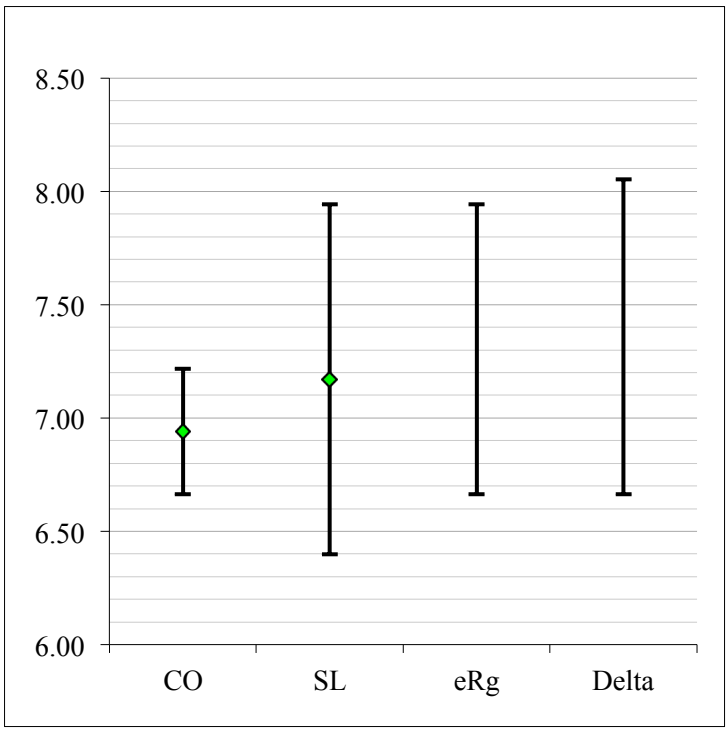

Figure 7: College and SL PSQI Comparison
95% ICIs for college and SL groups, eRg = 1.28, Δ = 1.39

RQ 4: Are sleep quality scores (PSQI) obtained via MTurk equivalent to those obtained from college students? ICIs were compared for the college group to the mean sleep score for the MTurk group (N = 198), $M = 6.99$, $SD = 3.67$, 95% CI [6.50, 7.48], and a Delta set at 1.39. As hypothesized, we also found statistical equivalence and no statistical difference between the college group and the MTurk group. Graphical representations of the ICIs are presented in Figure 8.

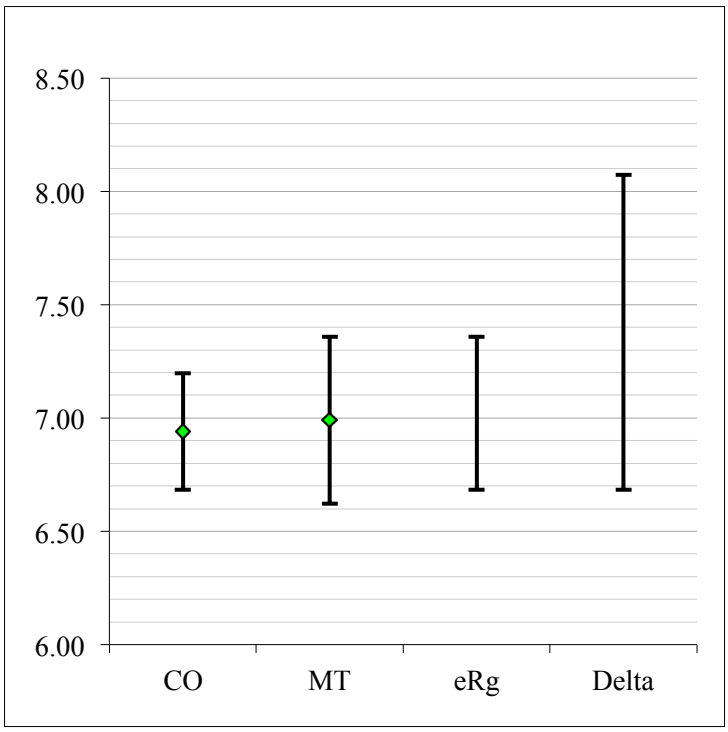

Figure 8. College and MTurk PSQI Comparison
95% ICIs for college and MTurk groups, eRg = 0.68, Δ = 1.39

Correlational Comparisons

RQ 5: Does the strength of the relationship between sleep and stress differ significantly between SL participants and college students? To assess the differences of the correlations between samples for PSS and PSQI scores, we followed the methods outlined in Weigold et al. (2013) adopted from Preckel and Thiemann (2003) who used Fisher's *r*-to-*z* transformation. Sleep and stress were significantly correlated in college students, $r(361) = .45$, $p < .001$, and SL participants, $r(64) = .39$, $p = .002$. The difference between these correlations was not statistically significant, $z = .54$, $p = .59$.

RQ 6: Does the strength of the relationship between sleep and stress differ significantly between MTurk participants and college students? Sleep and stress were significantly correlated

in college students, $r(361) = .45$, $p < .001$, and MTurk participants, $r(198) = .47$, $p < .001$. The difference between these correlations was not statistically significant, $z = .24$, $p = .81$.

Supplemental Analysis

The main focus of the study was to determine if data obtained from two groups recruited via emerging online platforms were equivalent to a traditional sample recruited from the college classroom. As a supplemental analysis, we wanted to see how the two groups from the online platforms compared to each other. Three additional analyses were performed to assess equivalence. The first compared the stress scores of the MTurk sample to those of the SL sample. The second test compared the same groups' sleep quality scores. With Deltas chosen at +/-20% of the MTurk group's mean stress score ($\Delta = 3.82$) and +/-20% of the MTurk PSQI mean score ($\Delta = 1.40$), and $\alpha = .05$, we found both tests to be statistically equivalent and not statistically different. Equivalence testing results including ICIs, eRgs, and Deltas are presented in Figures 9 and 10.

We also compared the relationship of stress and sleep for the MTurk group with the SL group. Sleep and stress were significantly correlated in the MTurk group, $r(198) = .47$, $p < .001$, and SL participants, $r(64) = .39$, $p = .002$. The difference between these correlations was not statistically significant, $z = .66$, $p = .51$.

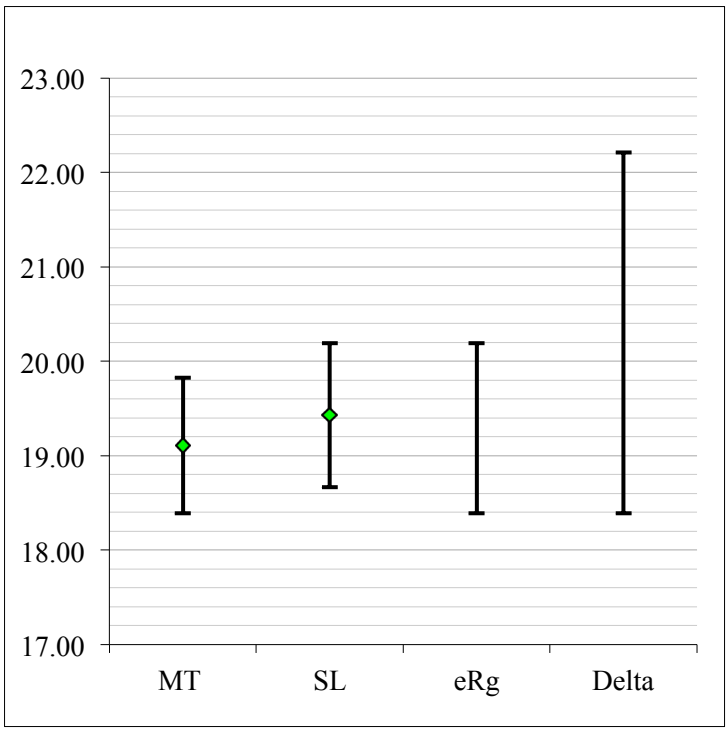

Figure 9. MTurk and SL PSS Comparison
Note: MTurk 95% ICI [18.39, 19.83], SL 95% ICI [18.67, 20.19], eRg = 1.80, Δ = 3.82

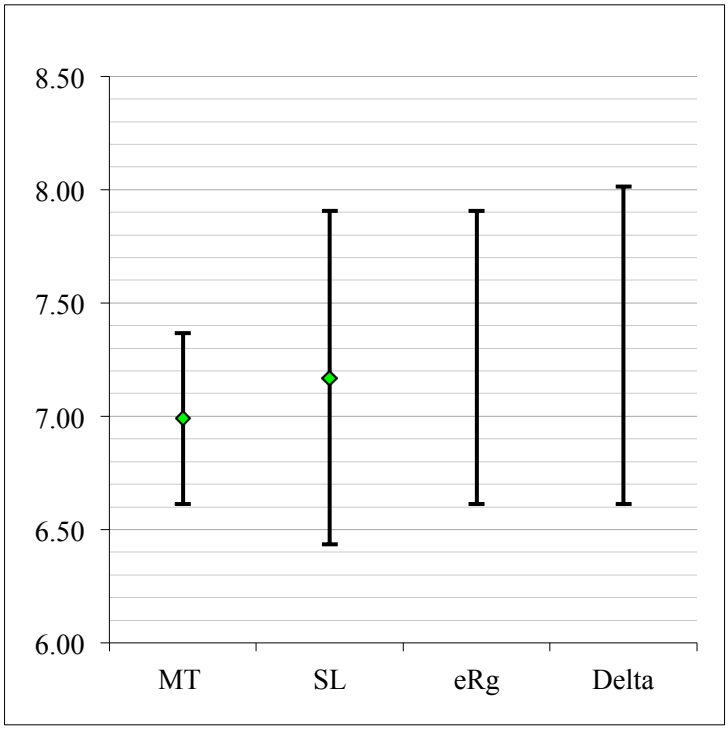

Figure 10. MTurk and SL PSQI Comparison
Note: MTurk 95% ICI [6.61, 7.37], SL 95% ICI [6.44, 7.91], eRg = 1.29, Δ = 1.40

CHAPTER V

DISCUSSION

The aim of this research was to examine the viability of using two existing online platforms (MTurk and SL) for online survey recruitment by comparing the resultant data to that collected via more conventional means (recruitment from undergraduate student classes). We believe that this is the only study that has simultaneously compared these two platforms to a traditional college recruitment method. Based on the means of two standardized measures, and on the observed correlations between those two measures, our results suggest that the data collected from these two online platforms is statistically equivalent to that obtained from the college student sample. These findings suggest that such methods might serve as an alternative to conventional classroom recruitment.

The Internet provides researchers with an alternative to the college student population for psychological study recruitment, but heavy reliance on college undergraduates for psychological research has not decreased in recent years despite these expanding technological resources (Gallander Wintre, North, & Sugar, 2001). The Internet provides both a platform for administration of online surveys (Couper, & Miller, 2008; Kraut et al., 2004; Riva, Teruzzi, & Anolli, 2003) and also as a tool for acquiring research participants. As tools for survey administration and participant recruitment are developed, however, researchers must be careful to evaluate the validity of these novel approaches. This is most commonly achieved by comparing new methods to more established administration procedures (e.g., pencil-and-paper)

and more traditional samples (e.g., undergraduate students). The present study examined whether data obtained from two emerging online recruitment platforms was equivalent to that obtained from a conventional college sample.

Based on the demographic data collected from our MTurk and SL groups, these samples appeared to show greater diversity than our college sample in terms of age, ethnicity, and education and a more balanced proportion of males to females. While Caucasians were overrepresented in the MTurk sample, the distribution of ethnicity/race was more diverse than our particular college student sample. Education was not directly assessed in the college sample, but the MTurk sample provided an assortment of education levels that were less homogeneous than the standard undergraduate subject pool. Taken together, the demographic data suggests that MTurk provides a more diverse sample than found with undergraduate students, steering researchers away from some of the concerns associated with lack of generalizability. Using a Delta value of 20%, these non-traditional samples were statistically equivalent to college students on measures of both stress and sleep quality. Additionally, a comparison of the correlation between these measures demonstrated no significant differences between any of the groups. As predicted, these findings support the notion of equivalence across recruitment platforms.

This study adds to the growing body of literature demonstrating the validity of MTurk as a viable alternative to traditional undergraduate pools for survey recruitment (Behrend et al., 2011; Buhrmester et al., 2011; Goodman et al., 2012; Paolacci et al., 2010). Our research replicates and extends previous studies of MTurk by using equivalence testing to compare the data obtained with traditional (undergraduate) samples; a statistical approach that we believe is more appropriate than the previously applied NHST procedures. We selected two widely-used

standardized measures as a basis for comparison: the Perceived Stress Scale (PSS; Cohen et al., 1983) and the Pittsburgh Sleep Quality Index (PSQI; Buysse, 1989). Based on these measures, MTurk produces data that is statistically equivalent to that obtained from college students.

Our findings also add to the limited research conducted on the use of SL as a research recruitment method. SL was not developed for research, but social scientists soon recognized its utility as a platform to study human interaction. Initial studies focused on how and where to recruit in SL, but few sought to compare it against an established method. To date, no study has applied equivalence testing to evaluate SL as a valid survey recruitment mode. While SL appeared to provide a more diverse sample than the college sample, as well as supportive evidence of equivalence for stress and sleep quality measures, data collection through this platform created some practical challenges.

MTurk versus Second Life: Practical Considerations

The continuing use of undergraduate students in psychological research is driven, in part, by practical considerations such as ease of access and low cost. To establish viability of these non-conventional approaches, their practicability must therefore be evaluated alongside issues of statistical equivalency.

Initial Setup

For researchers who are unfamiliar with the MTurk and SL platforms, the proposition of learning a new system can be a substantial hurdle that negates the transition from conventional classroom recruitment. Amazon's MTurk, though still rather underutilized by social science researchers, provides a relatively intuitive and user-friendly interface. The system has been designed to provide a mechanism through which tasks can be advertised to workers by requesters

and therefore recruitment for research is a natural fit with the system's existing structure. Second Life was designed as a 3D virtual world where users could interact with each other in real-time, primarily as a means of social networking. Though researchers soon recognized the value of SL as a venue for conducting research, such projects required independent development of methods and tools for recruitment and administration. While some have found the virtual environment fun and intriguing, others have declared it a labor-intensive task. Thus, SL can involve a steep learning curve for researchers, particularly for those who are not already familiar with the nature of the platform. While the use of SL may show some promise, researchers should carefully consider its practicality in relation to their own technical abilities.

Participant Payments and Research Costs

One of the reasons for the continued popularity of undergraduate subject pools may be their low financial cost. In the majority of universities, students participate in research either as part of a course requirement (e.g., Introductory Psychology Participant Pools) or for extra credit in college courses. Therefore, the financial cost of alternative methods may be an important consideration when evaluating practical viability. In the current study, we elected to use a fairly common rate for short MTurk tasks: a payment of $0.10. To maintain consistency across the two platforms, we offered an equivalent $0.10 payment for SL participants (converted to the SL currency of L$22 Linden Dollars). Due to a low response rate from SL participants, the SL payment was increased to L$250 (approximately $1.00 in U.S. currency) after 2 weeks. The payment process in MTurk is fully automated; participants simply entered a validation code that was presented at the end of the research survey in order to receive their payment from the researcher's MTurk account. Second Life does not have a comparable system for payments and therefore the process for paying SL participants was considerably more involved. Second Life

participants included their (unique) SL avatar name as part of the online survey. The Qualtrics survey responses were periodically checked for completed surveys and avatar names were recorded, then the appropriate Linden Dollar amount was transferred to the SL user's account using this avatar name. Though some studies have developed sophisticated computer coded scripts to better automate this task, ensuring reliable techniques requires advanced knowledge of SL's own scripting language.

Recruitment Success Rate and Survey Response Time

Online surveys provide a rapid means to collect data, contingent upon sufficient participation. Obtaining reasonable sample sizes using conventional college classroom recruitment may vary depending on the nature of the university and established policies within the department. For small colleges, or those with limited participation incentives, online recruitment may provide a reasonably cost-effective approach to obtaining adequate sample sizes. Based on the short survey used in the current study, our targeted sample size of 200 was obtained through MTurk within approximately two and a half weeks. Within that same period, only 27 individuals from SL had completed surveys.

Even after attempts to increase response rates by raising the payment and posting an advertisement on the SL Facebook page, we were unable to reach our SL sample size goal within a reasonable timeframe. After five and a half weeks, the SL sample size had only reached 68. Our study did not exhaust all recruitment methods within SL and future research may establish more effective approaches to recruitment within this platform, but such issues are beyond the scope of our current investigation.

Screening Respondents and Reducing Multiple Submissions

MTurk provides a system for screening workers based on various criteria and displaying listings of pre-qualified HITs to the individual workers. As such, MTurk provides a level of quality control (e.g., worker must meet a specific approval rating in order to view the research HIT) and flexibility in limiting samples based on specific demographics (e.g., must be a U.S. citizen). Second Life does not offer a similar function; screening criteria must be established through recruitment announcements and survey questions and relies entirely on the honesty of the participant.

The Qualtrics online survey platform includes its own tool for reducing "ballot stuffing" (restricting people from taking the survey more than once). MTurk also has a utility that, if used, prohibits multiple responses to a HIT, further protecting against repeat users. Second Life provides no added protection, thus increasing the possibility of multiple responses from a single participant. Second Life allows for multiple avatar accounts to be made with one email address, making it possible for one person to take a survey over multiple avatar accounts. Thus, while multiple responses are a limitation for all online surveys, MTurk offers an additional feature to combat this weakness.

Institutional Review Board Restrictions

Though policies may vary from institution to institution, human subjects research approval by Institutional Review Boards (IRBs) may create additional challenges. Online research faces a unique challenge in that the federal regulations surrounding human subjects protection were developed before such technologies existed. Thus, IRBs must develop their own policies regarding online research, policies that may be inconsistent across institutions. Our original study design did not place limits on the nationality of participants, but institutional IRB

policy regarding the withholding of taxes for payments made to non-U.S. citizens/nationals and the need to collect additional personal information from such individuals, even for $0.10 payments or payments in virtual currency, necessitated our setting more restrictive criteria for inclusion. As a result, our MTurk and SL samples were less diverse than we had originally intended.

Limitations and Future Directions

The current study is the first to simultaneously compare data obtained through MTurk and SL recruitment with data obtained through traditional classroom recruitment methods and, as such, provides valuable information about the utility of these platforms for research. Our study suffers from limitations common across many online survey studies: such as volunteer bias and social desirability bias. Additionally, online surveys can only provide information given by unverified respondents and thus quality of data depends entirely on the participant; we cannot assume that all information disclosed is credible (Duda & Nobile, 2010). However, the equivalency of the datasets suggests that this problem is not greater in MTurk or SL samples relative to conventional undergraduate samples.

Our study involved a small number of measures and therefore could be completed in a relatively short time: on average it took MTurk and SL participants approximately five minutes to complete. Whether or not the observed equivalency holds for longer or more complex studies would need to be determined through further research. Additional research will also help to determine whether the *nature* of the measures has any impact: our studies made use of relatively benign scales of self-perceived stress and sleep quality and it remains to be seen whether equivalency would extend to more sensitive measures such as abuse or depression.

The number of SL respondents that we obtained was lower than our original goal, in spite of advertising through additional venues and increasing the payment for participation. While the sample size may be seen as a limitation to the study, it also reflects the real-world difficulties associated with using SL as a recruitment medium. The low numbers tell a story of their own and suggest that SL may not be the best tool for recruitment. In relation to MTurk, SL is more complex to set up, more difficult to process payment, and does not generate as high a response rate. Second Life might offer some advantages over MTurk in other areas of research, given users' immersion in a 3D virtual environment, but its use as a simple recruitment tool may be limited.

Our study did not attempt to match samples based on demographic characteristics, opting to make comparisons based on the raw datasets. As such, the MTurk and SL samples contained a more diverse range of ages and a more balanced distribution of males and females than the undergraduate sample. It is important to recognize this lack of similarity between samples, but the fact that equivalencies were demonstrated *in spite of* these demographic differences provides stronger evidence for the utility of these non-conventional recruitment techniques. Additionally, given the range of ages represented on MTurk and the apparent ease of access to populations that differ from college undergraduates, MTurk may provide a reasonable route through which to obtain targeted populations such as middle-aged or elderly individuals.

Given the IRB requirements at our institution, our current MTurk sample was limited to MTurk workers who were U.S. residents or citizens and who, based on the demographics, appeared to be relatively well-educated and over-represented by Caucasians. Thus, some of the general concerns associated with lack of diversity in Western college student samples still hold. Our findings are limited in this regard, given that other researchers may have fewer restrictions

on the recruitment of non-U.S. nationals. It remains to be seen whether our findings would be replicated with a more diverse sample of worldwide participants.

Lastly, because methods to assess correlational equivalence have yet to be established, our analyses of obtained correlation coefficients was limited to more traditional NHST methods. As a result, the hypotheses being tested for our correlation comparisons do not fully equate to those being tested for our comparison of group means. However, the findings of equivalence and failure to find statistically significant differences appear to echo a consistent message.

Explaining Equivalency

Early concerns about sample bias due to individual restrictions to internet access (Buchanan & Smith, 1999; Kraut et al., 2004) may not be as relevant today, but MTurk and SL users may still be a biased sample given that they are self-selected (and arguably "*unusual*") groups. While equivalency of these two groups to college students suggests that online survey research may be a viable alternative to collecting data from undergraduate samples, the very fact that these groups did not differ may seem somewhat surprising. *Why* should these groups, comprised of people with different backgrounds, different distributions of sexes, and different age ranges, provide equivalent data? One reason may simply be that the measures used in our study are insensitive to such between-group differences; that the equivalencies demonstrated are nothing more than a reflection of poorly selected measures. However, we chose two measures that have been psychometrically validated and widely adopted and that show a good distribution of scores. They appear to be sensitive instruments. A second explanation may be that our chosen Delta was too generous. The determination of equivalency depends, in part, on the a-priori establishment of a reasonable Delta; a value that describes how far the two groups can differ

while still being considered equivalent. Our Delta was based on a recommended rule-of-thumb 20% value, but future research may benefit from more sophisticated determinations of Delta. As demonstrated in our analysis, reducing this value ultimately reaches a critical point at which equivalency fails to be shown. Thus, more conservative assessments based on smaller Deltas would have failed to demonstrate equivalency while simultaneously maintaining NHST findings of no statistically significant differences: groups would have been neither equivalent nor significantly different (a scenario of *statistical indeterminacy*; Tryon, 2001). Thirdly, it may be that these groups really are quite similar; that college students do not differ greatly from individuals who complete HITs on MTurk or who engage in SL.

Conclusions

In conclusion, data obtained from MTurk and SL samples appear to be statistically equivalent to that obtained from undergraduate samples. From a practical standpoint, MTurk may be a viable alternative for recruitment, particularly for those with limited access to college students, with low cost and a more diverse representation of demographics. The utility of Second Life may be more questionable, however, given the technical knowledge required, higher cost, and lower response rate. We hope that these findings, along with similar studies demonstrating MTurk's utility, foster further exploration of the platform as a tool for conducting survey research in psychology.

REFERENCES

Amazon Mechanical Turk (2013). *Amazon Mechanical Turk – welcome.* Retrieved June 12, 2013, from https://www.mturk.com/mturk/welcome

Anstadt, S., Bradley, S., & Burnette, A. (2013). Virtual worlds: In-world survey methodological considerations. *Journal of Technology in Human Services, 31*(2), 156-174. doi:10.1080/15228835.2013.784107

Bainbridge, W. S. (2007). The scientific research potential of virtual worlds. *Science, 317*(5837), 472-476. doi:10.1126/science.1146930

Barnes, S. J. (2011). Understanding use continuance in virtual worlds: Empirical test of a research model. *Information & Management, 48*(8), 313-319. doi:http://dx.doi.org/10.1016/j.im.2011.08.004

Beckstead, J. W. (2008). Inferential confidence intervals and null hypothesis testing. Retrieved April 1, 2014, from http://personal.health.usf.edu/jbeckste/ci.html

Behrend, T., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*(3), 800-813.

Bell, M. W. (2008). Toward a definition of "virtual worlds". *Journal of Virtual Worlds Research, 1*(1). doi:10.4101/jvwr.v1i1.283

Bell, M. W., Castronova, E., & Wagner, G. G. (2008). Virtual assisted self interviewing (VASI): An expansion of survey data collection methods to the virtual worlds by means of VDCI. *DIW Berlin, Research Notes of the German Council for Social and Economic Data (RatSWD), Data Documentation 37*.

Bell, M. W., Castronova, E., & Wagner, G. G. (2009). Surveying the virtual world: A large scale survey in Second Life using the virtual data collection interface (VDCI). *DIW Berlin, Research Notes of the German Council for Social and Economic Data (RatSWD), Data Documentation 44*.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." *Political Analysis*, doi:10.1093/pan/mpr057.

Bray, D. A., & Konsynski, B. R. (2007). Virtual worlds: Opportunities for multi-disciplinary research. *The DATA BASE for Advances in Information Systems, 38*(4), 17-25. doi:10.1145/1314234.1314239

Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, *90*(1), 125.

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

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index (PSQI): A new instrument for psychiatric research and practice. *Psychiatry Research, 28*(2), 193-213.

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of psychological stress. *Journal of Health and Social Behavior, 24*, 385–396.

Cohen, S., & Williamson, G. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan, & S. Oskamp (Eds), *The social psychology of health.* Newbury Park, CA: Sage.

Couper, M. P., & Miller, P. V. (2008). Web survey methods: Introduction. *Public Opinion Quarterly*, *72*(5), 831-835.

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*(1), 1-10.

Crowdsourcing. (n.d.). Merriam-Webster's online dictionary. Retrieved July 23, 2013, from http://www.merriam-webster.com/dictionary/crowdsourcing

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS ONE, 8*(3), 1-18. doi:10.1371/journal.pone.0057410

Dean, E., Cook, S., Murphy, J., & Keating, M. (2012). The effectiveness of survey recruitment methods in Second Life. *Social Science Computer Review, 30*(3), 324-338. doi:10.1177/0894439311410024

Dipko, S., Billington, C., & Brick, P. D. (2012). The effect of mode on participant responses to qualitative research in virtual worlds. Paper presented at the American Association for Public Opinion Research (AAPOR) 67th Annual Conference: *Evaluating New Frontiers in Public Opinion and Social Research.* Orlando, FL.

Duda, M., & Nobile, J. L. (2010). The fallacy of online surveys: No data are better than bad data. *Human Dimensions of Wildlife*, *15*(1), 55-64. doi:10.1080/10871200903244250

Epstein, J. J., Klinkenberg, W. D., Wiley, D. D., & McKinley, L. L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*(3), 339-346. doi:10.1016/S0747-5632(01)00002-4

Feise, R. J. (2002). Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology*, *2*, 1-4.

Fenner, Y., Garland, S., Moore, E., Jayasinghe, Y., Fletcher, A., Tabrizi, S., & ... Wark, J. (2012). Web-based recruiting for health research using a social networking site: an exploratory study. *Journal of Medical Internet Research*, *14*(1), e20. doi:10.2196/jmir.1978

Fetscherin, M., & Lattemann, C. (2008). User acceptance of virtual worlds. *Journal of Electronic Commerce Research, 9*(3), 231-242.

Gallander Wintre, M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology/Psychologie Canadienne*, *42*(3), 216-225. doi:10.1037/h008689

Goodman, J. K., Cryder, C. E., & Cheema, A. (2012). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213-224. doi: 10.1002/bdm.1753

Gosling, S. D., Sandy, C. J., John, O. P., & Potter, J. (2010). Wired but not WEIRD: The promise of the Internet in reaching more diverse samples [Peer commentary on the journal article "The weirdest people in the world?" by J. Henrich, S. J. Heine, A. Norenzayan]. *Behavioral and Brain Sciences, 33*(2-3), 61-135. doi:10.1017/S0140525X10000300

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

Grohol, J. M. (2010). Psychology secrets: Most psychology studies are college student biased. Retrieved from http://psychcentral.com/blog/archives/2010/08/26/psychology-secrets-most-psychology-studies-are-college-student-biased/

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry, 19*(2), 49-71. doi:10.1080/10478400802049936

Hirai, M., Vernon, L. L., Clum, G. A., & Skidmore, S. T. (2011). Psychometric properties and administration measurement invariance of social phobia symptom measures: Paper-pencil vs. internet administrations. *Journal of Psychopathology and Behavioral Assessment, 33*(4), 470-479. doi:10.1007/s10862-011-9257-2

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics, 14*(3), 399-425. doi:10.1007/s10683-011-9273-9

Inman, C., Wright, V. H., & Hartman, J. A. (2010). Use of Second Life in K-12 and higher education: A review of research. *Journal of Interactive Online Learning, 9*(1), 44-63.

Internet users in the world distribution by world regions. (2012). Internet World Stats - Usage and Population Statistics. Retrieved July 6, 2013, from http://www.internetworldstats.com/stats.htm

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk (Working Paper No. CEDER-10-01). NYU Working Paper Series. Retrieved from https://archive.nyu.edu/handle/2451/29585

Kaplowitz, M. D., Hadlock, T. D., & Levine, R. (2004). A Comparison of web and mail survey response rates. *Public Opinion Quarterly, 68*(1), 94-101. doi:10.1093/poq/nfh006

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *American Psychologist, 59*(2), 105-117. doi:10.1037/0003-066x.59.2.105

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile, and data quality. *Journal of Official Statistics, 18*(2), 257-273.

Lewis, I., Watson, B., & White, K. M. (2009). Internet versus paper-and-pencil survey methods in psychological experiments: Equivalence testing of participant responses to health-related messages. *Australian Journal of Psychology, 61*(2), 107-116. doi:10.1080/00049530802105865

Manfreda, K., Berzelak, J., Vehovar, V., Bosnjak, M., & Haas, I. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79-105.

Marge, M., Banerjee, S., Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In J. Hansen (Ed.), *Proceedings of the 2010 IEEE Conference on Acoustics, Speech, and Signal Processing* (pp. 5270-5273). Dallas, TX.

Martey, R. M., & Shiflett, K. (2012). Reconsidering site and self: Methodological frameworks for virtual-world research. *International Journal of Communication, 6*, 105-126.

Mason, W. A., & Watts, D. J. (2009). Financial incentives and the "performance of crowds". In *Proceedings of the HCOMP '09 ACM SIGKDD Workshop on Human Computation* (pp. 100-108). New York, NY.

Medlin, C. J., Roy, S., & Chai, T. H. (1999). World Wide Web versus mail surveys: A comparison and report. *Marketing in the third millennium: Proceedings of the Australian and New Zealand Marketing Academy Conference (ANZMAC)*. Sydney, New South Wales.

Morgan, A., Jorm, A., & Mackinnon, A. (2013). Internet-based recruitment to a depression prevention intervention: Lessons from the Mood Memos study. *Journal of Medical Internet Research*, *15*(2), 90-101. doi:10.2196/jmir.2262

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology*, *15*(6), 1044-1045. doi:10.1093/beheco/arh107

Ollesch, H., Heineken, E., & Schulte, F. P. (2006). Physical or virtual presence of the experimenter: Psychological online-experiments in different settings. *International Journal of Internet Science, 1*(1), 71-81.

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

Perneger, T. (1998). What's wrong with Bonferroni adjustments. *BMJ (Clinical Research Ed.)*, *316*(7139), 1236-1238.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*(3), 450-461.

Pontin, J. (2007). Artificial intelligence, with help from the humans. The New York Times. Retrieved August 4, 2013, from http://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html

Potoglou, D., Kanaroglou, P. S., & Robinson, N. (2012). Evidence on the comparison of telephone and Internet surveys for respondent recruitment. *Open Transportation Journal*, *6,* 11-22

Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil version of a high potential intelligence test. *Swiss Journal Of Psychology/Schweizerische Zeitschrift Für Psychologie/Revue Suisse De Psychologie*, *62*(2), 131-138. doi:10.1024//1421-0185.62.2.131

Qualtrics (2013). *Qualtrics*. Retrieved June 12, 2013, from http://www.qualtrics.com

Rogers, J. L., & Howard, K. I. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The Use of the Internet in Psychological Research: Comparison of Online and Offline Questionnaires. *Cyberpsychology & Behavior*, *6*(1), 73-80. doi:10.1089/109493103321167983

Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology (Cambridge, Mass.)*, *1*(1), 43-46.

Rusticus, S. A., & Lovato, C. Y. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation, 16*(7), 1-6.

Samuels, D. J., Zucco, C. (2013). Using Facebook as a subject recruitment tool for survey-experimental research. Retrieved October 16, 2013, from http://ssrn.com/abstract=2101458 or http://dx.doi.org/10.2139/ssrn.2101458

Schuirmann D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657-680. doi:http://dx.doi.org/10.1007/BF01068419

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology, 51*(3), 515-530. doi:http://dx.doi.org/10.1037/0022-3514.51.3.515

Second Life (2013). *Second Life.* Retrieved June 2, 2013, from http://secondlife.com/

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods, 20*(3), 249-271. doi:10.1177/1525822x08317085

Simons, D. J., & Chabris, C. F. (2012). Common (mis)beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PLoS ONE, 7*(12), 1-5. doi:10.1371/journal.pone.0051876

Steele, R. M., Mummery, W. K., & Dwyer, T. (2009). A Comparison of Face-to-Face or Internet-Delivered Physical Activity Intervention on Targeted Determinants. *Health Education & Behavior*, *36*(6), 1051-1064.

Stoddart, C. (2013). How brains see. Retrieved June 22, 2013, from https://www.youtube.com/watch?v=0oNHukxz5rI

SurveyMonkey (2013). *SurveyMonkey, Inc.* Retrieved June 12, 2013, from http://www.surveymonkey.com

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*(4), 371-386. doi:10.1037/1082-989X.6.4.371

Tryon, W. W., & Lewis, C. (2008). An inferential confidence interval method of establishing statistical equivalence that corrects Tryon's (2001) reduction factor. *Psychological Methods, 13*(3), 272-277.

Walker, E., & Nowacki, A. S. (2010). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine, 26*(2), 192-196.

Weigold, A., Weigold, I. K., & Russell, E. J. (2013). Examination of the equivalence of self-report survey-based paper-and-pencil and Internet data collection methods. *Psychological Methods, 18*(1), 53-70.

Wright, K. B. (2005). Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication, 10*(3), 00. doi:10.1111/j.1083-6101.2005.tb00259.x

APPENDIX A

APPENDIX A

PERCEIVED STRESS SCALE (PSS)

Perceived Stress Scale- 10 Item
The questions in this scale ask you about your feelings and thoughts during the last month. In each case, please indicate with a check how often you felt or thought a certain way.

1. In the last month, how often have you been upset because of something that happened unexpectedly?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

2. In the last month, how often have you felt that you were unable to control the important things in your life?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

3. In the last month, how often have you felt nervous and "stressed"?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

4. In the last month, how often have you felt confident about your ability to handle your personal problems?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

5. In the last month, how often have you felt that things were going your way?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

6. In the last month, how often have you found that you could not cope with all the things that you had to do?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

7. In the last month, how often have you been able to control irritations in your life?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

8. In the last month, how often have you felt that you were on top of things?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

9. In the last month, how often have you been angered because of things that were outside of your control?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

10. In the last month, how often have you felt difficulties were piling up so high that you could not overcome them?
___0=never ___1=almost never ___2=sometimes ___3=fairly often ___4=very often

APPENDIX B

APPENDIX B

PITTSBURGH SLEEP QUALITY INDEX (PSQI)

INSTRUCTIONS:
The following questions relate to your usual sleep habits during the past month only. Your answers should indicate the most accurate reply for the majority of days and nights in the past month.

Please answer all questions.

1. During the past month, what time have you usually gone to bed at night?
   BED TIME _____
2. During the past month, how long (in minutes) has it usually taken you to fall asleep each night?
   NUMBER OF MINUTES _____
3. During the past month, what time have you usually gotten up in the morning?
   GETTING UP TIME _____
4. During the past month, how many hours of actual sleep did you get at night? (This may be different than the number of hours you spent in bed.)
   HOURS OF SLEEP PER NIGHT _____

For each of the remaining questions, check the one best response. Please answer all questions.

5. During the past month, how often have you had trouble sleeping because you . . .

   a) Cannot get to sleep within 30 minutes

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

   b) Wake up in the middle of the night or early morning

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

   c) Have to get up to use the bathroom

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

   d) Cannot breathe comfortably

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

e) Cough or snore loudly

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

f) Feel too cold

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

g) Feel too hot

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

h) Had bad dreams

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

i) Have pain

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

j) Other reason(s), please describe_____
_____

How often during the past month have you had trouble sleeping because of this?

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

6. During the past month, how would you rate your sleep quality overall?
   Very good _____
   Fairly good _____
   Fairly bad _____
   Very bad _____

7. During the past month, how often have you taken medicine to help you sleep (prescribed or "over the counter")?

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

8. During the past month, how often have you had trouble staying awake while driving, eating meals, or engaging in social activity?

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

9. During the past month, how much of a problem has it been for you to keep up enough enthusiasm to get things done?
   - No problem at all _____
   - Only a very slight problem _____
   - Somewhat of a problem _____
   - A very big problem _____

10. Do you have a bed partner or room mate?
   - No bed partner or room mate _____
   - Partner/room mate in other room _____
   - Partner in same room, but not same bed _____
   - Partner in same bed _____

If you have a room mate or bed partner, ask him/her how often in the past month you have had . . .

a) Loud snoring

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

b) Long pauses between breaths while asleep

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

c) Legs twitching or jerking while you sleep

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

d) Episodes of disorientation or confusion during sleep

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

e) Other restlessness while you sleep; please describe _____
_____

Not during the past month _____ Less than once a week _____ Once or twice a week _____ Three or more times a week _____

Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ: Psychiatry Research, 28:193-213, 1989.

BIOGRAPHICAL SKETCH

Elizabeth M. Briones was born in Harlingen, Texas. She earned a Bachelor of Science degree in Psychology *cum laude* in 2011 and a Master of Arts degree in Experimental Psychology in 2014 from the University of Texas-Pan American. As an undergraduate she was placed on the Dean's List, became a member of Psi Chi the International Honor Society in Psychology, UTPA Ally, and served as an aid to David Sanchez's campaign for district court judge. As a graduate student Elizabeth received the STEM Success Scholarship, Stars Scholarship, and has been a dedicated member of the Center for Online Learning, Teaching and Technology staff at UTPA for 6 years.

Inquiries can be sent to ElzBriones@me.com or by mail to 808 N. Roberts St., Edinburg, TX 78541.