An examination of the equivalency of self-report measures obtained from crowdsourced versus undergraduate student samples

Elizabeth M. Briones¹ • Grant Benham¹

Published online: 23 February 2016
© Psychonomic Society, Inc. 2016

Abstract Increasingly, researchers have begun to explore the potential of the Internet to reach beyond the traditional undergraduate sample. In the present study, we sought to compare the data obtained from a conventional undergraduate college-student sample to data collected via two online survey recruitment platforms. In order to examine whether the data sampled from the three populations were equivalent, we conducted a test of equivalency using inferential confidence intervals—an approach that differs from the more traditional null hypothesis significance testing. The results showed that the data obtained via the two online recruitment platforms, the Amazon Mechanical Turk crowdsourcing site and the virtual environment of Second Life, were statistically equivalent to the data obtained from the college sample, on the basis of means of 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 for the use of these recruitment platforms, are discussed, and recommendations for other researchers who may be considering the use of these platforms are provided.

Keywords Crowdsourcing · Amazon Mechanical Turk · Second Life · Equivalence testing · Stress · Sleep

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. Early Web-based 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 2014, there were over three billion Internet users worldwide (“Internet Users in the World Distribution by World Regions,” 2014).

Though the methods are frequently conflated in discussions of online survey research, it is important to clearly distinguish between survey administration methods (e.g., pencil-and-paper vs. online survey) and survey recruitment methods (e.g., in-class announcement, posted flyer, e-mail, etc.). Online surveys have gained a foothold as a valid and reliable method for survey data collection (Gosling, Vazire, Srivastava, & John, 2004; Meyerson & Tryon, 2003) and may offer distinct advantages over traditional methods (Horton, Rand, & Zeckhauser, 2011; Lewis, Watson, & White, 2009; Ollesch, Heineken, & Schulte, 2006). With some notable exceptions (e.g., Buchanan et al., 2005), data collected via online survey methods do not appear to differ from the values obtained through traditional survey administration methods (De Beuckelaer & Lievens, 2009; Howell, Rodzon, Kurai, & Sanchez, 2010; Lewis et al., 2009). In concert with these new *methods* of survey administration, novel

✉ Grant Benham
grant.benham@utrgv.edu

¹ Department of Psychological Science, The University of Texas Rio Grande Valley, 1201 W. University Drive, Edinburg 78539, Texas, USA

approaches to *recruitment* of online participants have also emerged.

Emerging online recruitment venues

Online surveys, accessible to anyone with an Internet connection and appropriate hardware (e.g., smartphone, tablet, computer), have opened the door to a broader range of participants. As such, they have addressed some of the limitations of traditional methods, such as the overrepresented elderly and underrepresented younger populations commonly found in telephone surveys (Simons & Chabris, 2012). Though convenience samples of undergraduate students remain a staple of psychological research (Peterson, 2001), the Internet provides several venues through which survey participants may be recruited, affording access to groups of people that might otherwise be inaccessible (Dandurand, Shultz, & Onishi, 2008; Gosling, Sandy, John, & Potter, 2010; Wright, 2005). Some have explored the use of social media sites such as Facebook and search engine advertising (Google, Bing, or 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*: the ability to outsource services or tasks to large groups of people, specifically online communities (Crowdsourcing, n.d.).

Amazon's Mechanical Turk

Amazon's Mechanical Turk (MTurk; <https://www.mturk.com/mturk/welcome>) 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). A decade later, MTurk is still in its beta version. However, the site is now used by businesses and individuals alike to contract out duties that might otherwise be too costly to be freelanced by traditional means. 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 recruitment, 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 number of recent papers have supported the utility of MTurk as a mechanism for recruiting participants for online studies (Berinsky, Huber, & Lenz, 2012; Buhrmester, Kwang, & Gosling, 2011; Casler, Bickel, & Hackett, 2013; Crump, McDonnell, & Gureckis, 2013; Goodman, Cryder, & Cheema, 2012; Mason & Suri, 2012; Paolacci & Chandler, 2014; Paolacci, Chandler, & Ipeirotis, 2010; Shapiro, Chandler, & Mueller, 2013; Simons & Chabris, 2012). MTurk's workers have been found to be more attentive to instructions than unsupervised college students (Hauser & Schwarz, 2015) and 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 and 35 years old, these numbers widen the age gap by almost 10 years when compared to the traditional college-aged student, who is between the ages of 18 and 24 years old (Ipeirotis, 2010). It is not uncommon for MTurk workers to be paid US\$0.10 or similarly small incentives (Mason & Suri, 2012) for completing small tasks, and research suggests that these small payment incentives do not affect data quality (Buhrmester et al., 2011; Marge, Banerjee, & Rudnick, 2010; Mason & Watts, 2009).

Linden Lab's Second Life

Second Life (SL; <http://secondlife.com>) has been endorsed as one of the leading applications for immersive technology. Launched in 2003 as a “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. With over 3.2 billion U.S. dollars spent in virtual goods since 2003 (Linden Lab, 2013), 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.

Ten years after its inception, SL had garnered 36 million user accounts with an average of 400,000 new users monthly, and thus has been recognized for its substantial growth in

social media (Linden Lab, 2013). According to Linden Lab's latest *Key Metrics* report, the general SL population is made up of roughly 59 % males and 41 % females, with over 28 million hours being spent "in-world" (Linden, 2008). The United States is noted as having the highest number of active users, making up 35 % of the population, with Germany and United Kingdom following and accounting for approximately 16 %, combined. More generally, similar to MTurk's demographic population, the age distribution of SL users are slightly higher when compared to undergraduate populations, with approximately 25 % of users between 18 and 24 years old, 35 % between 25 and 34 years, and 39 % 35 years or older. Data on demographic composition are limited to self-reports and Linden Lab's released data policy, which was suspended in 2008. Similar gender and location demographics were obtained more recently in a survey of 330 SL users (Dean, Cook, Murphy, & Keating, 2012), though a slightly younger distribution of ages was observed. Additionally, Dean et al. found that 57 % of respondents were extroverted (self-report of how others would describe them in real life), and over 60 % had completed some college or had already graduated.

A number of researchers have examined SL as a venue for psychological research, including survey administration. Early research required the user to leave the SL interface and open an Internet browser to take part in an online survey, interrupting the user's virtual experience and potentially altering attitudes and perceptions that were trying to be assessed (Bell et al., 2009). A more sophisticated method, the virtual data collection interface (VDCI), was later developed by Bell, Castronova, and Wagner (2008) to provide a way to collect survey data while the user remained immersed in the virtual environment. As the SL platform evolved, it provided the tools to natively incorporate Web-based surveys without the need for specialized VDCI interfaces. SL users are encouraged to participate in research through payment in a virtual currency; although there is no established standard, the average amount paid to SL participants for survey research has generally been around L\$250, which is approximately equivalent to US\$1.00. Several studies have explored methodological considerations of research performed in SL, but few have examined the utility of SL as a recruitment platform for survey studies in relation to other established approaches (Anstadt, Bradley, & Burnette, 2013; Bell, 2008; Bell et al., 2009; Dean et al., 2012; Martey & Shiflett, 2012).

Equivalency testing

The aim of this study was to compare the data obtained from a conventional undergraduate college-student sample to those obtained via two online recruitment platforms. 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 find a significant difference between groups does not directly demonstrate *equivalence* between the 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 asks "What is the probability of having a difference this large, assuming the populations are actually the same?" Equivalency testing asks "What is the probability of having a difference this small assuming the populations are actually different?" This approach is not new to the field of biostatistics: For example, medical researchers must often 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). Though some social science researchers have adopted this statistical approach, equivalency testing remains relatively underutilized.

Thus, the purpose of our study was not to demonstrate that the obtained means were smaller or larger than those obtained from a traditional college sample, but to examine whether we obtained similar (equivalent) results when using these nontraditional methods of recruitment. A number of methods have been developed to test for statistical equivalence, including the two one-sided test procedure (Schuirmann, 1987). For our study, we elected to base our determination of equivalence on *inferential* confidence intervals (ICIs; Tryon, 2001; Tryon & Lewis, 2008), because this method provides a visual representation of the equivalency of groups while simultaneously evaluating statistical significance. It should be noted that this is one method for comparing results from different populations. In Behrend et al.'s (2011) detailed analysis of crowdsourced versus college-student survey responses, a variety of comparisons between the groups were made (including completion time and the psychometric properties of the scales). Our analysis, although restricted to the equivalency of means, may be viewed as an alternative path for researchers to follow toward the same goal.

Though a number of previous studies have examined the equivalency of traditional samples with MTurk and SL samples, there has been renewed emphasis on the importance of the reproducibility of findings in psychological research (Open Science Collaboration, 2015). Our study builds upon previous findings by Behrend et al. (2011) and Dean et al. (2012) by simultaneously examining MTurk and SL and by using a statistical approach that is infrequently used in behavioral sciences. Given the potential of both MTurk and SL to

serve as low-cost survey recruitment methods that connect with diverse populations, it is important to examine the extent to which data collected from these venues are similar to those obtained through a more common means: recruitment from undergraduate psychology courses. In this study, therefore, we compared a relatively novel online survey recruitment method (SL) and a more established method (MTurk) with a traditional college undergraduate approach. Our aim was to test the equivalency of the data obtained via these three samples, using two established measures selected from the field of health psychology: to examine whether the two crowdsourced recruitment methods offer a practical alternative to traditional college-student recruitment.

Method

Participants

Our participants included the following three groups:

- *Three hundred 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. The 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. The completion rate for the college sample was 94 %.
- *Two hundred individuals recruited through MTurk*. 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 completion rate of 94 % over two and a half weeks in February 2014.
- *Sixty-seven respondents¹ recruited through SL in spring 2014*. These participants ranged in age from 18 to 61 ($M = 29.1$, $SD = 11.2$) and were 55 % female and 30 % Hispanic. The completion rate for the SL survey was 87 % over five and a half weeks.

Measures

Two measures of sleep and stress were selected for from the field of health psychology. The measures were selected because (1) they were contained within an existing dataset of referent sample college students and (2) they are both well-validated and widely used measures within the field of health psychology, garnering over 10,000 citations combined since their publication in the late 1980s.

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

Perceived Stress Scale (PSS) The PSS (Cohen & Williamson, 1988) is a measure of self-perceived stress. The PSS is a 10-item Likert-type scale that asks respondents “In the last week, how often have you . . .” and completes the prompt with items such as “felt nervous and stressed?” and “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. The PSS has been reported as a better predictor of psychological symptoms, physical symptoms, and health service utilization than are life-event scales (Cohen et al., 1983). Possible scores range from 0 to 40, with higher scores representing more stress. Good internal reliability has been reported for the PSS, Cronbach’s $\alpha = .91$.

Pittsburgh Sleep Quality Index (PSQI) The PSQI (Buysse et al., 1989) 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. 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 SL users, and thus a location question was asked of the SL participants.

Procedures

College-student recruitment procedure The college-student data were extracted from an existing dataset based on an online survey study conducted in 2013. The 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 The 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. A worker’s approval rating is measured by the MTurk system as the percentage of

assignments submitted by the worker that have been approved by requesters—that is, what percentage of HITs they’ve successfully completed. The participant was also required to be 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. When they clicked on the link for the present study, a page with a short description of the survey, the survey link, and a verification code box was shown. If the worker chose to accept the HIT, the user was redirected to an external webpage for the survey.

SL recruitment procedure The SL participants were recruited through an advertisement published on SL’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 SL (Bell et al., 2009; Dean et al., 2012). This can be accessed through the game’s user interface or through SL’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’s SL island location, where a virtual kiosk was set up. A folder with instructions, a copy of the informed consent, along with the *heads-up display* (HUD) item was automatically given to the user when a box labeled “CLICK ME for survey” was touched. The HUD, once worn by the avatar, was seen as a pop-up on the user’s screen occupying a large portion of the monitor. Specific to SL, the pop-up presented the survey as it would have been seen in an external Web browser. This method provided added privacy; once the user had attached it, it was invisible to other residents.

Survey administration All participants were directed to complete the online survey, created and hosted through Qualtrics (Provo, UT; www.qualtrics.com). The survey included demographic questions, an Internet usage question, and the aforementioned stress and sleep scales.

The MTurk participant was presented with a consent form, which the user agreed to before proceeding with the survey. Once the participant had 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 US\$0.10 for completion of the survey.

For the SL participants, the HUD provided the survey just as the MTurk user viewed it, with the exception that participation was performed exclusively in SL’s interface. Both the MTurk and SL surveys were identical, with the exception of

two questions added to the SL survey version: (1) respondents were asked if they resided either inside or outside the United States, and (2) 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 US\$0.10) for completion of the survey. After two weeks, due to a low response rate, an announcement was placed on SL’s Facebook page advertising the survey, and the rate was increased to L\$250 (approximately US\$1.00).

Results

Data analysis

Equivalency 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 Fig. 1.

Rather than relying on standard descriptive CIs, Tryon’s procedure shortens these such that nonoverlapping ICIs are algebraically equivalent to NHST between two means. In Fig. 1, the 95 % ICIs of Groups 1 and 2 are presented. The first step in equivalency testing is to establish the Delta (Δ) that will be used: an a priori criterion of how far the two groups can differ while still being considered equivalent. Given the lack of prior research on which to base this Δ value, we elected to follow the criterion used by a number of other researchers, setting Δ to ± 20 % around the selected reference group (Cribbie et al., 2004; Epstein et al., 2001; Lewis et al., 2009; Rogers & Howard, 1993; Rusticus & Lovato, 2011; Steele, Mummery, & Dwyer, 2009; Weigold et al., 2013). Once ICIs for both groups are calculated, an equivalence range (eRg), or “maximum probable difference between the two means” (Tryon, 2001, p. 379), is determined. Statistical equivalence is then determined by measuring the eRg against the Δ . This can be visually determined by graphing the eRg and Δ measurement bars, with the Δ starting at the lower bound of the eRg.²

In the example represented in Fig. 1, the maximum probable difference between the two means (i.e., the equivalence range, eRg) extends from the lower CI limit of the lesser mean (Group 1) to the upper CI of the greater mean (Group 2). The value of Δ is set at 20 % of the mean of Group 1 (our selected reference group), equating to a Δ of 2.0. Delta is *anchored* at the lower limit of the eRg range (9.5) and extends 2.0 points,

² For readers interested in learning more about this method, we recommend the two excellent reports by Tryon (Tryon, 2001; Tryon & Lewis, 2008).

Fig. 1 Sample of equivalence test output using the inferential confidence interval method

to a value of 11.5. Because the equivalence range (eRg) is contained within the Δ range, the two groups are considered statistically equivalent.

Statistical difference is said to exist between the two means if the ICIs of Groups 1 and 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 (Δ)—that is, when eRg is contained within Δ . *Statistical indeterminacy* is said to exist when the means are neither statistically different nor equivalent (Beckstead, 2008; Tryon, 2001; Tryon & Lewis, 2008). In Fig. 1, the ICIs of Groups 1 and 2 overlap, indicating no statistical difference (the traditional NHST approach). Additionally, the equivalence range (eRg) fits within the 20 % Δ 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 those obtained from the traditional college-student samples.³ Given that our Δ values were based on a rule-of-thumb criterion, we also determined the Δ value, or critical Δ , for which equivalence would no longer hold (i.e., the Δ value that would result in a lack of equivalency between the two groups).

In addition to comparisons of the measurement means, we wanted to examine whether the correlations between stress and sleep quality would differ between the 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.

³ In our analyses, we used our college-student sample as the reference group because they were viewed as the more conventional sample.

Given the small number of a priori tests being conducted, and the debate regarding the appropriateness of such methods (Feise, 2002; Nakagawa, 2004; Perneger, 1998; Rothman, 1990), 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 the sample characteristics for each group according to age, gender, ethnicity, and education. The age data were nonnormal and heteroscedastic; therefore, the Welch test was applied to the ranked age values (as was recommended by Cribbie, Wilcox, Bewell, & Keselman, 2007). An omnibus test demonstrated that age differed significantly between the groups, $F(2, 382.55) = 350.91, p < .001$. Because our primary goal was to evaluate the equivalency of

Table 1 Demographic sample percentages for characteristics by group

	College Group (<i>N</i> = 363) %	MTurk Group (<i>N</i> = 200) %	SL Group (<i>N</i> = 67) %
Age			
18–25	82.4	23.5	52.2
26–35	13.8	37.0	23.9
36–45	3.3	14.0	10.4
46–55	0.6	12.0	10.4
Above 55	0.0	13.5	3.0
No response	0.0	0.0	0.0
Gender			
Male	21.5	38.0	44.8
Female	78.5	62.0	55.2
No response	0.0	0.0	0.0
Ethnicity			
Caucasian	1.9	69.5	46.3
Hispanic	92.3	10.0	29.9
African American	1.4	10.0	6.0
Asian	2.2	6.0	1.5
Other	2.0	4.5	16.4
No response	0.3	0.0	0.0
Education			
Less than high school	–	0.0	4.5
High school/GED	–	7.5	34.3
Some college	–	31.5	29.9
2-year degree	–	10.0	13.4
4-year degree	–	36.5	13.4
Master's degree	–	12.0	3.0
Doctoral degree	–	0.5	1.5
Professional degree	–	2.0	0.0
No response	–	0.0	0.0

Education level was not assessed for the college group, denoted by the dashes

data obtained via the crowdsourced samples (MTurk and SL) relative to the reference (college-student) sample, post-hoc analyses were restricted to those comparisons. Post-hoc Welch t tests on ranked age revealed statistically significant differences between the college students and MTurk workers, Welch $t(560.95) = 11.94$, $p < .001$, and between the college students and SL participants, Welch $t(427.93) = 24.84$, $p < .001$. As a group, the college students tended to be younger (median age = 21.0) than the participants recruited through MTurk (median age = 32.0) or SL (median age = 25.0).

An omnibus Pearson's chi-square test demonstrated a significant relationship between recruitment group and gender, $\chi^2(2, N = 630) = 25.80$, $p < .001$. Subsequent post-hoc 2×2 contingency tables revealed that, relative to the college undergraduates, there were significant relationships between recruitment group and gender for both MTurk, $\chi^2(1, N = 563) = 17.70$, $p < .001$, and SL, $\chi^2(1, N = 431) = 15.62$, $p < .001$. Both the MTurk and SL samples contained higher proportions of males.

An omnibus chi-square analysis indicated a significant relationship between recruitment group and ethnicity, $\chi^2(2, N = 629) = 392.17$, $p < .001$. Relative to the college undergraduates, post-hoc testing revealed that this relationship was significant for both MTurk, $\chi^2(1, N = 562) = 377.24$, $p < .001$, and SL, $\chi^2(1, N = 430) = 158.44$, $p < .001$. Both the MTurk and SL samples contained lower proportions of Hispanics. Similar differences were shown with race: Overall, we found a significant relationship between recruitment group and race, $\chi^2(2, N = 629) = 70.47$, $p < .001$. Post-hoc testing revealed that, relative to the college undergraduates, this relationship was significant for both MTurk, $\chi^2(1, N = 562) = 59.48$, $p < .001$, and SL, $\chi^2(1, N = 430) = 23.68$, $p < .001$. Both the MTurk and SL samples contained higher proportions of Caucasians.⁴ Information about the highest level of education obtained was not assessed for the college group. However, given the setting, we considered it reasonable to categorize all of the students as having achieved either some college hours or a 2-year degree. Furthermore, the independent categories were collapsed down into "No college," "Some college or 2-year degree," and "bachelor's or advanced degree." An omnibus test demonstrated a significant relationship between recruitment group and education level, $\chi^2(4, N = 626) = 377.92$, $p < .001$. Relative to the college sample, post-hoc testing revealed that this relationship was significant for both MTurk, $\chi^2(1, N = 559) = 262.31$, $p < .001$, and SL, $\chi^2(1, N = 430) = 225.84$, $p < .001$. Overall, the MTurk group had a higher level of education than the undergraduate sample, with 50 % of MTurk participants having obtained either a bachelor's or advanced degree. Although almost 20 % of the SL participants had also obtained a bachelor's or advanced

degree, 39 % of them reported not having taken any college hours. In sum, both of the crowdsourced samples differed from the college-student sample on a number of basic demographic variables. Our college sample contained a higher proportion of Hispanic and a lower proportion of Caucasian participants than the crowdsourced samples, which is largely explained by the fact that they were recruited from a Hispanic Serving Institution in a predominantly Hispanic region of the country.

Equivalence tests

Tests of statistical equivalence and statistical difference were conducted to examine whether the MTurk and SL participants' stress and sleep quality scores were the same as those obtained from the college students.⁵ The ICI statistics, equivalence determinations, and the critical Δ value for equivalence no longer held, and our statistical difference determinations are summarized in Table 2 for each comparison.

Research Question (RQ) 1: Are the PSS scores obtained via SL equivalent to those obtained from college students? ICIs were compared for the mean PSS scores of the college group ($N = 363$), $M = 20.06$, $SD = 6.16$, 95 % CI⁶ [19.42, 20.69], and of the SL group ($N = 67$), $M = 19.43$, $SD = 4.46$, 95 % CI [18.35, 20.52]. Using a value of $\Delta = 4.01$ ($20.06 \times .20$), we found statistical equivalence and no statistical difference between the groups. The equivalence testing results are graphically depicted in Fig. 2.

RQ 2: Are the PSS scores obtained via MTurk equivalent to those obtained from college students? For our second research question, we compared the ICIs of the mean PSS score of the college group with the mean PSS score of the MTurk group ($N = 200$), $M = 19.11$, $SD = 7.35$, 95 % CI [18.09, 20.13]. As is shown in Fig. 2, with a Δ set at 4.01, we found statistical equivalence and no statistical difference between the groups.

RQ 3: Are the sleep quality (PSQI) scores obtained via SL equivalent to those obtained from college students? The third research question attempted to assess the equivalence for the sleep quality scores between the college and SL samples. Delta was computed at ± 20 % of the college group's mean PSQI score, $\Delta = 1.39$. The results indicated that the college group ($N = 361$), $M = 6.94$, SD

⁴ To ensure that all expected cell values were above 5, and to simplify the analysis, race was dichotomized into Caucasian and non-Caucasian.

⁵ Though an argument might be made for the creation of matched samples, on the basis of 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.

⁶ These descriptive confidence intervals should not be confused with the ICIs used in the statistical equivalence analyses.

Table 2 Inferential confidence intervals and equivalence testing results

	95 % ICIs	Statistical Equivalence at $\Delta = 20$ %	Critical Δ	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]			

CO = college group, MT = MTurk group, SL = Second Life group; 95 % ICIs = inferential confidence intervals [lower bound, upper bound]; Critical Δ = values of Δ for which equivalence would no longer hold

= 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 2 presents the ICI comparison.

RQ 4: Are the sleep quality (PSQI) scores obtained via MTurk equivalent to those obtained from college students? ICIs were compared for the mean sleep scores of the college group and the MTurk group ($N = 198$), $M = 6.99$, $SD = 3.67$, 95 % CI [6.50, 7.48], and Δ was 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 Fig. 2.

Correlational comparisons

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

PSS score comparison of ICIs at 95% for college and SL groups, eRg = 1.88, $\Delta = 4.01$.

PSS score comparison of ICIs at 95% for college and MTurk groups, eRg = 2.15, $\Delta = 4.01$.

PSQI score comparison of ICIs at 95% for college and SL groups, eRg = 1.28, $\Delta = 1.39$.

PSQI score comparison of ICIs at 95% for college and MTurk groups, eRg = 0.68, $\Delta = 1.39$.

Fig. 2 Graphical depictions of CO = college, MT = MTurk, and SL = Second Life PSS and PSQI score comparisons for both statistical difference and equivalence

and college students? To assess the differences of the correlations between samples for our 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 the college students, $r(361) = .45, p < .001$, and in the SL participants, $r(64) = .39, p = .002$. The difference between these correlations was not statistically significant, $z = 0.53, p = .60$. RQ 6: Does the strength of the relationship between sleep and stress differ significantly between the MTurk participants and college students? Sleep and stress were significantly correlated in the college students, $r(361) = .45, p < .001$, and in the MTurk participants, $r(198) = .47, p < .001$. The difference between these correlations was not statistically significant, $z = 0.29, p = .78$.

Supplemental analysis

The main focus of the study was to determine whether data obtained from two groups recruited via emerging online platforms were equivalent to those from 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 Δ s chosen at $\pm 20\%$ of the MTurk group's mean stress score ($\Delta = 3.82$) and $\pm 20\%$ of the MTurk PSQI mean score ($\Delta = 1.40$), and $\alpha = .05$, we found both tests to be statistically equivalent and not statistically different.

We also compared the relationships between stress and sleep for the MTurk group and the SL group. Sleep and stress were significantly correlated in the MTurk group, $r(198) = .47, p < .001$, and in the SL participants, $r(64) = .39, p = .002$. The difference between these correlations was not statistically significant, $z = 0.67, p = .50$.

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 those collected via more conventional means (recruitment from undergraduate student classes). For the purposes of this study, we considered the viability of the samples based on their practicability for researchers (ease of use, recruitment success, response costs, etc.), alongside their statistical equivalency. Though there are other methodological and statistical paths toward the evaluation of viability and equivalency (e.g., those

of Behrend et al., 2011; Crump et al., 2013; Germine et al., 2012; and Sprouse, 2011), we believe that our findings provide a unique contribution. To our knowledge, this is the first study to simultaneously compare the MTurk and SL platforms to a traditional undergraduate sample and to use a test of statistical equivalency toward that end. On the basis of the means of two standardized measures, and the observed correlations between those two measures, our results indicate that the data collected from these two online platforms are statistically equivalent to those obtained from the college-student sample. These findings suggest that such methods might serve as an alternative or supplement to conventional classroom recruitment.

The demographic data collected from our MTurk and SL groups indicated that these nontraditional samples showed greater diversity than our college sample in terms of age, ethnicity, race, and education, and a more balanced proportion of males to females. These data are concordant with previous findings (Berinsky et al., 2012; Buhrmester et al., 2011; Paolacci & Chandler, 2014; Paolacci et al., 2010) and suggest that inclusion of such crowdsourced samples might ameliorate some of the concerns related to lack of sample generalizability in psychological research: For those concerned with college populations being predominantly *WEIRD*—or Western, educated, industrialized, rich, and from democratic societies (Henrich, Heine, & Norenzayan, 2010)—the use of online crowdsourced samples could be used as a complementary or alternative recruitment venue. Indeed, Paolacci et al. (2010) have argued that MTurk workers may be more representative of the U.S. population as a whole than are college students, matching the general population more closely on gender, race, age, and education. This is certainly true in relation to our obtained samples. However, such crowdsourced populations may be unusual in their own right and do not substitute the need for additional research on stress and sleep with specialized populations such as medical professionals, military personnel, cancer patients, and shift-workers. Additionally, as interest in recruiting study participants through MTurk grows, researchers need to remain cognizant of the fact that research-naïve MTurk workers may be a limited commodity. In their recent analysis, Stewart et al. (2015) estimated that the number of MTurk workers who participate in research at any given time is only 7,300—considerably smaller than the total number of MTurk workers advertised by Amazon. Although turnover is comparable to what one finds in a university setting, with approximately half of the workers being replaced every seven months, 36 % of MTurk workers complete HITS from more than one lab (Stewart et al.). As the system becomes more populated with research studies, legitimate concerns over threats to internal validity may be raised due to a lack of participant naiveté.

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 the administration of online surveys (Couper, & Miller, 2008; Kraut et al., 2004; Riva, Teruzzi, & Anolli, 2003) and a tool for acquiring more diverse 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., online surveys vs. pencil-and-paper) and more traditional samples (e.g., online samples vs. undergraduate students). In the present study, we examined whether the data obtained from two emerging online recruitment platforms were equivalent to those obtained from a conventional college sample.

Our study does not negate the continued utility of undergraduate participant pools for survey recruitment, but it adds to the growing body of literature demonstrating the validity of MTurk as a viable alternative (e.g., Behrend et al., 2011; Buhrmester et al., 2011; Goodman et al., 2012; Paolacci & Chandler, 2014). Our research has replicated and extended previous studies of MTurk by using equivalence testing to compare the data obtained with traditional (undergraduate) samples, a statistical approach that is underutilized, but appropriate for, the examination of group equivalency. We selected two widely used standardized measures from the field of health psychology as a basis for comparison: the PSS (Cohen et al., 1983) and the PSQI (Buysse et al., 1989). On the basis of these measures, MTurk produces data that are statistically equivalent to those 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. A number of the studies conducted have tended to focus on examining the unique characteristics of SL participants (Hooi & Cho, 2014; McLeod, Liu, & Axline 2014), the behavior of SL avatars in virtual worlds (Grinberg, Careaga, Mehl, & O'Connor, 2014; Hooi & Cho, 2013), the use of SL for online instruction within educational institutions (Halvorson, Ewing, & Windisch, 2011; Inman, Wright, & Hartman, 2010), or the feasibility of conducting social experiments and experimental manipulations in SL (Greiner, Caravella, & Roth, 2014; Lee, 2014; Tawa, Negrón, Suyemoto, & Carter, 2015). A handful of studies have used SL as a tool for recruitment, though generally the recruitment goals are directed at specific populations (Keelan et al., 2015; Swicegood & Haque 2015). Thus, SL provides a unique platform for conducting certain types of research and may be useful for researching specific populations, but its usefulness as a recruitment platform, relative to MTurk, may be limited. Recruitment through this platform also created some practical challenges.

MTurk versus SL: 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 the viability of these nonconventional 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. SL was designed as a 3-D 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: labor-intensive setup that may be unreasonable to researchers unfamiliar with the platform. Thus, SL can involve a steep learning curve for researchers, particularly for those who are not already familiar with the nature of the platform. Although the use of SL may show some promise, researchers should carefully consider its feasibility in relation to their own technical abilities and should recognize that this unfamiliarity may also be an issue for participants who are not already SL users (Keelan et al., 2015), narrowing its utility further.

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 present study, we elected to use a fairly common rate for short MTurk tasks: a payment of US\$0.10.⁷ To maintain consistency across the two platforms, we offered an equivalent US\$0.10 payment for our

⁷ Our payment rate was consistent with those in a number of prior studies (e.g., Bergvall-Kåreborn & Howcroft, 2014; Buhrmester et al., 2011; Goodman et al., 2012; Horton et al., 2011; Mason & Suri, 2012; Mason & Watts, 2009; Paolacci et al., 2010). Although such payments appear to be effective, some have begun to ask whether such wages are ethically appropriate (Irani & Silberman, 2013). Due to the contemporary nature of crowdsourcing, however, minimum wage standards have yet to be established.

SL participants (converted to the SL currency of L\$22). Due to a low response rate from the 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 enter a validation code that is presented at the end of the research survey in order to receive their payment from the researcher's MTurk account. SL does not have a comparable system for payments, and therefore the process for paying SL participants is considerably more involved. SL participants included their (unique) SL avatar name as part of the online survey. The Qualtrics survey responses were periodically checked for completed surveys and the 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.

When considering the use of commercial companies for recruitment, researchers should recognize that they have no control over pricing increases for such services. In June 2015, Amazon implemented substantial price increases for their MTurk platform, doubling the standard commission it takes from 10 % to 20 % of the total paid and adding an additional 20 % commission for requesters who use ten or more MTurk workers. Thus, the service charge has quadrupled for researchers who need ten or more participants for their study (at the time of writing, Amazon does not offer separate MTurk fee schedules for commercial vs. academic research). Such costs may lead researchers to consider existing "micro-batching" workarounds to the ten-or-more commission fee, or to explore competing academic-focused online participant recruitment systems. Despite concerns over these price increases, MTurk remains a relatively inexpensive and established platform for online recruitment.

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 crowdsourced recruitment may provide a reasonably cost-effective approach to obtaining adequate sample sizes. On the basis of the short survey used in the present 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. The additional efforts did advance the

SL response rate from 17 responses in 2 weeks to 51 responses in 3.5 weeks. However, by that time it was apparent that the SL venue was an inefficient option in relation to MTurk, and data collection was ended. It should be noted that our study did not exhaust all recruitment methods within SL. Although we selected the most effective recruitment method (SL classifieds) on the basis of previous research (Dean et al., 2012), and followed up with a post on the SL Facebook page, we did not exhaust all SL recruitment methods. Our approach was to balance effort against reward and, in this respect, SL recruitment was a less successful method. Future studies may improve the recruitment approach by investing in paid advertisements that increase an event's visibility in SL's advertising network or through underexplored venues. Recently, for example, Swicegood and Haque (2015) successfully recruited nearly 1/3 of their participants through an SL "New World Notes" blog page.

Screening respondents and reducing multiple submissions

MTurk provides a system for screening workers on the basis of various criteria and displaying listings of prequalified HITs to the individual workers. As such, MTurk provides a level of quality control (e.g., workers must meet a specific approval rating in order to view the research HIT) and flexibility in limiting samples on the basis of specific demographics (e.g., must be a U.S. citizen). SL does not offer a similar function; screening criteria must be established through recruitment announcements and survey questions and rely 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. SL provides no added protection, thus increasing the possibility of multiple responses from a single participant. SL allows for multiple avatar accounts to be made with one e-mail address, making it possible for one person to take a survey over multiple avatar accounts. Though multiple responses are a limitation for all online surveys, MTurk offers an additional feature to combat this weakness.

Institutional review board (IRB) restrictions

Though policies may vary from institution to institution, human subjects research approval by 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 US\$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 present 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 suffered 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 the 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 than in 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 5 min to complete. Whether the observed equivalency would hold 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 study 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. Our comparison of the means obtained from these two scales also limits our ability to claim equivalencies beyond survey methodologies. However, a substantial body of research now suggests that MTurk and SL may be viable tools for online experiments (Berinsky et al., 2012; Casler et al., 2013; Horton et al., 2011; Keelan et al., 2015; Shapiro et al., 2013; Simcox & Fiez, 2014; Sprouse, 2011; Yee, Bailenson, Urbanek, Chang, & Merget, 2007). In a comprehensive study by Crump et al. (2013), for example, the research team replicated ten cognitive psychology experiments with MTurk participants, concluding that the data quality was reasonably high and compared well with laboratory studies. Tools for conducting replicable behavioral experiments on MTurk, such as the open-source project “psi-Turk” (Gureckis et al., 2015) are likely to aid in this endeavor. In order to add greater support to the viability of online platforms for experimental research, we encourage researchers to consider incorporating statistical tests of equivalency as a complement to the more traditional NHST.

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. Although 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, involves more difficulty when processing payments, and does not generate as high a response rate.

In our study, we did not attempt to match samples on the basis of demographic characteristics, opting instead 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 did the undergraduate sample. It is important to recognize this lack of similarity between the samples, but the fact that equivalencies were demonstrated *in spite of* these demographic differences provides stronger evidence for the utility of these nonconventional recruitment techniques. 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. Equally, our findings suggest that researchers who limit their samples to undergraduate students may feel somewhat more confident about the generalizability of their data beyond the traditional college sample.

Given the IRB requirements at our institution, our present MTurk sample was limited to MTurk workers who were U.S. residents or citizens and who, on the basis of demographics, appeared to be relatively well-educated and overrepresented 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 face 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 international participants.

Finally, because methods to assess correlational equivalence have yet to be established, our analyses of the obtained correlation coefficients was limited to more traditional NHST methods. As a result, the hypotheses being tested for our correlation comparisons did 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 on Internet access (Buchanan & Smith, 1999; Kraut et al., 2004) may not be as relevant today, but MTurk and SL users may still be biased samples, given that they are self-selected (and arguably “unusual”) groups. Although the equivalency of these two groups to college students suggests that crowdsourced recruitment may be a viable alternative to

recruiting from undergraduate courses, the very fact that these groups did not differ may seem somewhat surprising. *Why* should these groups, composed 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 have shown good distributions of scores. They appear to be sensitive instruments. A second explanation may be that our chosen Δ was too generous. The determination of equivalency depends, in part, on the a priori establishment of a reasonable Δ , a value that describes how far the two groups can differ while still being considered equivalent. Our Δ was based on a recommended rule-of-thumb 20 % value, but future research may benefit from more sophisticated determinations of Δ . As we 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 Δ s would have failed to demonstrate equivalency while simultaneously maintaining NHST findings of no statistically significant differences: The groups would have been neither equivalent nor significantly different (a scenario of *statistical indeterminacy*; Tryon, 2001; Tryon & Lewis, 2008). Third, it may be that these groups really were quite similar—that college students do not differ greatly from the individuals who complete HITs on MTurk or who engage in SL. Indeed, on the basis of our samples and closely matching previous studies (Dean et al., 2012; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010), 90 % of MTurk workers and 62 % of SL respondents reported having at least some college hours.

In conclusion, the data obtained from MTurk and SL samples appear to be statistically equivalent to those 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 a low cost and a more diverse representation of demographics. The utility of SL 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, will foster further exploration of the platform as a tool for conducting survey research in psychology.

Author note The authors thank Darrin Rogers and Mark Winkel for their feedback on an earlier draft of the manuscript. This research was supported in part by funds from the Center for Online Learning, Teaching and Technology at the University of Texas–Pan American.

References

- Anstadt, S., Bradley, S., & Burnette, A. (2013). Virtual worlds: In-world survey methodological considerations. *Journal of Technology in Human Services*, 31, 156–174. doi:10.1080/15228835.2013.784107
- 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. S., Sharek, D. J., Meade, A. W., & Wiebe, E. N. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, 43, 800–813. doi:10.3758/s13428-011-0081-0
- Bell, M. W. (2008). Toward a definition of “virtual worlds.” *Journal of Virtual Worlds Research*, 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*. German Council for Social and Economic Data Research Notes, no. 37.
- Bell, M., Castronova, E., & Wagner, G. (2009). *Surveying the virtual world-A large scale survey in Second Life using the virtual data collection interface (VDCI)*. German Council for Social and Economic Data Research Notes, no. 40.
- Bergvall-Kåreborn, B., & Howcroft, D. (2014). Amazon Mechanical Turk and the commodification of labour. *New Technology, Work and Employment*, 29, 213–223. doi:10.1111/ntwe.12038
- Berinsky, A., Huber, G., & Lenz, G. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, 20, 351–368.
- Buchanan, T., Ali, T., Heffernan, T., Ling, J., Parrott, A., Rodgers, J., & Scholey, A. (2005). Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire. *Behavior Research Methods*, 37, 148–154. doi:10.3758/BF03206409
- Buchanan, T., & Smith, J. L. (1999). Using the Internet for psychological research: Personality testing on the World Wide Web. *British Journal of Psychology*, 90, 125–144.
- 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*, 6, 3–5. doi:10.1177/1745691610393980
- Buyssse, 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, 193–213.
- Casler, K., Bickel, L., & Hackett, E. (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, 2156–2160. doi:10.1016/j.chb.2013.05.009
- 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. M. (1988). Perceived stress in a probability sample of the United States. In S. Spacapan & S. Oskamp (Eds.), *The social psychology of health* (pp. 31–67). Newbury Park: Sage.
- Couper, M. P., & Miller, P. V. (2008). Web survey methods: Introduction. *Public Opinion Quarterly*, 72, 831–835.
- Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, 60, 1–10.
- Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, 6, 117–132.

- Crowdsourcing. (n.d.). Merriam-Webster's Online Dictionary. Retrieved July 23, 2013, from 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(57410), 1–18. doi:10.1371/journal.pone.0057410
- Dandurand, F., Shultz, T., & Onishi, K. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, 40, 428–434. doi:10.3758/BRM.40.2.428
- De Beuckelaer, A., & Lievens, F. (2009). Measurement equivalence of paper-and-pencil and Internet organisational surveys: A large scale examination in 16 countries. *Applied Psychology*, 58, 336–361. doi:10.1111/j.1464-0597.2008.00350.x
- Dean, E., Cook, S., Murphy, J., & Keating, M. (2012). The effectiveness of survey recruitment methods in Second Life. *Social Science Computer Review*, 30, 324–338. doi:10.1177/0894439311410024
- Duda, M., & Nobile, J. L. (2010). The fallacy of online surveys: No data are better than bad data. *Human Dimensions of Wildlife*, 15, 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, 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, e20. doi:10.2196/jmir.1978
- Gallander Wintre, M., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology*, 42, 216–225. doi:10.1037/h008689
- Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., & Wilmer, J. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review*, 19, 847–857. doi:10.3758/s13423-012-0296-9
- 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, 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. *Behavioral and Brain Sciences*, 33, 94–95. 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, 93–104. doi:10.1037/0003-066x.59.2.93
- Greiner, B., Caravella, M., & Roth, A. (2014). Is avatar-to-avatar communication as effective as face-to-face communication? An Ultimatum Game experiment in First and Second Life. *Journal of Economic Behavior & Organization*, 108, 374–382. doi:10.1016/j.jebo.2014.01.011
- Grinberg, A., Careaga, J., Mehl, M., & O'Connor, M. (2014). Social engagement and user immersion in a socially based virtual world. *Computers in Human Behavior*, 36, 479–486. doi:10.1016/j.chb.2014.04.008
- Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., . . . Chan, P. (2015). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0642-8
- Halvorson, W., Ewing, M., & Windisch, L. (2011). Using Second Life to teach about marketing in Second Life. *Journal of Marketing Education*, 33, 217–228. doi:10.1177/0273475311410854
- Hauser, D. J., & Schwarz, N. (2015). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-015-0578-z
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33, 61–83. doi:10.1017/S0140525X0999152X
- Hooi, R., & Cho, H. (2013). Deception in avatar-mediated virtual environment. *Computers in Human Behavior*, 29, 276–284. doi:10.1016/j.chb.2012.09.004
- Hooi, R., & Cho, H. (2014). Avatar-driven self-disclosure: The virtual me is the actual me. *Computers in Human Behavior*, 39, 20–28. doi:10.1016/j.chb.2014.06.019
- Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, 14, 399–425. doi:10.1007/s10683-011-9273-9
- Howell, R., Rodzon, K., Kurai, M., & Sanchez, A. (2010). A validation of well-being and happiness surveys for administration via the Internet. *Behavior Research Methods*, 42, 775–784. doi:10.3758/BRM.42.3.775
- 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, 44–63.
- Internet Users in the World Distribution by World Regions. (2014). Internet world stats—Usage and population statistics. Retrieved March 16, 2015, from 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>
- Irani, L. C., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)* (pp. 611–620). New York, NY: ACM Press. doi:10.1145/2470654.2470742
- Keelan, J., Beard Ashley, L., Morra, D., Busch, V., Atkinson, K., & Wilson, K. (2015). Using virtual worlds to conduct health-related research: Lessons from two pilot studies in Second Life. *Health Policy and Technology*, 4, 232–240. doi:10.1016/j.hlpt.2015.04.004
- 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, 105–117. doi:10.1037/0003-066x.59.2.105
- Lee, J. (2014). Does virtual diversity matter?: Effects of avatar-based diversity representation on willingness to express offline racial identity and avatar customization. *Computers in Human Behavior*, 36, 190–197. doi:10.1016/j.chb.2014.03.040
- 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, 107–116. doi:10.1080/00049530802105865
- Linden, M. (2008, February 22). Key economic metrics through January 2008 [Msg 1]. Official Second Life Blog. Message posted to <https://community.secondlife.com/t5/Features/Key-Economic-Metrics-through-January-2008/ba-p/611953>
- Linden Lab. (2013). Second Life celebrates 10-year anniversary [Press release]. Retrieved from www.lindenlab.com/releases/second-life-celebrates-10-year-anniversary
- 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). Piscataway: IEEE Press.
- 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., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44, 1–23. doi:10.3758/s13428-011-0124-6
- Mason, W., & 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: ACM Press.
- McLeod, P., Liu, Y., & Axline, J. (2014). When your Second Life comes knocking: Effects of personality on changes to real life from virtual world experiences. *Computers in Human Behavior*, 39, 59–70. doi:10.1016/j.chb.2014.06.025
- Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, 35, 614–620. doi:10.3758/BF03195541
- 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, 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, 1044–1045. doi:10.1093/beheco/arl107
- 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, 71–81.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. doi:10.1126/science.aac4716
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, 23, 184–188. doi:10.1177/0963721414531598
- Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5, 411–419.
- Perneger, T. (1998). What's wrong with Bonferroni adjustments. *BMJ*, 316, 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, 450–461.
- Pontin, J. (2007, March 25). Artificial intelligence, with help from the humans. *The New York Times*. Retrieved August 4, 2013, from www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html
- Preckel, F., & Thiemann, H. (2003). Online- versus paper-pencil version of a high potential intelligence test. *Swiss Journal of Psychology*, 62, 131–138. doi:10.1024/1421-0185.62.2.131
- Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the Internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior*, 6, 73–80. doi:10.1089/109493103321167983
- Rogers, J. L., & Howard, K. I. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113, 553–565.
- Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers? Shifting demographics in Mechanical Turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). New York, NY: ACM Press.
- Rothman, K. (1990). No adjustments are needed for multiple comparisons. *Epidemiology*, 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., Jr. (2013). *Using Facebook as a subject recruitment tool for survey-experimental research*. Unpublished manuscript. doi:10.2139/ssrn.2101458
- Schuurmann, 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, 657–680. doi:10.1007/BF01068419
- Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, 1, 213–220. doi:10.1177/2167702612469015
- Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and Adobe Flash. *Behavior Research Methods*, 46, 95–111. doi:10.3758/s13428-013-0345-y
- 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(51876), 1–5. doi:10.1371/journal.pone.0051876
- Sproule, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, 43, 155–167. doi:10.3758/s13428-010-0039-7
- 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 and Behavior*, 36, 1051–1064.
- Stewart, N., Ungemach, C., Harris, A. J. L., Bartels, D. M., Newell, B. R., Paolacci, G., & Chandler, J. (2015). The average laboratory samples a population of 7,300 Amazon Mechanical Turk workers. *Judgment and Decision Making*, 10, 479–491.
- Swicegood, J., & Haque, S. (2015). Lessons from recruiting Second Life users with chronic medical conditions: Applications for health communications. *Journal for Virtual Worlds Research*, 8. doi:10.4101/jvwr.v8i1.7097
- Tawa, J., Negrón, R., Suyemoto, K. L., & Carter, A. S. (2015). The effect of resource competition on Blacks' and Asians' social distance using a virtual world methodology. *Group Processes and Intergroup Relations*, 18, 761–777. doi:10.1177/1368430214561694
- 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, 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, 272–277.
- Walker, E., & Nowacki, A. S. (2010). Understanding equivalence and noninferiority testing. *Journal of General Internal Medicine*, 26, 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, 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). doi:10.1111/j.1083-6101.2005.tb00259.x
- Yee, N., Bailenson, J. N., Urbanek, M., Chang, F., & Merget, D. (2007). The unbearable likeness of being digital: The persistence of nonverbal social norms in online virtual environments. *CyberPsychology & Behavior*, 10, 115–121. doi:10.1089/cpb.2006.9984