Personality and Individual Differences 75 (2015) 47–52

Contents lists available at ScienceDirect

Personality and Individual Differences

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

# Crowdsourcing and personality measurement equivalence: A warning about countries whose primary language is not English

Jennifer Feitosa [a,*], Dana L. Joseph [b,1], Daniel A. Newman [c,2]

[a] University of Central Florida, Institute for Simulation and Training, 3100 Technology Parkway, Orlando, FL 32826, United States
[b] University of Central Florida, Department of Psychology, 4000 Central Florida Blvd., Orlando, FL 32816, United States
[c] University of Illinois at Urbana-Champaign, Department of Psychology, 603 East Daniel Street, Champaign, IL 61820, United States

ARTICLE INFO

*Article history:*
Received 9 September 2014
Received in revised form 30 October 2014
Accepted 3 November 2014

*Keywords:*
Crowdsourcing
Measurement
Data collection techniques
Survey methods
Invariance testing

ABSTRACT

In the search to find cheaper, faster approaches for data collection, crowdsourcing methods (i.e., online labor portals that allow independent workers to complete surveys for compensation) have risen in popularity as a tool for personality researchers, despite a lack of evidence regarding the equivalence of crowdsourcing with traditional data collection methods. The purpose of this study was to evaluate crowdsourcing as a data collection tool by examining the measurement equivalence of crowdsourced data (i.e., from Amazon.com's MTurk) with more traditional samples (i.e., an undergraduate sample and a sample of organizational employees). Our results (using a popular measure of Big Five personality) provided evidence of measurement equivalence across all three samples, with one important exception: crowdsourced data (from MTurk) only exhibited measurement invariance with traditional data collection methods when responses were restricted to participants from native-English speaking countries. Although MTurk appears to be an easy, cost-effective data collection tool, our results suggest that MTurk data are similar to traditionally-collected data only when the MTurk sample is restricted to IP addresses from English-speaking countries.



## 1. Introduction

As part of the growing trend toward the use of faster and cheaper data collection methods, more studies are using the internet as a means to efficiently reach out to large, online samples via "crowdsourcing." This method involves the recruitment of participants using specific websites (e.g., Amazon.com's Mechanical Turk [MTurk]) that are designed to administer surveys/tasks to a global population of independent workers, who are then paid for their participation. Our knowledge of the equivalence of these crowdsourcing methods with more traditional data collection methods (e.g., undergraduate students, field samples) appears limited, and it is unclear whether participants recruited via crowdsourcing interpret a given personality measure in a conceptually similar manner to participants recruited via traditional data collection techniques. The current study investigates this issue by examining the measurement equivalence/invariance (ME/I) of a Big Five personality measure across three samples, including a crowdsourced sample, an undergraduate sample, and a field/employee sample.

### 1.1. Crowdsourcing

Crowdsourcing has been defined as "the paid recruitment of an online, independent global workforce for the objective of working on a specifically defined task or set of tasks" (Behrend, Sharek, Meade, & Wiebe, 2011, p. 801). MTurk, the most inexpensive and commonly used crowdsourcing website, allows researchers from all over the world to post surveys which "workers" (i.e., anyone who has set up an MTurk account through Amazon.com, which only requires an email address) complete in return for compensation (a September 2014 search of MTurk showed 1798 available surveys with 53% payment below $0.50; see also Sun, Wang, & Peng, 2011). Researchers using MTurk are allowed to withhold compensation from survey respondents who provide low-quality data, thus allowing researchers to collect presumably high-quality data quickly (e.g., the current study collected 754 participants with a payment of $0.25 each in 17 days). Some scholars have even concluded, based on high internal consistency and 3-week retest reliability estimates, that the psychometric quality of data from MTurk seems to meet or exceed standards typically found in published

* Corresponding author. Tel.: +1 (407) 882 1329; fax: +1 (407) 882 1550.
*E-mail addresses:* jfeitosa@ist.ucf.edu (J. Feitosa), dana.joseph@ucf.edu (D.L. Joseph), d5n@uiuc.edu (D.A. Newman).
[1] Tel.: +1 (409) 823 3912; fax: +1 (407) 823 5862.
[2] Tel.: +1 (217) 244 2512; fax: +1 (217) 244 5876.

http://dx.doi.org/10.1016/j.paid.2014.11.017
0191-8869/

research (Buhrmester, Kwang, & Gosling, 2011). However, it is worth noting that in order to ensure high-quality data, researchers may have to include quality control questions, additional bonus payments for quality responses, and reminders that data will be reviewed prior to payment in order to counteract participants' inclination to underestimate the importance of the study (Barger & Sinar, 2011).

Evidence has started to accumulate regarding the relative similarity between web-based/crowdsourced samples and more traditional methods of data collection (Gosling, Vazire, Srivastava, & John, 2004; Sprouse, 2011). Specifically, MTurk samples seem to be more diverse, younger, and more educated than the general population (Chandler, Mueller, & Paolacci, 2014), and are more diverse, older, and have more work experience than undergraduate samples (Behrend et al., 2011). In addition, Behrend et al. (2011) have demonstrated measurement equivalence of an MTurk sample and an undergraduate sample. It is important to note that the MTurk sample collected in Behrend et al.'s study was ethnically similar to their undergraduate sample (i.e., both samples were approximately 80% Caucasian), and is uncharacteristic of most MTurk samples, which tend to greatly over-sample participants from India (Walsh, 2011). Therefore, although Behrend et al. helpfully demonstrated that ME/I exists between crowdsourced and undergraduate samples, their crowdsourced sample is not representative of the type of sample typically collected through MTurk. Thus, one important question that remains is whether ME/I exists between traditional samples and typical MTurk samples that are geographically and linguistically diverse.

Moreover, Behrend et al. only examined ME/I between crowdsourced samples and undergraduate samples, leaving the extent to which crowdsourced samples display ME/I with field samples (e.g., employee samples) unknown. In the current study, we seek to investigate the ME/I of crowdsourced samples against not only undergraduate samples, but also against field/employee samples, and using a crowdsourced sample that is more representative of the typical (geographically diverse) MTurk sample.

### 1.2. Measurement equivalence/invariance

ME/I can be defined as similarity in the conceptualization of a given construct across two or more groups (Meade, Johnson, & Braddy, 2008; Vandenberg & Lance, 2000). When ME/I is established, it does not refer to subgroup differences in means or shapes of distributions, but it rather refers to similarity in the relationships between the observable and latent variables across the subgroups (Drasgow, 1984). If ignored, measurement *non*equivalence can lead to erroneous conclusions. For example, if a researcher collected an initial sample of undergraduates and then collected a second sample using MTurk in order to quell concerns that the findings might not replicate beyond the undergraduate sample, any comparisons across these groups would be limited if ME/I were not first established.

There are two primary methods of establishing ME/I: Mean and Covariance Structure Analysis (MACS) and Item-Response Theory (IRT). For the purposes of this study, we will apply the MACS method, because researchers have identified MACS as the most adequate way to assess measurement invariance across subgroups when the measure at hand is multidimensional (Meade & Lautenschlager, 2004) and when sample size varies between 500 and 1000 (Stark, Chernyshenko, & Drasgow, 2006), as in the current study. Consequently, we will follow recommendations from Vandenberg and Lance (2000) for best practices in ME/I evaluation. Ultimately, we seek to examine the ME/I of a popular measure of Big Five personality across an MTurk sample, an undergraduate sample, and a field sample; in hopes of testing the extent to which participants interpret the personality measure in a similar conceptual manner across the three data collection methods.

## 2. Method

### 2.1. Participants and procedure

In order to examine personality ME/I across modes of data collection, we administered the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) to three samples: an MTurk (crowdsourced) sample, a field sample, and an undergraduate sample. Consistent with past studies that compared MTurk samples against more traditional samples (see Berinsky, Huber, & Lenz, 2012), we found statistically significant demographic differences across data sources (see Table 1). Specifically, females, Whites, older and less educated participants were overrepresented in the field; younger adults were overrepresented in undergraduate samples; and Asians were overrepresented in MTurk samples.

#### 2.1.1. MTurk sample

754 participants were recruited through the MTurk website. In order to collect a sample that represents a typical MTurk sample, the default MTurk setting of surveying only participants who have a 95% approval rate (i.e., 95% of their prior survey results have been approved for payment by the researcher) was used in the current study (consistent with prior MTurk research practices; Barger & Sinar, 2011). No additional participant requirements were specified (i.e., MTurk default settings). The current data collection involved the use of two quality control items (e.g., "For quality control purposes, please enter 'D' as the answer to this question"), with one quality control question placed at the middle, and one at the end, of the survey (as recommended by Barger & Sinar, 2011). Participants who failed to correctly answer both items were excluded from analysis. The final sample consisted of 687 adults. Participants who correctly answered both quality control items were paid $0.25 for participation in the survey (this price was set to be consistent with surveys of similar length posted on MTurk).

#### 2.1.2. Field/employee sample

The field sample consisted of 263 employees recruited from multiple organizations located in the Midwest, including healthcare employees, administrative staff, supervisors, property managers, and custodial staff. Participants were paid for their participation with a $10 Amazon.com gift card, and all surveys were completed online.

#### 2.1.3. Undergraduate sample

The undergraduate sample consisted of 233 undergraduate students from a Southwestern university, who were recruited through the psychology department's experiment participation system.

### 2.2. Measures

#### 2.2.1. Big Five Inventor

Participants completed the BFI, a 44-item measure designed to assess extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience (5-point Likert scale: 1 (*strongly disagree*) to 5 (*strongly agree*). This measure has shown similar psychometric properties across languages (e.g., Benet-Martinez & John, 1998; Denissen, Geenen, van Aken, Gosling, & Porter, 2008) and educational levels (Rammstedt, Goldberg, & Borg, 2010). Cronbach's alphas and intercorrelations among the five dimensions for each sample are presented in Table 2. Example items include "I see myself as someone who…" "Is outgoing, sociable," (extraversion) and "Does things efficiently" (conscientiousness).

**Table 1**
Sample demographics.

| Demographics | Sample | | | $\chi^2$ across samples |
|---|---|---|---|---|
| | MTurk $N = 687$ | Field $N = 263$ | Undergraduate $N = 233$ | |
| Sex | | | | 186.358* |
| Female | 38.7% | 85.6% | 65.7% | |
| Male | 60.7% | 14.4% | 30.5% | |
| Race/ethnicity | | | | 606.788* |
| African American | 2.9% | 2.7% | 12.9% | |
| Asian | 66.1% | 1.5% | 5.2% | |
| White | 23.7% | 89.7% | 62.2% | |
| Hispanic | 0.9% | 3.0% | 13.7% | |
| Other | 5.8% | 2.3% | 2.6% | |
| Age | | | | 443.519* |
| Under 30 | 68.3% | 16.0% | 87.1% | |
| 30–39 | 21.7% | 24.3% | 6.0% | |
| 40–49 | 5.8% | 20.5% | 2.1% | |
| 50–59 | 2.5% | 28.9% | 0.4% | |
| Over 60 | 1.5% | 9.1% | 0.4% | |
| Education | | | | 108.571* |
| Some high school | 2.0% | 0% | ‡ | |
| High school degree | 11.8% | 24.3% | ‡ | |
| Associate's degree | 8.0% | 25.9% | ‡ | |
| Bachelor's degree | 51.7% | 37.6% | ‡ | |
| Master's degree | 24.7% | 8.0% | ‡ | |
| Doctoral degree | 0.9% | 1.9% | ‡ | |

* $p < .05$.
‡ We did not record whether undergraduate participants had completed an Associate's degree, although all undergraduate participants had completed a high school degree.

**Table 2**
Intercorrelations across dimensions of the Big Five Inventory.

| Variable | *M* | *SD* | 1. | 2. | 3. | 4. | 5. |
|---|---|---|---|---|---|---|---|
| MTurk sample | | | | | | | |
| 1. Extraversion | 3.22 | .61 | (.71) | | | | |
| 2. Agreeableness | 3.56 | .60 | .23 | (.68) | | | |
| 3. Conscientiousness | 3.57 | .61 | .36 | .58 | (.76) | | |
| 4. Neuroticism | 2.83 | .67 | −.39 | −.47 | −.53 | (.81) | |
| 5. Openness to experience | 3.53 | .49 | .30 | .39 | .42 | −.16 | (.61) |
| Field sample | | | | | | | |
| 1. Extraversion | 3.51 | .83 | (.87) | | | | |
| 2. Agreeableness | 4.21 | .50 | .23 | (.68) | | | |
| 3. Conscientiousness | 4.26 | .52 | .26 | .54 | (.78) | | |
| 4. Neuroticism | 2.40 | .75 | −.34 | −.42 | −.41 | (.86) | |
| 5. Openness to experience | 3.63 | .63 | .48 | .18 | .31 | −.25 | (.83) |
| Undergraduate sample | | | | | | | |
| 1. Extraversion | 3.23 | .78 | (.89) | | | | |
| 2. Agreeableness | 3.81 | .62 | .27 | (.79) | | | |
| 3. Conscientiousness | 3.69 | .62 | .12 | .43 | (.81) | | |
| 4. Neuroticism | 2.84 | .64 | −.35 | −.29 | −.31 | (.80) | |
| 5. Openness to experience | 3.60 | .58 | .14 | .29 | .30 | −.09 | (.78) |

*Note.* MTurk sample: $N = 687$; Field sample: $N = 263$; Undergraduate sample: $N = 233$. Cronbach's $\alpha$ in diagonal; $M$ = mean, $SD$ = standard deviation.

## 3. Results and discussion

Before conducting ME/I analyses, we first performed a confirmatory factor analysis (CFA) on the BFI to confirm the oblique five-factor structure. Item parcels were created (items were randomly assigned to three parcels per factor; Landis, Beal, & Tesluk, 2000) to test the oblique five-factor structure of the BFI, due to known problems fitting item-level indicators (Hopwood & Donnellan, 2010; Joseph & Newman, 2010). In order to set the scale for the latent factors, one loading per factor was set to 1.0 (to select the five referent parcels, we used parcels with the most similar loadings across samples, as determined from a CFA specified with all loadings free and the variance of each latent variable fixed to 1.0). LISREL 8.80 was used to estimate all models. The five factor solution fit adequately in all samples (undergraduate sample: CFI = .933, RMSEA = .092; field sample: CFI = .973, RMSEA = .063; MTurk sample: CFI = .970, RMSEA = .064), and all parcels loaded strongly onto the intended factors (all loadings $\geqslant$.56, with the exception of an openness parcel within the MTurk sample for which the loading was .30; see Table 3 for CFA results).

ME/I analyses were then performed to assess the following sample comparisons: (a) undergraduate vs. field, (b) field vs. MTurk, and (c) undergraduate vs. MTurk. Following Vandenberg and Lance's (2000) measurement invariance procedures, for each of the three sample comparisons, we first assessed *configural invariance* to determine whether the same pattern of factor loadings was obtained in both samples. After configural invariance was established, we constrained the factor loadings to be of equal magnitude across samples as a test of *metric invariance*. If metric invariance was confirmed, scalar invariance could next be assessed

**Table 3**
Confirmatory factor analysis of the Big Five Inventory.

| Observed variable | Undergraduate sample ($N$ = 233) | | | | Field sample ($N$ = 263) | | | | MTurk sample ($N$ = 687) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Factor loadings | | | | Factor loadings | | | | Factor loadings | | | |
| Extraversion 1 | .95 | | | | .94 | | | | .86 | | | |
| Extraversion 2 | .94 | | | | .95 | | | | .95 | | | |
| Extraversion 3 | .80 | | | | .74 | | | | .56 | | | |
| Agreeableness 1 | .65 | | | | .68 | | | | .66 | | | |
| Agreeableness 2 | .78 | | | | .67 | | | | .57 | | | |
| Agreeableness 3 | .82 | | | | .74 | | | | .81 | | | |
| Conscientiousness 1 | .82 | | | | .60 | | | | .75 | | | |
| Conscientiousness 2 | .73 | | | | .76 | | | | .69 | | | |
| Conscientiousness 3 | .75 | | | | .83 | | | | .76 | | | |
| Neuroticism 1 | .81 | | | | .84 | | | | .69 | | | |
| Neuroticism 2 | .75 | | | | .83 | | | | .72 | | | |
| Neuroticism 3 | .67 | | | | .73 | | | | .72 | | | |
| Openness 1 | .75 | | | | .79 | | | | .78 | | | |
| Openness 2 | .91 | | | | .82 | | | | .77 | | | |
| Openness 3 | .58 | | | | .72 | | | | .30 | | | |
| Factor intercorrelations | | | | | | | | | | | | |
| Extraversion | | | | | | | | | | | | |
| Agreeableness | .38 | | | | .31 | | | | .39 | | | |
| Conscientiousness | .14 | .55 | | | .32 | .70 | | | .46 | .76 | | |
| Neuroticism | −.39 | −.34 | −.40 | | −.36 | −.57 | −.51 | | −.48 | −.61 | −.69 | |
| Openness | .23 | .43 | .41 | −.12 | .50 | .28 | .40 | −.28 | .45 | .60 | .63 | −.31 |
| Fit indices | | | | | | | | | | | | |
| $\chi^2$ (df) | 222.86 (80) | | | | 147.06 (80) | | | | 156.28 (80) | | | |
| RMSEA/SRMR | .092/.073 | | | | .063/.052 | | | | .064/.058 | | | |
| TLI/CFI | .912/.933 | | | | .965/.973 | | | | .961/.970 | | | |

*Note.* Completely standardized solution; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; TLI = Tucker–Lewis index; CFI = comparative fit index.

by constraining the indicator intercepts to be equal across samples. Comparisons of model fit between nested models were made using indices of practical fit (rather than using change in $\chi^2$ between models, which has strong dependence on sample size and thus signals trivial fit departures in large samples), and following the recommendation to interpret ΔCFI > .01 (Cheung & Rensvold, 2002) as indicating non-equivalence.

Results of the ME/I analyses are presented in Table 4. Regarding ME/I across traditional data collection techniques (i.e., field sample vs. undergraduate sample), our results show support for configural, metric, and scalar equivalence. Configural invariance was established by the acceptable fit of Model 1 (CFI = .953, RMSEA = .081, TLI = .938, SRMR = .073), suggesting undergraduates and employees have similar frames of reference when responding to the Big Five Inventory (see Riordan & Vandenberg, 1994). The metric invariance model also showed adequate fit (CFI = .950, RMSEA = .081, TLI = .938, SRMR = .081) and did not result in significantly worse fit in comparison to the configural model (ΔCFI = −.003). In the subsequent test of scalar invariance, Model 3 demonstrated adequate fit (CFI = .947, RMSEA = .081, TLI = .938, SRMR = .082), and the model did not notably worsen the fit compared to the metric invariance model (ΔCFI = −.003). These results suggest a finding of strong ME/I between undergraduate and employee/field samples (Vandenberg & Lance, 2000); a finding which lends support to (or at least, is not inconsistent with) the use of undergraduate samples as proxies for field samples in psychological research on personality measurement.

Upon comparing the field sample against the MTurk sample, our results supported configural invariance (RMSEA = .080, TLI = .941, CFI = .955, SRMR = .057), but the data did not support metric invariance (ΔCFI = −.011, RMSEA = .086, TLI = .931, CFI = .944, SRMR = .068). Cheung and Rensvold (2002) advise that ΔCFI > .01

**Table 4**
BFI measurement equivalence across field, undergraduate, and MTurk samples.

| Measurement equivalence model | $\chi^2$ | *df* | RMSEA (90% CI) | TLI (NNFI) | CFI | SRMR | ΔCFI |
|---|---|---|---|---|---|---|---|
| *Field vs. undergraduate* | | | | | | | |
| 0. Invariant covariance matrices | 217.120 | 120 | .058 (.045–.070) | 0.968 | 0.982 | 0.080 | – |
| 1. Configural invariance | 402.674 | 160 | .081 (.071–.091) | 0.938 | 0.953 | 0.073 | – |
| 2. Metric invariance | 428.882 | 170 | .081 (.072–.091) | 0.938 | 0.950 | 0.081 | −0.003 |
| 3. Scalar invariance | 452.681 | 180 | .081 (.072–.090) | 0.938 | 0.947 | 0.082 | −0.003 |
| *Field vs. MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 428.761 | 120 | .074 (.070–.082) | 0.949 | 0.971 | 0.063 | – |
| 1. Configural invariance | 641.867 | 160 | .080 (.074–.090) | 0.941 | 0.955 | 0.057 | – |
| 2. Metric invariance* | 763.688 | 170 | .086 (.080–.093) | 0.931 | 0.944 | 0.068 | −0.011 |
| *Undergraduate vs. MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 352.881 | 120 | .065 (.057–.073) | 0.959 | 0.977 | 0.041 | – |
| 1. Configural invariance | 688.667 | 160 | .085 (.079–.092) | 0.931 | 0.947 | 0.057 | – |
| 2. Metric invariance* | 546.783 | 170 | .098 (.089–.107) | 0.902 | 0.920 | 0.075 | −0.027 |

* Invariance not supported (Cheung & Rensvold, 2002; ΔCFI ⩽ .01 indicates measurement equivalence).

signals measurement nonequivalence. For diagnostic purposes, we next note that the BFI parcel factor loadings that seemed to be the source of the metric misfit appeared to originate with four parcels: extraversion 3, agreeableness 2, neuroticism 1, and openness to experience 3; which had lower factor loadings in the MTurk sample, in comparison to the field sample (field $\lambda$ = .738, .668, .844, .696 respectively; MTurk $\lambda$ = .508, .548, .679, .269, respectively).

Similarly, upon comparing the undergraduate versus the MTurk sample, our results show support for configural invariance (RMSEA = .085, TLI = .931, CFI = .947, SRMR = .057), but not metric invariance ($\Delta$CFI = −.027, RMSEA = .098, TLI = .902, CFI = .920, SRMR = .075). The BFI parcel factor loadings that seemed to be the source of the metric misfit appeared to originate with four parcels: extraversion 3, agreeableness 2, openness 2, and openness 3; which had lower factor loadings in the MTurk sample (undergraduate $\lambda$ = .798, .782, .913, .576, respectively; MTurk $\lambda$ = .526, .541, .746, .283, respectively). To summarize, even though the undergraduate and field/employee samples displayed ME/I with each other, the MTurk sample exhibited evidence of metric nonequivalence with both the undergraduate and field samples, suggesting that when measuring personality, MTurk respondents are not interpreting the same constructs in the same manner as are undergraduate and field/employee samples.

After consistently finding evidence of nonequivalence of MTurk samples versus samples from traditional data collection techniques, we sought to explain potential reasons why our results differed from Behrend et al. (2011), who largely found strong ME/I between MTurk and an undergraduate sample). One key difference between our design and Behrend et al.'s design was that, in contrast to Behrend, 60% of our MTurk sample used IP addresses that indicated the participants were from India, where participants likely do not speak English as a first language. As a result, we decided to divide the MTurk sample into separate sub-samples to see whether the nonequivalence could be due to potential language barriers. Therefore, we separated the MTurk sample – a similar dichotomization method used by Lyness and Kropf (2007) – into two sub-samples: native English speakers (i.e., U.S. and United Kingdom participants) and non-native English speakers (i.e., all other countries; of whom 81% were Indian).

In order to test the assumption that non-native English speakers were driving the nonequivalence of the MTurk sample, we first evaluated ME/I across English speaking versus non-English speaking participants from the MTurk sample (Table 5). As expected, the comparison of English and non-native English speaking MTurk samples only showed evidence of configural invariance (CFI = .945, RMSEA = .087, TLI = .928, SRMR = .063), and no support for metric invariance ($\Delta$CFI = −.257, CFI = .688, RMSEA = .206, TLI = .614, SRMR = .422). Consequently, we decided to re-run our previous ME/I analyses comparing the MTurk sample with undergraduate and field samples, by conducting four separate sets of analyses. In particular, we assessed ME/I of the English-speaking MTurk sample versus: (a) undergraduate and (b) field samples; as well as ME/I of the non-native English-speaking MTurk sample vs. (c) undergraduate and (d) field samples.

Consistent with our rationale that a language barrier might explain the observed metric nonequivalence in MTurk samples, the results in Table 5 show nonequivalence of MTurk samples *only when the sample includes non-native English speakers*. When the MTurk sample consisted only of native English speakers, the MTurk samples showed evidence of measurement equivalence with both the field and undergraduate samples. Therefore, it appears that MTurk samples may display ME/I with traditional data collection techniques (undergraduate and field samples) *only when the MTurk sample is restricted to respondents with IP addresses from native-English speaking countries*.

## 4. Conclusion

Considering the importance of assessing ME/I prior to estimating between-groups differences (e.g., gender differences in personality, longitudinal change in personality; Meade et al., 2008; Vandenberg & Lance, 2000), our findings provide an optimistic outlook regarding the similarity of MTurk crowdsourcing to more traditional methods (i.e., undergraduate college student and field/

**Table 5**
BFI measurement equivalence of native English and non-native English-speaking MTurk samples.

| Measurement equivalence model | $\chi^2$ | *df* | RMSEA (90% CI) | TLI (NNFI) | CFI | SRMR | $\Delta$CFI |
|---|---|---|---|---|---|---|---|
| *English-speaking MTurk vs. non-native English-speaking MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 310.543 | 120 | .068 (.059–.078) | 0.956 | 0.975 | 0.047 | – |
| 1. Configural invariance | 575.317 | 160 | .087 (.080–.095) | 0.928 | 0.945 | 0.063 | – |
| 2. Metric invariance* | 2549.319 | 170 | .206 (.196–.210) | 0.614 | 0.688 | 0.422 | −0.257 |
| *Field vs. English-speaking MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 214.036 | 120 | .085 (.067–.104) | 0.945 | 0.968 | 0.120 | – |
| 1. Configural invariance | 316.283 | 160 | .095 (.080–.111) | 0.931 | 0.948 | 0.077 | – |
| 2. Metric invariance | 332.972 | 170 | .094 (.079–.109) | 0.933 | 0.945 | 0.086 | −0.003 |
| 3. Scalar invariance | 363.727 | 180 | .097 (.083–.112) | 0.928 | 0.938 | 0.086 | −0.007 |
| *Field vs. non-native English-speaking MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 442.284 | 120 | .102 (.092–.112) | 0.909 | 0.948 | 0.134 | – |
| 1. Configural invariance | 600.356 | 160 | .081 (.075–.088) | 0.937 | 0.952 | 0.063 | – |
| 2. Metric invariance* | 757.981 | 170 | .091 (.085–.098) | 0.920 | 0.936 | 0.078 | −0.016 |
| *Undergraduate vs. English-speaking MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 202.760 | 120 | .080 (.061–.100) | 0.940 | 0.966 | 0.135 | – |
| 1. Configural invariance | 398.514 | 160 | .094 (.082–.106) | 0.918 | 0.937 | 0.073 | – |
| 2. Metric invariance | 411.820 | 170 | .092 (.081–.103) | 0.921 | 0.936 | 0.073 | −0.001 |
| 3. Scalar invariance | 446.816 | 180 | .094 (.083–.105) | 0.918 | 0.930 | 0.074 | −0.006 |
| *Undergraduate vs. non-native English-speaking MTurk* | | | | | | | |
| 0. Invariant covariance matrices | 442.631 | 120 | .082 (.074–.090) | 0.933 | 0.962 | 0.056 | – |
| 1. Configural invariance | 526.326 | 160 | .099 (.090–.109) | 0.909 | 0.931 | 0.073 | – |
| 2. Metric invariance* | 615.542 | 170 | .106 (.097–.115) | 0.896 | 0.916 | 0.079 | −0.015 |

* Invariance not supported (Cheung & Rensvold, 2002; $\Delta$CFI $\leqslant$ .01 indicates measurement equivalence). English-speaking MTurk sample: $N$ = 109; Non-native English speaking MTurk sample: $N$ = 578; Field sample: $N$ = 263; Undergraduate sample: $N$ = 233.

employee samples), *as long as the crowdsourced sample is restricted to participants from native English-speaking countries only*. These results can hopefully continue to decrease some of the skepticism surrounding crowdsourcing, while identifying a potential limitation of crowdsourcing that has not been previously identified using psychometric methods. Thus, this study begins to concretize prior speculations that crowdsourcing can at times be just as good as traditional samples – with an important caveat/restriction about respondents' first languages (i.e., countries of origin). Given that MTurk default settings collect data from multinational samples of respondents, this caveat (i.e., that MTurk samples only display personality measurement equivalence with traditional samples when they are collected from native English-speaking countries) appears to be of critical importance. Moreover, several recent forums/listserves (i.e., RMNET; http://turkrequesters.blogspot.com; http://www.behind-the-enemy-lines.com) have suggested that even if an MTurk researcher uses the advanced settings in MTurk to restrict the sample to U.S. participants, 5–10% of the sample may still be non-U.S., as indicated by IP addresses (i.e., it appears that non-U.S. participants have discovered several loopholes to the U.S. participant restriction, including registering for an Amazon.com account with U.S. social security numbers of deceased individuals). In other words, due to the default MTurk settings and recent reports that U.S.-restricted MTurk samples are not actually U.S.-only samples, our findings illuminate critical steps that should be taken in the MTurk data collection process: one should (a) remove participants from non-native English speaking countries on the basis of IP addresses, or alternatively (b) establish measurement equivalence across native and non-native English speakers before including these participants in the subsequent analyses.

Nonetheless, the current study has several limitations. First, we caution that additional research is needed (particularly using measures other than the BFI) before we can provide a universal endorsement of, or warning against, particular crowdsourcing methods. Recent measurement equivalence research on a variety of different psychological measures has provided promising results in this regard (Behrend et al., 2011; who compared a 93% U.S. and Canadian MTurk sample against a U.S. undergraduate sample to exhibit good evidence of ME/I). Further, our current index of whether a respondent hails from a primarily-English-speaking country (i.e., we inferred this from the IP addresses used on MTurk) is admittedly imperfect. It is entirely possible that some of the respondents with IP addresses from India might have indeed been native English speakers. On the other hand, the use of IP addresses is a potentially useful general method for MTurk data screening, because all MTurk users have direct access to this information.

Furthermore, our proscription against the use of non-native English-speaking MTurk subsamples might be limiting for researchers seeking to conduct cross-cultural research using MTurk. Indeed, cross-cultural research is one of the primary areas where the need to establish measurement equivalence across languages is most important (see Hui & Triandis, 1985; Nye, Roberts, Saucier, & Zhou, 2008; Steenkamp & Baumgartner, 1998). In essence, our current plea – for future researchers to establish ME/I for non-native English IP addresses in MTurk samples – is really just a natural implication of the aforementioned work on cross-cultural ME/I, and serves to reinforce past advice that ME/I must be established, rather than assumed, in multinational samples (Schaffer & Riordan, 2003).

## Acknowledgement

The undergraduate sample used in the current study is the same sample as that of Nye, Newman, & Joseph (2010) and Joseph and Newman (2010).

## References

Barger, P. B., & Sinar, E. F. (April, 2011). *Psychological data from Amazon.com's MTurk: Rapid and inexpensive – but high-quality?* Presented at 26th Annual SIOP Convention, Chicago, IL.

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

Benet-Martinez, V., & John, O. P. (1998). *Los Cinco Grandes* across cultures and ethnic groups: Multitrait–multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75*(3), 729.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. Political Analysis, 20(3), 351–368. *Human Relations, 61*(8), 1139–1160.

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*(1), 3–5.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods, 46*(1), 112–130.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255.

Denissen, J. J., Geenen, R., Van Aken, M. A., Gosling, S. D., & Potter, J. (2008). Development and validation of a Dutch translation of the Big Five Inventory (BFI). *Journal of Personality Assessment, 90*(2), 152–157.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. *Psychological Bulletin, 95*(1), 134–135.

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.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332–346.

Hui, C. H., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131–152.

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

Joseph, D. L., & Newman, D. A. (2010). Discriminant validity of self-reported emotional intelligence: A multitrait-multisource study. *Educational and Psychological Measurement, 70*(4), 672–694.

Landis, R. S., Beal, D. J., & Tesluk, P. E. (2000). A comparison of approaches to forming composite measures in structural equation models. *Organizational Research Methods, 3*(2), 186–207.

Lyness, K. S., & Kropf, M. B. (2007). Cultural values and potential nonresponse bias: A multilevel examination of cross-national differences in mail survey response rates. *Organizational Research Methods, 10*(2), 210–224.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568–592.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7*(4), 361–388.

Nye, C. D., Newman, D. A., & Joseph, D. L. (2010). Never say "always"? Extreme item wording effects on scalar invariance and item response curves. *Organizational Research Methods, 13*(4), 806–830.

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42*(6), 1524–1536.

Rammstedt, B., Goldberg, L. R., & Borg, I. (2010). The measurement equivalence of Big-Five factor markers for persons with different levels of education. *Journal of Research in Personality, 44*(1), 53–61.

Riordan, C. M., & Vandenberg, R. J. (1994). A central question in cross-cultural research: Do employees of different cultures interpret work-related measures in an equivalent manner? *Journal of Management, 20*(3), 643–671.

Schaffer, B. S., & Riordan, C. M. (2003). A review of cross-cultural methodologies for organizational research: A best-practice approach. *Organizational Research Methods, 6*(2), 169–215.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43*(1), 155–167.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292–1306.

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*(1), 78–107.

Sun, Y., Wang, N., & Peng, Z. (2011). Working for one penny: Understanding why people would like to participate in online tasks with low payment. *Computers in Human Behavior, 27*(2), 1033–1041.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–69.

Walsh, B. (2011). Pennies for your thoughts. *Time, 177*(4), 55–56.