DOI 10.3758/s13428-011-0081-0

The viability of crowdsourcing for survey research

Tara S. Behrend · David J. Sharek · Adam W. Meade · Eric N. Wiebe

Published online: 25 March 2011
© Psychonomic Society, Inc. 2011

Abstract Online contract labor portals (i.e., crowdsourcing) have recently emerged as attractive alternatives to university participant pools for the purposes of collecting survey data for behavioral research. However, prior research has not provided a thorough examination of crowdsourced data for organizational psychology research. We found that, as compared with a traditional university participant pool, crowdsourcing respondents were older, were more ethnically diverse, and had more work experience. Additionally, the reliability of the data from the crowdsourcing sample was as good as or better than the corresponding university sample. Moreover, measurement invariance generally held across these groups. We conclude that the use of these labor portals is an efficient and appropriate alternative to a university participant pool, despite small differences in personality and socially desirable responding across the samples. The risks and advantages of crowdsourcing are outlined, and an overview of practical and ethical guidelines is provided.

Keywords Crowdsourcing · Mechanical Turk · Industrial/organizational psychology · Survey research · Sampling · Personality

The past decade has seen data collection in survey research migrate from paper-and-pencil measures to online surveys. Conducting survey research using online media is often more convenient and flexible, permitting the researcher to quickly and easily obtain data from a large number of participants (Truell, Bartlett, & Alexander, 2002). Other benefits include lower cost (Kraut, Olson, Banaji, Bruckman, Cohen, & Couper, 2004), fewer physical resources, simplified logistics, and the elimination of data entry errors. Initially, there was some concern over the equivalence of online data collection, as compared with in-person paper-and-pencil methods. However, there appears to be some consensus that these two approaches are largely equivalent in terms of the psychometric properties that can be expected (Cole, Bedeian, & Field, 2006; De Beuckalaer & Lievens, 2009; Meade, Michels, & Lautenschlager, 2007; Meyerson & Tryon, 2003; Stanton, 1998), as well as impression management/social desirability (Booth-Kewley, Edwards, & Rosenfeld, 1992) and data completeness (Stanton, 1998).

One limitation of this previous work, however, is that some studies have conflated differences in administration medium with differences in the population reached by using the medium. For example, Booth-Kewley et al. (1992) found differences between administration formats on an attitudes survey when a college population was used, with no differences found in a sample of professional Navy recruits. Moreover, Manfreda, Bosnjak, Berzelak, Haas, and Vehovar (2008) reported meta-analytic results of 45 studies comparing Web-based and other survey media, finding that differences in criteria such as response rate or dropout rate were dependent on the sample recruitment base in question. Online panels (i.e., pools of respondents who have agreed to be contacted for multiple survey opportunities) behaved differently than one-time respondents, generally showing a smaller difference across media. Other authors have also noted that the effects of administration medium may

T. S. Behrend (✉)
Organizational Sciences & Communication,
George Washington University,
Washington, DC, USA
e-mail: behrend@gwu.edu

D. J. Sharek · A. W. Meade · E. N. Wiebe
Psychology, North Carolina State University,
Raleigh, NC, USA

depend on the purpose and significance of the questionnaire (e.g., Ployhart, Weekley, Holtz, & Kemp, 2003; Richman, Kiesler, Weisband, & Drasgow, 1999).

These studies underscore an important and often unrealized benefit of using online media for organizational research: the potential to reach a wider and more diverse population (Barchard & Williams, 2008; Dandurand, Shultz, & Onishi, 2008). Often, survey research relies on a homogeneous sample of undergraduates from *Western*, *educated*, *industrialized*, *rich*, and *democratic* societies (WEIRD; Henrich, Heine, & Norenzayan, 2010a, 2010b). This is of special concern when the survey in question relates to job- or career-related constructs of interest to organizational researchers (Anderson, 2003; Landy, 2008; Locke, 1986; Ward, 1993). Some past work has shown that even among "particularistic" research (e.g., research that is concerned with narrowly defined independent and dependent variables), differences exist between students and working adults, and these differences can severely limit the generalizability of findings (Ward, 1993).

The recent mainstream introduction of globally-reaching Internet technologies such as *crowdsourcing* may be a solution to the limited participant pool with which researchers must sometimes work (Gosling, Sandy, John, & Potter, 2010). Previous research has begun to shed light on the general demographic makeup of Mechanical Turk workers (Ipeirotis, 2010). Additionally, research has compared the quality of data from acceptability judgment experiments between Mechanical Turk workers and in-lab participants (Sprouse, 2011). Unfortunately, few studies have examined the types of people from crowdsourcing communities that participate in organizational psychology-research. Thus, the goal of the present study is to provide a primer on the use of crowdsourcing for organizational survey research. In this article, we provide an overview of crowdsourcing, examine the demographic makeup of a crowdsourced sample, and systematically investigate the viability of crowdsourcing for providing quality data for survey research. Specifically, we compare group means from crowdsourced and university samples with respect to several commonly used measures in organizational research. We assess the quality of the data garnered from both samples by examining social desirability, reliability of scales, completion time, length of open-ended responses, and data consistency and completeness. We also compare the psychometric functioning of the measures across samples via invariance tests.

Crowdsourcing

The etymology of the term *crowdsourcing* can be traced to a Wired magazine article where the term *outsourcing* was modified to describe the recruitment of a global online workforce without the need for a traditional outsourcing company (Howe, 2006).

Although Howe (2006) did not clearly define crowdsourcing when he coined the term, he indicated that it was limited to for-profit businesses leveraging the Internet workforce. The term has been recently defined as "the intentional mobilization for commercial exploitation of creative ideas and other forms of work performed by consumers" (Kleemann, Voß, & Rieder, 2008, p. 22). On the basis of the current and emerging uses of crowdsourcing technologies, these definitions have become too narrow and should be expanded to include other uses for leveraging an independent global workforce. For example, when adventurer Steve Fossett was reported missing in 2007 after failing to return from a solo plane ride over the Sierra Nevada Mountains, an Internet-based initiative was employed where thousands of independent, unpaid workers were tasked with using recently uploaded satellite images to search for any signs of a crash site ("Turk and Rescue," 2007). Recently, academic uses of crowdsourcing such as online research studies have become increasingly popular (e.g., Kittur, Chi, & Suh, 2008; Little, Chilton, Goldman, & Miller, 2009). For example, Heilman and Smith (2010) recruited 178 participants who produced 6,000 ratings on the quality of computer-generated questions on subject matter sourced from Wikipedia articles. In less than 24 h, Sharek (2010) used a crowdsourcing service to recruit 169 people to participate in an online study that used a video game to help measure user engagement. In another study, designed to investigate how people interpret line drawings and shaded images, 560 crowdsourced participants were asked to orient 250,000 gauges onto 3-D objects (F. Cole, Sanik, DeCarlo, Finkelstein, Funkhouser, Rusinkiewicz, & Singh, 2009). In the past, similar but noncrowdsourced studies were limited to a small number of motivated participants that were willing to spend up to 12 h in order to place the large number of gauges.

For the purposes of this article, crowdsourcing is operationally 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*. The key features of this definition are that (1) workers are paid, (2) they can be recruited online from any geographic location, and (3) they are hired only to complete a defined task or set of tasks. In some ways, workers mirror undergraduate research pools in that the interaction between researcher and participant is of short duration and participants are assumed to be motivated primarily by extrinsic factors (e.g., financial compensation for workers or course credit for undergraduates).

The mechanisms by which individuals are recruited and the examples described in the literature indicate that there is reason to believe that crowdsourcing can be a vehicle for recruiting respondents that are more representative of the working adult population than is a university participant pool. Not only are large samples of participants readily available to complete surveys at a relatively low cost, but also it is likely that many of these participants will have more relevant work experience in career-oriented jobs than a typical sample composed largely of college freshmen and sophomores. A crowdsourced pool may also be more ethnically and educationally diverse. In sum, the use of crowdsourcing for organizational psychology research is a promising approach to collecting more representative samples, as compared with the commonly used undergraduate participant pool. However, the use of this crowdsourcing raises a number of important questions that need to be empirically addressed. Specifically, we investigate the following questions:

Research question 1	What are the demographic characteristics of respondents from a crowdsourcing pool? Do they differ from the characteristics of a university participant pool?
Research question 2	How does the quality of the data obtained using crowdsourcing compare with that of a university participant pool?
Research question 3	Do the psychometric properties of commonly used organizational research surveys differ across undergraduate and crowdsourcing samples?
Research question 4	Do mean differences across undergraduate and crowdsourcing samples exist with respect to personality traits and attitudes of interest to organizational researchers?
Research question 5	Why do users participate in crowdsourcing?

Mechanical Turk

The most well-known crowdsourcing Website is Amazon's Mechanical Turk (Amazon.com, 2010). It was chosen for the present study due to its growing popularity as a viable means for recruiting participants in academic research. Although initially an internal tool, Amazon's impetus for releasing the Mechanical Turk service to the public in 2005 was based on the idea that there are many tasks that people can do better than computers, such as identifying and listing objects in a photograph. Traditionally, those tasks required large-scale, costly outsourcing initiatives. Mechanical Turk provides a means for businesses or individuals (known as *requesters*) to outsource small tasks referred to as *human intelligence tasks* (HITs) to a global workforce. For example, a business launching an online shopping Website may need to provide descriptive tags for potentially millions of product images, a task difficult for computer algorithms. Rather than hire temporary employees, the business could source individuals through Mechanical Turk and pay them a few cents per image description. Recent data from the Mechanical Turk Website revealed that there were over 270,000 available HITs, ranging from US $0.01 to US $13.00 (Amazon.com, 2010).

Once a business or individual signs up for a requester account, a job request (HIT) can be completed using either a blank template or adjusting the supplied example HIT templates (e.g., Standard Survey) to suit the task. Requesters then enter a title, task description, relevant keywords, and how much money they will pay for each assignment. All of this information is supplied to the workers when they search for HITs to complete. Additionally, requesters must enter how many assignments (unique workers) they need for the HIT, expiration date for the HIT, and length of time before a worker's submission is automatically approved. Requesters can also filter workers by location (at the country level) and HIT approval rate, which reflects the typical quality of the worker's submission as indicated by previous requesters.

Individuals 18 years and older can sign up for a free worker account that allows them access to view and participate in the HITs. A worker is allowed only one account, and each worker is assigned an alphanumeric worker ID that is used to track his or her performance and payment records. Workers can search for HITs by keyword, date, compensation amount, and time allotted to complete the HIT. Additionally, workers can browse all available HITs and read task descriptions before deciding to participate. Some HITs cannot be accessed until a related qualification exam has been taken. Requesters may design these qualification exams to ensure that workers possess a certain degree of skill or proficiency before they are allowed to participate in an HIT. Generally, if a worker produces low-quality work, it is up to the discretion of the requester to reject the work and not pay the worker. If this happens, the worker's HIT approval rate is lowered, and the transaction is reflected in the requester's statistics.

Method

Participants

Two samples were collected; the first was from a traditional psychology participant pool, and the second was from Mechanical Turk. The undergraduate sample consisted of 270 undergraduate students enrolled in an entry-level psychology course at a large Southeastern research university in the U.S. Participants were compensated with course credits, as per standard university practice. The Mechanical Turk sample contained 270 adults, who were paid US $0.80 each for their participation. This level of compensation was chosen in an attempt to be close to the median pay rate for HITs requiring similar time commitments available at the time of data collection, although no centralized database exists to identify the true distribution of HIT compensation levels. Demographic information for each sample is presented in Table 1, including, age, gender, nationality, location, employment status, tenure, education, and profession.

Procedure

Mechanical Turk A HIT was created that contained a brief description of the study and a link to an online informed consent form and questionnaire.[1] After the questionnaire was completed, a completion code was presented to the participant. In order for the participant to receive compensation, he or she had to enter the completion code on the Mechanical Turk Website. A useful feature of Mechanical Turk is that it provides an administrative page that reveals real-time submission statistics and completion codes. Once a completion code had been entered, the experimenter reviewed and approved the code, thus automatically sending compensation to the participant's account. This method ensured that identifying information connected to their worker ID was not connected to their responses.

Undergraduates The study description was posted on a university Website managed by the psychology department, using language identical to that for the Mechanical Turk HIT. As with the Mechanical Turk sample, the undergraduates completed their work with a computer online in an asynchronous manner from a location of their choosing. Participants who chose to sign up after viewing the study description were given an HTML link to the questionnaire and informed consent. To receive course credit for participation, participants entered their e-mail address, using an independent questionnaire link (i.e., e-mail addresses were not connected to their responses).

Measures

A number of measures were included for their widespread usage among organizational researchers, while others were included for their relevance to previous research on online survey behavior. Reliability estimates were calculated separately for each sample; this information is presented in Table 2. Unless indicated otherwise, responses were given on a 5-point scale with anchors from *strongly disagree* to *strongly agree*.

Internet knowledge Internet knowledge was measured with a 13-item scale from Potosky (2007). An example item is, "I am familiar with html."

Computer attitudes Attitudes toward computers were measured with a 19-item scale from Garland and Noyes (2004). An example item is, "People who like computers are often not very sociable" (reverse coded).

Computer knowledge and experience Computer knowledge and experience was measured with a 12-item scale from Potosky and Bobko (1998). An example item is, "I know how to recover deleted or 'lost data' on a computer or PC."

Goal orientation Learning goal orientation, performance-prove goal orientation, and performance avoid goal orientation (four items each) were measured with VandeWalle's (1997) scale.

Personality Extraversion, agreeableness, neuroticism, openness, and conscientiousness (i.e., the Big 5) were measured with 20 items each from the International Personality Item Pool (Goldberg, 1999) version of the NEO-PI–R.

Open-ended questions A number of open-ended questions were included at the end of the survey. These questions included the following: "Why did you take this survey?" "What was the best/worst thing about this survey?" "Would you be interested in participating in future studies on this topic? Why/why not?"

Mechanical Turk experience For the Mechanical Turk sample only, a number of open-ended questions were included to assess experience, motivation, and usage patterns for the Mechanical Turk Website. These questions were the following: "How did you first hear of Mechanical Turk?" "How many HITs have you completed?" "How long have you been using Mechanical Turk?" "How many hours per month do you spend using Mechanical Turk?""Why do you use Mechanical Turk?" Responses were content-coded by two raters; after discussion, there were no discrepancies in the coding.

Demographic measures We also asked several demographic questions, such asage, gender, ethnicity, nationality, education level, profession, years of work experience, and current employment status.

[1] Additional information about the HIT, including survey templates, can be obtained from the second author.

Table 1 Sample characteristics

		Undergraduate Pool		MTurk Pool		Chi-sq	df	p
		N	%	N	%			
Age						t = 21.48*	527	< .001
	Under 18	10	3.79	0	0.00			
	18-25	250	94.70	87	32.58			
	26-35	4	1.52	81	30.34			
	36-45	0	0.00	61	22.85			
	46-55	0	0.00	30	11.24			
	Above 55	0	0.00	7	2.62			
	No response	1	0.38	1	0.37			
Gender						0.19	2	.908
	Male	100	37.88	97	36.33			
	Female	161	60.98	169	63.30			
	No response	3	1.14	1	0.37			
Ethnicity						13.19	5	.022
	Caucasian	217	82.20	213	79.78			
	African American	16	6.06	8	3.00			
	Asian	22	8.33	22	8.24			
	Hispanic	3	1.14	14	5.24			
	Other/multiple	4	1.52	6	2.25			
	No response	2	0.76	4	1.50			
Education completed						211.93	5	< .001
	Middle school	0	0.00	1	0.37			
	High school	243	92.05	84	31.46			
	2-year degree	7	2.65	53	19.85			
	4-year degree	12	4.55	95	35.58			
	Master's or equivalent	0	0.00	28	10.49			
	Ph.D. or equivalent	0	0.00	6	2.25			
Currently employed?						168.08	2	< .001
	No	170	64.39	81	30.34			
	Yes–part-time	88	33.33	49	18.35			
	Yes–full-time	4	1.52	137	51.31			
	No response	2	0.76	0	0.00			
Years employed at current job						t = 5.96*	291	< .001
	< 1–2	65	24.62	82	30.71			
	3–4	14	5.30	46	17.23			
	5–6	1	0.38	27	10.11			
	7–8	2	0.76	13	4.87			
	9–10	0	0.00	15	5.62			
	> 10	0	0.00	28	10.49			
	No response	182	68.94	56	20.97			
Profession						206.84	28	< .001
	No response	115	43.56	17	6.37			
	Student	108	40.91	27	10.11			
	Unemployed	2	0.76	22	8.24			
	Self-employed	0	0.00	1	0.37			
	Retired	0	0.00	2	0.75			
	Business & management	1	0.38	38	14.23			
	Computer, mathematical & engineering	2	0.76	34	12.73			

Table 1 (continued)

		Undergraduate Pool		MTurk Pool		Chi-sq	df	p
		N	%	N	%			
	Life, physical, and social science	0	0.00	2	0.75			
	Community and social services	0	0.00	6	2.25			
	Legal	0	0.00	5	1.87			
	Education, training, and library	0	0.00	17	6.37			
	Arts, design, entertainment, sports, and media	5	1.89	19	7.12			
	healthcare	3	1.14	9	3.37			
	Sales, service & food	15	5.68	28	10.49			
	Office and administrative support	4	1.52	29	10.86			
	Construction and maintenance	1	0.38	5	1.87			
	Production	3	1.14	3	1.12			
	Transportation and material moving	1	0.38	2	0.75			
	Military & protective service	4	1.52	1	0.37			
Region of U.S. (U.S. residents only)						168.55	8	< .001
	Midwest	2	0.76	61	22.85			
	Northeast	64	24.24	48	17.98			
	South	75	28.41	36	13.48			
	Southeast	81	30.68	28	10.49			
	Southwest	1	0.38	18	6.74			
	West	1	0.38	34	12.73			
	Northwest		0.00	18	6.74			
	Other		0.00	2	0.75			
	No response	40	15.15	22	8.24			
Nationality						38.62	9	< .001
	U.S. & Canada	219	82.95	248	92.88			
	Brazilian	0	0.00	0	0.00			
	UK & Western Europe	0	0.00	8	3.00			
	Central & South America	0	0.00	2	0.75			
	Asia	0	0.00	1	0.37			
	No response	45	17.05	8	3.03			
Reason for taking survey						106.37	13	< .001
	Boredom	0	0.00	14	5.24			
	Compensation	207	78.41	121	45.32			
	Curious	8	3.03	27	10.11			
	Ease	0	0.00	4	1.50			
	Enjoys surveys	1	0.38	21	7.87			
	Fun	0	0.00	4	1.50			
	To help research	0	0.00	8	3.00			
	Introspection	1	0.38	1	0.37			
	Education	1	0.38	1	0.37			
	Subject	6	2.27	35	13.11			
	No response	40	15.15	31	11.61			

*Age and tenure were assessed by asking open-ended time-based questions. Thus, while the variables are presented here as categorical, the underlying variables were continuous, and thus, mean differences were compared using a *t* test

Table 2 Scale reliabilities, mean comparisons, and descriptive statistics by sample

	Coefficient Alpha			Mechanical Turk			University Subject Pool						
	MTurk	Subj Pool	*p*	*N*	*M*	*SD*	*N*	*M*	*SD*	*t*	*df*	*p*	Cohen's *d*
N	.92	.91	.18	198	19.64	10.51	206	20.32	9.98	-0.67	402	0.51	-0.07
E	.93	.90	.04	199	20.06	10.09	207	23.01	8.94	-3.12	404	0.00	-0.31
O	.82	.84	.58	200	17.93	5.33	202	13.11	5.90	8.58	400	< .0001	0.86
C	.88	.88	.83	200	20.40	8.08	204	20.11	8.04	0.36	402	0.72	0.04
A	.87	.87	.96	197	15.34	5.84	209	16.27	5.60	-1.65	404	0.10	-0.16
LGO	.85	.73	.00	210	3.67	1.96	214	2.62	1.72	5.87	422	< .0001	0.57
PPGO	.79	.68	.02	209	5.75	2.98	213	5.52	2.52	0.86	420	0.39	0.08
PAGO	.85	.73	.00	209	3.95	3.24	213	5.34	2.99	-4.57	420	< .0001	-0.44
IntKnowl	.88	.91	.03	206	4.01	0.58	209	3.67	0.71	5.41**	413	0.00	-0.48
CompAtt	.81	.81	1.0	198	3.86	0.41	206	3.70	0.41	4.05**	402	0.00	-0.39
CompKnow	.78	.81	.28	201	3.72	0.55	212	3.45	0.57	4.85**	411	0.00	-0.47

Note. N = neuroticism, E = extraversion, O = openness to experience, C = conscientiousness, A = agreeableness, LGO = learning goal orientation, PPGO = performance–prove goal orientation, PAGO = performance–avoid goal orientation., IntKnowl = Internet knowledge, CompAtt = computer attitudes, CompKnow = computer knowledge

Response behavior Several measures of survey-taking behavior were collected to gain additional information about the quality of each sample's responses. Response time was assessed as the number of minutes spent completing the survey, as calculated by using time stamps at survey initiation and completion. Also, the length of open-ended comments, in number of words, was assessed by obtaining the total word count for all open-ended responses. Finally, responses were flagged for deletion if respondents exited the survey before completion or if the total time spent working on the survey was less than 10 min. The proportion of cases that were flagged was then calculated.

Data quality Excessive response consistency was assessed by selecting a pair of Likertscale items that should have opposite responses (i.e., *psychometric antonyms*; Goldberg & Kilkowski, 1985): "seldom feel blue" and "often feel blue." Cases were flagged if their responses to these two items were identical. Next, random responders were identified by selecting a pair of Likertscale items that should have similar responses: "do things according to a plan" and "make a plan and stick to it." Cases were flagged if their responses to these two items were more than 2 points apart. The total proportion of cases flagged under either rule was then computed. Finally, the Long String Index (Johnson, 2005) was calculated; this index measures the longest continuous string of identical responses for a given participant (e.g., selecting "strongly agree" for 15 consecutive items), giving an additional measure of inattentive responding.

Social desirability Socially desirable responding was measured with the 33-item scale from Crowne and Marlowe (1960), with a true–false response format. Example items are, "I never hesitate to go out of my way to help someone in trouble" and "There have been times when I have been quite jealous of the good fortune of others" (reversed). High scores on this scale indicate a desire to "fake good" and respond to survey items in a socially desirable manner; low scores indicate more honest responding.

Scale reliability Cronbach's coefficient alpha was calculated for each sample to obtain a measure of internal consistency. The differences between coefficient alpha values across samples were compared via a chi-square statistic described by Feldt, Woodruff, and Salih (1987), using the AlphaTest program (Lautenschlager & Meade, 2008).

Measurement invariance analysis

The measurement invariance of each of the personality and goal orientation scales was investigated using an item response theory (IRT) based likelihood ratio test. Prior to IRT analysis, items were reverse coded where appropriate, and flagged cases (as described above) were deleted. Cases were also deleted on a scale-by-scale basis if they contained any missing data. In order to perform invariance analyses using IRT, it was necessary to first establish the dimensionality of each scale. A principal components analysis (PCA) was performed separately for each scale, and for seven of the eight scales, scree plots and eigenvalues suggested that the scales were clearly unidimensional (e.g., first eigenvalues well above 1.0, second eigenvalues near or below 1.0, with scree plots showing a clear drop). However, the conscientiousness scale suggested that more than one factor may fit the data. A follow-up exploratory factor analysis with principal factors extraction and oblique rotation revealed that item factor loadings produced factors that were not sufficiently conceptually distinct. In the end, for all eight scales, we opted to use the PCA results, retaining only one factor and using only items that loaded more than .40 on that factor in the IRT analyses. The total number of items retained for subsequent analyses for each scale is reported in Table 3.

Data analysis involved using the IRT graded response model (Samejima, 1969), in which one item *a* parameter and one fewer than the number of response option *b* parameters are estimated for each item (see Embretson & Reise, 2000, for a description). The *b* parameters identify the boundary of probability of responding with one option (e.g., 1) with those of the next option and higher (e.g., 2 through 5). Preliminary analysis suggested that when fewer than 20 respondents in either group endorsed a given response option, standard errors associated with the associated *b* parameter were quite large. As a result, prior to the analyses, response categories were collapsed into a single category for such items. For instance, if 15 persons responded with a "1" for a given item, those persons' responses were recoded as "2." As a result, different items had a different number of (recoded) response options. All items were then recoded such that the lowest response option was "0," as required by the software used for these tests.

The invariance of the eight scales was examined, one at a time, with the likelihood ratio test (LRT; Thissen, Steinberg, & Wainer, 1993), using the IRTLRDIF program (Thissen, 2001). The IRTLRDIF program first estimates a baseline model in which all item parameters are constrained to be equal across groups for a given scale. This baseline is then compared with a series of augmented models in which the parameters for a single item are free to vary across groups. The improvement in model fit associated with freeing these constraints is distributed as chi-square with degrees of freedom corresponding to the number of freed parameters. As with all chi-square-based statistics, the LRT is sensitive to sample size and has very high power when samples are large (Rivas, Gabriel, Stark, & Chernyshenko, 2009), potentially detecting even trivial noninvariance (Meade, 2010). As such, we also computed invariance effect sizes to indicate the practical importance of a lack of invariance (or differential functioning [DF]). Meade recently developed a taxonomy of potential invariance effect size measures based on item and scale expected scores. The most basic of these is the signed test difference in the sample (STDS), which can be interpreted as simply the difference in the two groups' mean scores expected because of DF alone. A second index is the unsigned expected test score difference in the sample (UETSDS), which can be interpreted as the difference in scale scores due to DF alone, had the differences in scale scores uniformly "favored" one of the groups. The UETSDS is equivalent to the square root of Raju et al.'s (1995) NCDIF index. Additionally, the expected test score

Table 3 Differential item functioning

	N	E	O	C	A	LGO	PPGO	PAGO
Number of scale items	18	18	18	17	17	4	5	5
Total sample size	404	406	402	404	406	424	422	422
Sample size of Turk high group focal	198	199	200	200	197	210	209	209
# DIF items, using p < .01	2	0	4	3	2	1	1	0
Signed test difference in the sample (STDS)*	0.009	0.171	0.312	0.151	-0.445	0.024	-0.029	-0.139
Unsigned ETS difference in sample (UETSDS)	0.170	0.172	0.350	0.249	0.445	0.166	0.094	0.233
Expected test score stand. difference (ETSSD)*	0.001	0.018	0.065	0.020	-0.085	0.015	-0.013	-0.052
Possible scale range	43	40	29	39	30	7	12	11

Note. *MTurk expected to score higher than undergraduate sample due to lack of invariance. N = neuroticism, E = extraversion, O = openness to experience, C = conscientiousness, A = agreeableness, LGO = learning goal orientation, PPGO = performance–prove goal orientation, PAGO = performance–avoid goal orientation standardized difference (ETSSD) is reported as a test-level DF version of Cohen's *d* (Meade, 2010).

Results

Research question 1 concerned the demographic makeup of the crowdsourcing sample, as compared with a university sample. Table 1 contains a comparison of the two samples for demographic characteristics, including age, gender, ethnicity, nationality, education completed, employment status, and profession. The samples were similar in terms of gender and ethnicity; both samples were predominantly female and Caucasian. The crowdsourced sample was markedly more diverse in terms of education, employment status, and profession, with a wide range of professions and education levels represented. There were significant mean differences in age, such that the crowdsourcing sample (M = 32.93, SD = 10.68) was significantly older than the undergraduate sample (M = 18.68, SD = 1.35), $t(527) = 21.48$, $p < .001$, $d = 1.87$, consistent with expectations. A dramatically higher percentage of the crowdsourced sample was employed, either full-time or part-time. Additionally, for respondents from either sample who were employed, tenure in their current job was considerably longer in the crowdsourced sample (M = 5.11 years, SD = 5.33) than in the university sample (M = 1.54, SD = 1.51), $t(291) = 5.96$, $p < .001$, $d = .78$. As a whole, this information suggests that the crowdsourced sample was more attractive in terms of generalizability for organizational researchers.

Research question 2 concerned differences in response quality, as measured by differences in social desirability, reliability of scales, completion time, length of open-ended responses, and data consistency and completeness. It was demonstrated by *t* tests that the crowdsourced sample was significantly higher in social desirability (see Table 4). However, internal consistency estimates tended to be higher in the Mechanical Turk sample than in the undergraduate sample (see Table 2), with the exception of the Internet knowledge measure. No significant differences were found with respect to completion time or word count. The Long String Index showed no differences between samples. Finally, a similar proportion of cases in each sample were flagged, due to incompleteness or data consistency. As a whole, this information indicates that the data were of equal or perhaps better quality in the crowdsourcing sample, although slightly more susceptible to socially desirable responding.

Research question 3 concerned the measurement invariance of commonly used scales—namely, Big 5 measures of personality and goal orientation. As can be seen in Table 3, on the whole, most items tended to function equivalently across samples, with only one or two DF items per scale. Exceptions to these general findings were the openness (four DF items) and conscientiousness (three DF items) scales. Items in these scales displaying DF were as follows: "am full of ideas," "have a rich vocabulary," "love to read challenging reading material," "have difficulty imagining things," "get chores done right away, "am exacting in my work," and "shirk my duties." An examination of these items suggests that individuals in the crowdsourced sample with more work experience might reasonably interpret these items differently than those with little work experience. In contrast, items such as "tend to vote for liberal political candidates" would not be expected to vary in their interpretation on the basis of work experience, and indeed, items like these did not display DF in the present study.

Despite these statistically significant differences, across all scales, DF effect sizes were quite small. For instance, for the conscientiousness scale, the potential scale score range was from 17 to 56 (range = 39). However, the expected mean difference in observed scores between the two groups

Table 4 Data quality

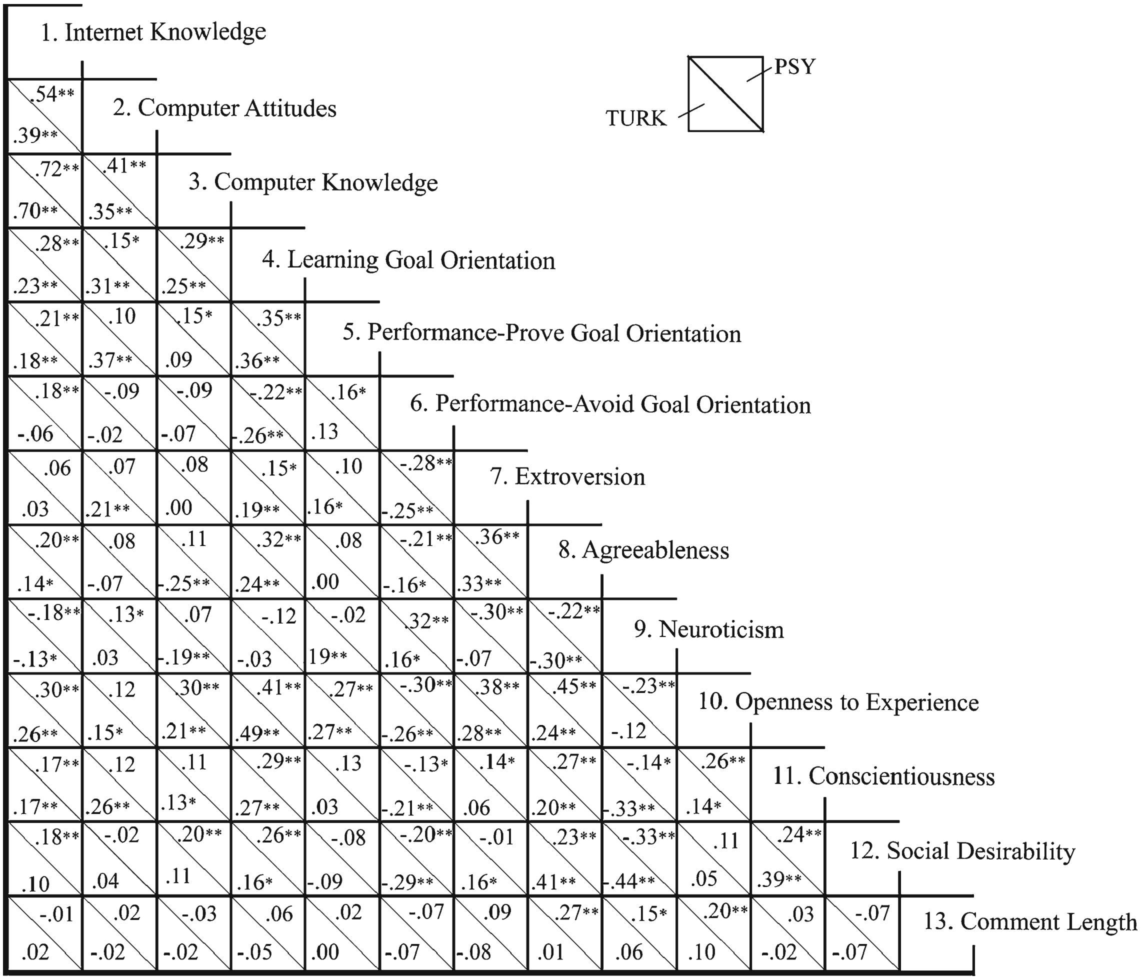
Variable	University Sample			MTurk Sample						
	N	*M*	*SD*	*N*	*M*	*SD*	*T*	*df*	*p*	Cohen's *d*
Data completeness flag	283	.07	.25	280	.05	.21	-1.06	561	.29	0.09
Data quality flag	283	.19	.40	280	.24	.43	1.19	561	.23	-0.12
Long string index	283	9.93	8.72	280	10.18	8.88	-0.33	561	.74	-0.03
Minutes, all data	272	28.21	20.73	274	26.45	30.25	-0.79	483.32	.43	0.07
Minutes, screened data	210	29.71	19.68	210	27.70	33.95	-0.75	335.26	.46	0.10
Word count, all data	283	32.61	22.57	280	29.17	22.09	-1.83	561	.07	0.15
Word count, screened data	216	33.08	20.90	210	30.72	22.46	-1.12	424	.26	0.11
Social desirability (screened only)	203	8.02	0.75	200	8.30	0.91	3.37**	401	.00	-0.37

**p < .01

due to DF alone was .151, less than one fifth of one scale point out of a potential 39. Effect sizes were slightly higher for the openness and agreeableness scales, although still not especially large. For instance, the ETSSD indices for the openness scale indicated that group mean differences would be expected to be 0.065 *SD* higher in the Mechanical Turk sample due to DF alone.

Given the minimal role of DF in the observed data, we were able to examine research question 4, which concerned mean differences with respect to individual differences, including personality, attitudes, and computer knowledge/experience. The Mechanical Turk sample was significantly higher in computer and Internet knowledge. The Mechanical Turk sample was also higher in openness to experience and learning goal orientation and was lower in extraversion (see Table 2). Effect sizes (*d*) were typically small, according to Cohen's (1969) criteria. Bivariate correlations by sample are presented in Fig. 1.

Finally, research question 5 concerned the primary motivations for persons' participating in crowdsourcing. Most respondents indicated that financial incentives were the primary reason for using Mechanical Turk, though educational and entertainment benefits were also listed (see Table 5). The majority of respondents self-identified as casual users, although a very small subset of users indicated having completed over 1,000 individual HITs and having spent over 100 h per month on the site, making their experience equivalent to part-time employment. Thus, the financial element of participating in crowdsourcing is important to users, although participation is still voluntary and the attractiveness of a given study may carry more weight in participation decisions than does the precise dollar amount of the compensation. For long, involved, or repetitive studies, one may need to compensate participants at a higher rate, while for engaging and interesting studies, one may be able to pay participants slightly less.



1. Internet Knowledge												
2. Computer Attitudes	.54** .39**											
3. Computer Knowledge	.72** .70**	.41** .35**										
4. Learning Goal Orientation	.28** .23**	.15* .31**	.29** .25**									
5. Performance-Prove Goal Orientation	.21** .18**	.10 .37**	.15* .09	.35** .36**								
6. Performance-Avoid Goal Orientation	.18** -.06	-.09 -.02	-.09 -.07	-.22** -.26**	.16* .13							
7. Extroversion	.06 .03	.07 .21**	.08 .00	.15* .19**	.10 .16*	-.28** -.25**						
8. Agreeableness	.20** .14*	.08 -.07	.11 -.25**	.32** .24**	.08 .00	-.21** -.16*	.36** .33**					
9. Neuroticism	-.18** -.13*	.13* .03	.07 -.19**	-.12 -.03	-.02 19**	.32** .16*	-.30** -.07	-.22** -.30**				
10. Openness to Experience	.30** .26**	.12 .15*	.30** .21**	.41** .49**	.27** .27**	-.30** -.26**	.38** .28**	.45** .24**	-.23** -.12			
11. Conscientiousness	.17** .17**	.12 .26**	.11 .13*	.29** .27**	.13 .03	-.13* -.21**	.14* .06	.27** .20**	-.14* -.33**	.26** .14*		
12. Social Desirability	.18** .10	-.02 .04	.20** .11	.26** .16*	-.08 -.09	-.20** -.29**	-.01 .16*	.23** .41**	-.33** -.44**	.11 .05	.24** .39**	
13. Comment Length	-.01 .02	.02 -.02	-.03 -.02	.06 -.05	.02 .00	-.07 -.07	.09 -.08	.27** .01	.15* .06	.20** .10	.03 -.02	-.07 -.07

* $p < .05$

** $p < .01$

Fig. 1 Bivariate correlations

Table 5 Mechanical Turk usage patterns and motivations

	N	%
How respondent heard of Turk		
Turk advertisement	4	1.50
Article	89	33.33
Forums	12	4.49
Other Web site	94	35.21
TV/radio	7	2.62
Personal referral	33	12.36
Other	19	7.12
No response	9	3.37
Why respondent uses Turk		
Boredom	14	5.24
Compensation	189	70.79
Curious	15	5.62
Ease	2	0.75
Fun	20	7.49
Education	4	1.50
Reputation	3	1.12
Habit	2	0.75
No response	18	6.74
How long respondent has used Turk (months)		
< 1–4	218	81.65
5–8	8	3.00
9–12	12	4.49
13–16	0	0.00
17–20	1	0.37
21–24	5	1.87
25–28	0	0.00
29–32	2	0.75
33–36	2	0.75
No response	10	3.75
Hits completed		
0–9	54	20.22
10–99	116	43.45
100–999	43	16.10
1,000–9,999	24	8.99
10,000–99,999	8	3.00
100,000–999,999	1	0.37
No response	21	7.87
Time spent on Turk (hours/month)		
< 1–10	64	23.97
11–20	17	6.37
21–30	18	6.74
31–40	6	2.25
41–50	6	2.25
51–60	16	5.99
61–70	0	0.00
71–80	2	0.75
81–90	4	1.50
> 100	9	3.37
No response	125	46.82

*$p < .05$

**$p < .01$

Discussion

In this study, we sought to identify whether crowdsourcing is a viable alternative to the use of university subject pools, which are often criticized for their homogeneous makeup and limited work experience (Anderson, 2003; Landy, 2008; Locke, 1986; Ward, 1993). We administered a survey to samples drawn from both crowdsourcing and university participant pools to examine the quality of data gathered from each source, as well as to understand more about the crowdsourcing participants. Overall, the crowdsourcing sample behaved similarly to participants from a traditional psychology participant pool, a finding that is consistent with Sprouse's (2011) comparison of the quality of acceptability judgment data between Mechanical Turk workers and in-lab participants. Where differences existed, effect sizes were typically small. A few noticeable differences were found in terms of data quality, with the slightly higher levels of social desirability in the crowdsourcing sample offset by the slightly better reliability of the data from that sample. Additionally, there were some clear advantages gained from using crowdsourcing; namely, the resulting sample was more diverse, was older, and had more relevant experience, making them an attractive pool for organizational researchers. Thus, it would seem that crowdsourcing tools are a viable option for organizational researchers.

Given the relatively small monetary compensation offered in this study, we were also interested in exploring the motivation of crowdsourcing participants. Approximately 70% of respondents indicated that their primary motivation for using Mechanical Turk, in general, was financial, although the remainder of respondents listed other benefits, such as entertainment or education. It is worth noting that when asked why they volunteered to complete this particular survey, almost all respondents from the undergraduate sample listed course credit as their primary reason, significantly more than in the Mechanical Turk sample, in which respondents gave a number of other reasons, such as an interest in taking surveys and a general interest in personality and related topics.

Implications

Practical Our findings show that the use of crowdsourcing can be a potentially viable resource for researchers wishing to collect survey data on many types of organizational phenomena. This is especially relevant to those researchers who commonly recruit participants from undergraduate populations. Specifically, the ability to select participants at the country level may reduce the barriers of recruiting only WEIRD populations. It is important to point out, however, that Amazon's Mechanical Turk can disburse earnings only to U.S. or Indian bank accounts. Workers from other countries are paid through Amazon.com gift certificates. The respondents to this survey were, on average, employed, had several years of full-time work experience, and came from a wide range of organizations and occupations, making them a more representative population for many types of organizational studies. In addition, these portals provide access to researchers who either have small participant pools or no access to an undergraduate psychology pool or its equivalent. Other practical benefits should also be noted; primary among these is the significant time savings that can be realized by using crowdsourcing. In the present study, over 250 surveys were completed in 2 days, while a similar sample size drawn from the undergraduate psychology pool took several weeks. Moreover, the availability of crowdsourcing participants is not limited by semester schedules or the size of a university's undergraduate population.

Ethical Mechanical Turk users are evaluated by the requesters, and their rating is visible to others who post work on the site. Thus, there is the potential that Mechanical Turk users will feel undue pressure to complete the survey even if they want to exit. This makes the use of informed consent forms that much more important. However, there is some evidence that online informed consent documentation is not always read carefully (Stanton & Rogelberg, 2001), leaving open the possibility that Mechanical Turk workers will not fully understand that their rating and compensation are not connected to their survey responses. Institutional review boards may or may not be familiar with crowdsourcing as a participant source, and researchers will need to work carefully to make sure participants are treated ethically. A cursory review of psychology surveys posted on Mechanical Turk discovered that almost one third had no informed consent information posted at all.

Additionally, conducting survey or other research is not the intended purpose of labor portals such as Mechanical Turk. The observed increase in the social desirability of Mechanical Turk responses could be due to perceptions that compensation will be based on the nature of the responses given; that is, Mechanical Turk workers may respond in socially desirable ways to avoid being docked pay. This type of subtle coercion has long been cited as a concern in the use of university participant pools (e.g., Rosnow & Rosenthal, 1976), although the problems may be exacerbated in an online labor portal setting.

Considerations for researchers wishing to use Mechanical Turk

Technical Although most features of Mechanical Turk can be accessed using a Web browser interface, some features are accessible only from text-only programming commands or using Amazon Web Service's application programming interface. To fully utilize Mechanical Turk, some degree of computer programming proficiency is required. This is especially true when using Mechanical Turk as a medium for implementing psychology-based experiments that go beyond the scope of simple input field data collection.

Financial The benefits of using Mechanical Turk or other crowdsourcing options do not come for free. Participants recruited in this way must be offered financial compensation. In the present study, participants were paid US $0.80 to complete a survey that took approximately 30 min to complete. Using this level of compensation, we were able to recruit several hundred participants in less than 48 h. However, more involved surveys or experimental studies may require higher levels of compensation. Mechanical Turk workers are generally aware of what the fair wage for a given task should look like and respond favorably to requesters who match or exceed this rate.

Oversurveying Given the relative ease and short timeframe promised by the use of crowdsourcing, the possibility of oversurveying becomes a concern (Thompson & Surface, 2007; Tippins, 2002). While members of a given labor portal are never obligated to accept a particular survey, it is possible that an excessive number of surveys will make the marketplace less attractive. However, anecdotal comments from Mechanical Turk participants in this study indicated that surveys were often welcome jobs and were relatively more engaging than other tasks available on the site.

Limitations and future research

The use of Internet-based psychological research has been discussed for well over a decade (Reips, 2002; Stanton, 1998), and many studies have highlighted the advantages and disadvantages of implementing online research. The next step in this discussion needs to bean investigation into how the online recruitment process can be improved both from a technological standpoint and, perhaps more importantly, from a psychometric standpoint where the generalizability of participants is held in a higher regard.

It is important to acknowledge the limitations of this form of data collection. As with any method involving monetary compensation, the motives of participants may be called into question. Indeed, an important avenue for future study is the effect of varying levels of compensation. In the present study, we selected a compensation level slightly above the median for the timeframe required. It is possible that paying much less than that would result in fewer participants signing up, while paying much more would attract participants who were not truly interested in completing the survey. Alternatively, paying less may also impart to participants a feeling that they have less of an obligation to to provide thoughtful responses, potentially resulting in low-quality data. Future research is required to explore the potential relationship between compensation and data quality and methods for embedding internal checks for undesirable motivations on the part of participants.

It will also be important to identify the degree to which these results are idiosyncratic to the specific community of Mechanical Turk users and to what degree they are generalizable across users of all crowdsourcing marketplaces. Additionally, the current lack of random sampling techniques for Internet users needs to be addressed so that Internet-sourced data can be more confidently generalized across Internet users as a whole (Kraut et al., 2004).Along the same lines, it will be important to discover the utility of this tool for investigating research questions that do not rely on survey methodology (e.g., experimental designs, diary studies).

In conclusion, the promise of crowdsourcing tools will go unrealized if researchers cannot be confident in the quality of the data they will obtain. This study provides initial evidence that data quality is as good as that from undergraduate pools and that diverse samples can be obtained, using these tools.

References

Amazon.com. (2010). Mechanical Turk, Retrieved from http://www.mturk.com/

Anderson, N. R. (2003). Applicant and recruiter reactions to new technology in selection: A critical review and agenda for future research. *International Journal of Selection and Assessment, 11*, 121–136. doi:10.1111/1468-2389.00235

Barchard, K., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods, 40*, 1111–1128. doi:10.3758/BRM.40.4.1111

Booth-Kewley, S., Edwards, J. E., & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *The Journal of Applied Psychology, 77*, 562–566.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cole, F., Sanik, K., DeCarlo, D., Finkelstein, A., Funkhouser, T., Rusinkiewicz, S., et al. (2009). How well do line drawings depict shape? *ACM Transactions on Graphics, 28*, 1–9. doi:10.1145/1531326.1531334

Cole, M. S., Bedeian, A. G., & Field, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership. *Organizational Research Methods, 9*, 339–368. doi:10.1177/1094428106287434

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods, 40*, 428–434.

DeBeuckalaer, A., & Lievens, F. (2009). Measurement equivalence of paper-and-pencil and internet organizational surveys: A large-scale examination in 16 countries. *Applied Psychology, 58*, 336–361. doi:10.1111/j.1464-0597.2008.00350.x

Embretson, S. E., & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93–103. doi:10.1177/014662168701100107

Garland, K. J., & Noyes, J. M. (2004). Computer experience: A poor predictor of computer attitudes. *Computers in Human Behavior, 20*, 823–840. doi:10.1016/j.chb.2003.11.010

Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & F. Ostendorf (Eds.), *Personality psychology in Europe vol. 7* (pp. 7–28). Tilburg, The Netherlands: Tilburg University Press.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*, 82–98. doi:10.1037/0022-3514.48.1.82

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. *Brain and Behavior Sciences, 33*, 94–95. doi:10.1017/S0140525X10000300

Heilman, M., & Smith, N. A. (2010). *Rating computer-generated questions with Mechanical Turk.* Paper presented at the NAACL Workshop on Creating Speech and Language Data With Amazon's Mechanical Turk, Los Angeles, CA

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. *Nature, 466*, 2929. doi:10.1038/466029a

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? *The Behavioral and Brain Sciences, 33*, 61–135. doi:10.1017/S0140525X0999152X

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), Retrieved from http://www.wired.com/wired/archive/14.06/crowds.html

Ipeirotis, P. G. (2010). Demographics of Mechanical Turk. *Center for digital economy research working papers*, 10. Retrieved from http://hdl.handle.net/2451/29585

Johnson, J. A. (2005). Ascertaining the validity of individual protocols from Web-based personality inventories. *Journal of Research in Personality, 39*, 103–129. doi:10.1016/j.jrp.2004.09.009

Kittur, A., Chi, E., Suh, B. (2008). *Crowdsourcing user studies with Mechanical Turk*. Paper presented at the Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy

Kleemann, F., Voß, G., & Rieder, K. (2008). Un(der)paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science Technology & Innovation Studies, 4*, 5–26.

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. *The American Psychologist, 59*, 105–117. doi:10.1037/0003-066X.59.2.105

Landy, F. J. (2008). Stereotypes, bias, and personnel decisions: strange and stranger. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 379–392. doi:10.1111/j.1754-9434.2008.00071.x

Lautenschlager, G. J., & Meade, A. W. (2008). AlphaTest: A windows program for tests of hypotheses about coefficient alpha. *Applied Psychological Measurement, 32*, 502–503. doi:10.1177/0146621607312307

Little, G., Chilton, L. B., Goldman, M., Miller, R. C. (2009). *TurKit: Tools for iterative tasks on Mechanical Turk*. Paper presented at the ACM SIGKDD Workshop on Human Computation, Paris, France

Locke, E. A. (Ed.). (1986). *Generalizing from laboratory to field settings*. Lexington, MA: Lexington Books.

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other modes. *International Journal of Market Research, 50*, 79–104.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential item functioning of items and scales. *The Journal of Applied Psychology, 95*, 728–743.

Meade, A. W., Michels, L. C., & Lautenschlager, G. J. (2007). Are Internet and paper-and-pencil personality tests truly comparable? An experimental design measurement invariance study. *Organizational Research Methods, 10*, 322–345. doi:10.1177/1094428106289393

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.

Ployhart, R. E., Weekley, J. A., Holtz, B. C., & Kemp, C. F. (2003). Web-based and paper-and-pencil testing of applicants in a proctored setting: are personality, biodata, and situational judgment tests comparable? *PersonnelPsychology, 56*, 733–752. doi:10.1111/j.1744-6570.2003.tb00757.x

Potosky, D. (2007). The Internet knowledge (iKnow) measure. *Computers in Human Behavior, 23*, 2760–2777. doi:10.1016/j.chb.2006.05.003

Potosky, D., & Bobko, P. (1998). The computer understanding and experience scale: A self-report measure of computer experience. *Computers in Human Behavior, 14*, 337–348.

Raju, N. S., van der Linden, W., & Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. *Applied Psychological Measurement, 19*, 353–368.

Reips, U. D. (2002). Standards for internet-based experimenting. *Experimental Psychology, 49*, 243–256. doi:10.1027//1618-3169.49.4.243

Richman, W., Kiesler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *The Journal of Applied Psychology, 84*, 754–775.

Rivas, L., Gabriel, E., Stark, S., & Chernyshenko, O. S. (2009). The effects of referent item parameters on differential item functioning detection using the free baseline likelihood ratio test. *Applied Psychological Measurement, 33*, 251–265.

Rosnow, R. L., & Rosenthal, R. (1976). The volunteer subject revisited. *Australian Journal of Psychology, 28*, 97–108.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, *17*.

Sharek, D. (2010). *The influence of flow in the measure of engagement*.Unpublished Master's thesis. Retrieved from http://www.lib.ncsu.edu/theses/available/etd-02182010-132225/

Sprouse, 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

Stanton, J. M. (1998). An empirical assessment of data collection using the internet. *Personnel Psychology., 51*, 709–725.

Stanton, J. M., & Rogelberg, S. G. (2001). Using internet/intranet web pages to collect organizational research data. *Organizational Research Methods, 4*, 200–217.

Thissen, D. (2001). *Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning*. Chapel Hill: Universityof North Carolina at Chapel Hill, L. L. Thurstone Psychometric Laboratory.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Thompson, L. F., & Surface, E. A. (2007). Employee surveys administered online: Attitudes toward the medium, nonresponse, and data representativeness. *Organizational Research Methods, 10*, 241–261. doi:10.1177/1094428106/294696

Tippins, N. T. (2002). Organization development and IT: Practicing OD in the virtual world. In J. Waclawski & A. H. Church (Eds.), *Organization development: A data-driven approach to organizational change* (pp. 245–265). San Francisco: Jossey-Bass.

Truell, A. D., Bartlett, J. E., & Alexander, M. W. (2002). Response rate, speed, and completeness: A comparison of Internet-based and mail surveys. *Behavior Research Methods, Instruments, & Computers, 34*, 46–49.

Turk and rescue. (2007). *Economist*, 384(8547), 97.

VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement, 57*, 995–1015. doi:10.1177/0013164497057006009

Ward, E. A. (1993). Generalizability of psychological research from undergraduates to employed adults. *The Journal of Social Psychology, 133*, 513–519.