Proceedings, The Fourth AAAI Conference on
Human Computation and Crowdsourcing

Validating the Quality of Crowdsourced Psychometric Personality Test Items

Bao Sheng Loe,[1] Francis Smart,[2] Lenka Firtova,[3]
Corinna Brauner,[4] Laura Lueneborg,[5] David Stillwell[1]

[1]University of Cambridge, bsl28@cam.ac.uk; d.stillwell@jbs.cam.ac.uk
[2]State Michigan University, smartfra@msu.edu
[3]University of Economics, lenkafirtova@seznam.cz
[4]University of Muenster, corinna.brauner@web.de
[5]University of Bonn, laura.lueneborg@gmail.com

Abstract

Personality questionnaire items are normally written by experts in personality theory or psychometrics. This makes personality test development expensive, especially since as many as 40% of the items are discarded after quantitative analysis. Could the crowd create personality items that are as good as the ones developed by professionals? We analyze the psychometric quality of 164 Big Five personality items generated by the crowd (n=82) and show that non-expert item writers are capable of designing high quality personality items. The analysis revealed that 59% of the crowdsourced items had adequate psychometric properties, and thus could be used as an alternative instrument to measure Big Five personality traits. We conclude that crowdsourcing personality items is a possible strategy to generate new items rapidly. We also discuss some of the challenges of crowdsourcing general psychological items.

Introduction

Psychological items are normally generated by professional item writers, psychologists or psychometricians, with a background in educational measurement, psychology, or a related discipline. Item writers usually undergo specialized training prior to item development and thus are deemed to have appropriate expertise. Once they have completed training, they may go on to write items for targeted clients (i.e. assessment companies), often at a high price. The items are then employed in pilot testing and the data is analysed in order to decide which items have acceptable psychometric properties.

Traditional item development requires that experts use fixed test specifications and follow item-writing guidelines to author each item. This process is costly and time-consuming. Rudner (2010) estimated that an operational item used in high-stakes testing could cost from $1,500 to $2,000. We list three reasons why it is necessary to generate large numbers of test items.

(1) As many as 40% of expertly created items fail to perform as intended when tested psychometrically (Haladyna 2012). Items that fail psychometric analysis need to be either revised or discarded (Haladyna and Rodriguez 2013).

(2) With smartphones and the Internet it is easier than ever for test-takers to compromise test security and share tests with other prospective test-takers. Some test developers create parallel forms of the same test so that each can be administered in a different locality or one of the forms can be randomly chosen to be administered to each candidate.

(3) Advances in computer processing power mean that it is now possible to use computer-adaptive testing (CAT) in real testing situations. CAT is a more efficient way of administering a test where items are selected for each candidate based on their previous responses, with the goal of maximising the expected information from each item (van der Linden and Glas 2000). CAT may reduce the test length by up to 50%-90% while keeping the same accuracy (Gibbons et al. 2008), and so it is desirable to deploy CAT tests in education and elsewhere, given that many people believe that children are over-tested (Coughlan, 2016). The

challenge for item developers is that CAT item banks are ideally four or five times larger than the number of items that each test-taker answers, so for a 50-item test the item bank should contain about 250 items. In short, test developers that adopt CAT have to invest heavily to create a sufficient number of new items for their testing programs or seek other alternatives to develop their item banks.

Crowdsourcing utilizes the collective knowledge of the public to complete tasks in return for financial incentives. Crowdsourcing platforms such as Amazon's MTurk (Mechanical Turk 2009) or CrowdFlower (CrowdFlower 2007) connect people and organize them to complete different sets of tasks. We examine whether test creators could utilize these platforms to have non-experts create operational personality items. If crowdsourcing is effective for developing high quality items, then it could decrease the cost of item development and increase the viability of new test administration procedures like CAT.

There is reason to be optimistic about the ability for non-experts to develop reasonable psychometric items. Sharpley and Rogers (1985) showed that anxiety scale items developed by non-psychologists (students) were just as reliable as those developed by psychologists who were expert item writers. Nevertheless, while the study demonstrates that non-experts can write items, it does not demonstrate crowdsourcing in the modern sense of the word, which enables test-makers to reach a large number of people via the Internet, who are both geographically and educationally diverse. Virtual job marketplaces such as crowdsourcing technology did not exist until the introduction of the Web 2.0 (Kleemann, Voß, and Rieder 2008). The aim of this research is to test whether crowdsourcing can provide a cost-effective and rapid method of generating new items to use not only in existing instruments but in new psychological instruments as well.

Why Crowdsourced Personality Items

There are two main reasons why the creation of personality items is an exciting direction for crowdsourcing research: (1) The item development phase can be quicker and more cost effective than when using traditional means; (2) Crowd-techniques expand the pool of item writers, and thus the background of those who are creating tests. Items generated by more diverse groups are more likely to reflect a wider range of real-world experiences and so may be more varied in content than those generated by a relatively small team of professional item writers. Both of these reasons are examined in further detail in the next sections.

Crowdsourced Items Can Speed Up Item Development Process

The development of an item bank via traditional means is often time-consuming and costly (Rudner 2010), particularly if there are no professional item writers within the immediate environment to help create items. Typically, in such situations, companies have to invest more resources in order to employ remote professional item writers to help in the item development process. Furthermore, if the piloting of new items were completed in an offline environment, there is a great deal of reliance on the number of participants one can recruit to complete the pilot test. Thus, extending the time taken to analyze the psychometric properties of the new items. Therefore, the goal of our research is to provide solutions that reduce the problem of test exposure and to devise an effective approach to the production of personality items in parallel forms with an optimized pricing scheme.

Scale development remains a thriving industry in the psychology discipline. Clark and Watson (1995) suggest that there is no restriction to the number of operationalized scales because of the complexity in human psychology. A search in the PsycNET database using the key words 'scale development' and 'test construction' revealed 2,557 articles published under psycARTICLES. The majority of these articles often report the development of new scales, and the psychometric analysis of existing scales. Under the psycINFO classification, there are 2,223 personality scales and inventories alone, not even accounting for other areas within psychology. From 2006 to 2012, the number of articles relating to the key word "test construction" increased from 41 to 97. Clearly, even with the existing scales available, the creation of new psychological assessment still plays a central role within the field.

With the employment of crowdsourcing platforms, we are able to reach out to a wider group of individuals who may not have the experience in writing items but may still have the knowledge to develop the item content of interest. By providing appropriate instructions, even non-expert item writers might be able to produce items at a much faster pace requiring smaller monetary rewards. The crowdsourcing approach effectively skips the process of having to train individuals to an expert level and only provides the necessary information that will suffice in generating the item content.

Since workers aim to strategically maximize their own benefit and rely on the reputation they have in these crowdsourcing platforms (Zhang and van der Schaar 2012), they are typically driven by the financial rewards to participate and perform well in these tasks (Mason and Watts 2010). Fortunately, to limit the number of workers from free-riding and false reporting (Feldman et al. 2004, Zhang et al. 2014), most crowdsourcing platforms include

a reputation mechanism. A straightforward approach in MTurk is that workers can be selected based on the percentage of successful jobs that they have previously completed. Disputes with workers can also be worked out individually, allowing workers to justify their actions, or on the positive side where superlative work has been done, workers can be given individual financial bonuses.

Scaling Creativity Using Collective Intelligence

Crowds are not just cheap labor – they may be higher quality labor, with more creative potential and collective wisdom than individual workers. For example, Surowiecki (2004) has explored the potential of crowds in his book *The Wisdom of Crowds* and concluded that crowds may be remarkably intelligent (see also Kosinski et al. 2012). Brabham (2008) cites several case studies in order to demonstrate that crowdsourcing is a viable way to make use of human creativity, from *Threadless.com*, a company that crowdsources T-shirt designs, to *iStockphoto*, which is a platform selling photographs uploaded by the general public. In fact, the application of crowdsourcing techniques has been so advantageous that it has been employed even in fields such as genealogy (Barreto, Fastovsky, and Sheehan 2003) or astronomy (McLaughlin 2014).

However, the following question remains: even if crowdsourced items are more diverse and varied compared to those created by a handful of experts, do they have satisfactory psychometric qualities? Previous research has shown this is indeed the case. For example, Christoforaki and Ipeirotis (2014) showed that crowdsourcing knowledge items based on item skeletons provided using an existing publicly available knowledge bank (Stackoverflow) produced affordable items which performed equally well or even better than those developed by professional item writers.

A final creative advantage afforded by crowdsourcing is the potential for enhanced cultural and social variety of the generated items. People's life experience in interaction with others may give them an intrinsic ability to create items that are specific to measuring a particular personality trait while accounting for cultural and societal influences. A case study by Kim et al. (2010) highlighted four major characteristics of item writers that influence the item writing process: previous experience, language background, personality and preferences. These characteristics played a large role in determining the type of items that were designed and could lead to limitations and potential biases in the item content. By crowdsourcing item creation to item writers with diverse backgrounds, we are able to reduce the constraints imposed by each item-writer's cultural, social and behavioral background.

Producing Crowdsourced Personality Items

We selected Roid and Haladyna's (1982) generic five-step procedure to guide us in the item writing process as it seemed most appropriate for generating personality items using crowdsourcing platforms. The five-steps are: (1) instructional intent; (2) specifying the domain; (3) item development; (4) item review; and (5) test development.

Step 1: The workers were given explicit instructions on the item specification prior to the creation of an item (Figure 1). In order for the workers on Amazon's MTurk (Mechanical Turk 2009) to create parallel forms of personality items, we presented them with established personality items taken from the 100-item IPIP representation of the Five Factor NEO-PI-R personality scale (Costa and McCrae 1992, Goldberg et al. 2006). The 5 NEO-PI-R measures five broad personality traits: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A) and Neuroticism (N). **Step 2:** There were 100 items in total, with 20 items each measuring a different personality trait. Of the 20 items, 10 items were positively worded and 10 items were negatively worded. Each participant was randomly assigned to create items for a specific trait. They were then presented with 10 positively and 10 negatively worded sample items measuring that specific trait. **Step 3:** To complete the task, participants had to design both a positive and a negative item. At the end of the task, we asked for demographic information such as gender, age, marital status, ethnicity, nationality and first language.

Quality Check

Step 4: The items were reviewed by a group of research psychologists after collection on MTurk. There are currently no recommended guidelines to the evaluation of crowdsourced psychological items. Nevertheless, we employed Hambleton and Rogers' (1991) recommendation by focusing on three features when reviewing item content: (1) item validities, (2) technical quality, and (3) representativeness. We also included a fourth guideline, which is to ensure that the content of the crowdsourced items is to be kept as original as possible. Therefore, an editing document was created with two goals in mind; the first was to make grammatical and stylistic changes where necessary, and the second was to remove items whose content did not make sense with respect to the targeted trait.

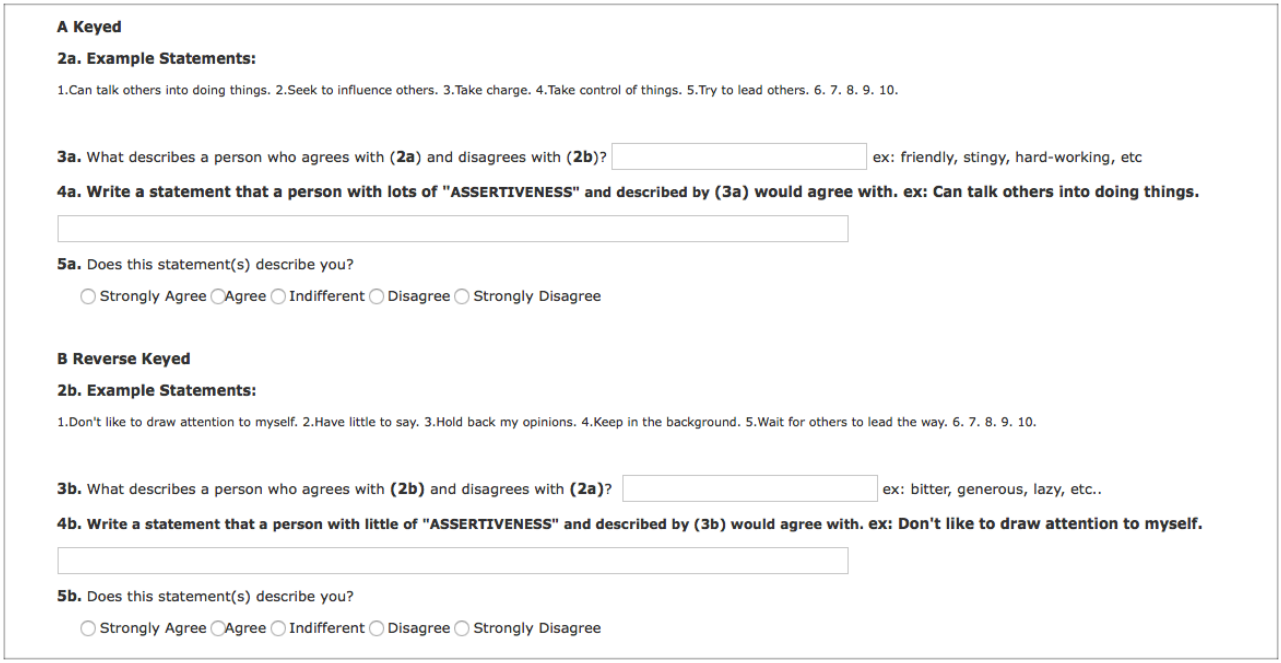

Figure 1. Example of how the task was presented to the worker.

For the first goal, there were ten different editing rules (*e.g. awkward wording, strong language or spelling mistakes*) that the reviewers used as a guide to revise the items. For example, an item measuring conscientiousness submitted by the item-writer as *"I hate burden and work"* was revised to *"I dislike burden and work"* by the reviewer. As for the second goal, six different editing rules (*e.g. no relation to the trait at all, repeated crowdsourced statements or items too culture specific*) were used to remove items. An example may be an item formulated as *"This is a dull book"*, which was removed from the list of crowdsourced items.

Crowdsourced Item Writers

Eighty-two crowdsourced item writers were recruited using MTurk. All MTurk data was collected over a period of two weeks. The workers were allowed to complete the assignment under the condition that they had completed at least 500 assignments on MTurk and had a minimum of 95 per cent completion rate. The workers were allowed no more than 30 minutes to complete each HIT. 41 crowdsourced item writers were females, 13 were males, and 28 did not respond. The crowdsourced item writers' age ranged from 18 to 65, with an average age of 37 years (SD = 10.3). The first language of 48 item writers was English, with 15 item writers not having English as their first language. 32 did not respond to this question. 26 item writers were Americans while 36 item writers were made up of various other nationalities. 20 item writers chose not to specify their nationality.

Result

The average time taken to read the instructions and create both a positive and negative item was 4 minutes and 30 seconds. In total, 164 items were created, with 82 negatively worded and 82 positively worded items.

In the following week, three research psychologists reviewed the crowdsourced items to check the quality of the items measuring the targeted traits with respect to the above-mentioned guidelines. There were no monetary incentives for the research psychologists to partake in the research. Following the two main goals of the editing guidelines, 23 negatively worded items and 18 positively worded items were corrected. 39 items were removed, as they were not designed to measure the targeted trait. What remained were 15 new items for the openness trait; 33 new items for the conscientiousness trait; 30 new items for the extraversion trait; 27 new items for the agreeableness trait; and 20 new items for the neuroticism trait. This resulted in a final set of 125 crowdsourced personality items after qualitative review.

Item Validation

Step 5: For empirical testing and evaluation in a newly collected dataset, validity analysis was conducted by examining the relationship between the crowdsourced and established items. Therefore, data was collected from a total of 225 items including both crowdsourced and

established items. The scale was created using an open source testing platform (Concerto). The analysis was completed in R using the psych package developed by Reveille (2007).

We employed psychometric means for item validation, as it is important to demonstrate the factorial validity of the crowdsourced items for each personality trait (Kline 1993). Therefore, we conducted Exploratory Factor Analysis for each trait separately in order to evaluate the quality of these crowdsourced items. We removed the crowdsourced items with poor psychometric estimates, and maintained the good items within the scale. Following the suggestions made by Kline (1993), items that had a factor loading of less than 0.3 are considered weakly correlated, and thus, should be rejected. Therefore, in our study, items that had factor loadings (>.30) were retained while those that were (< .30) were removed. A 5-factor model was subsequently imposed on the good items to evaluate the overall factor structure of the items.

Cronbach's alpha (Cronbach 1951) was used to measure the internal consistency of the newly developed scale as an indicator for reliability and to compare the alpha index against the established items and those reported in the literature. The internal consistency of a scale is described by the amount of inter-relatedness of the items within the test and is widely used in psychological scale construction (Santos 1999). Every established psychological scale in the literature measuring a particular trait reports an absolute Cronbach's alpha index ranging between 0 and 1. As a convention, Cronbach's alpha has to be at least .7 for the scale to be considered acceptable (Nunnally 1978).

Lastly, we ran a correlation analysis to observe the relationship between the crowdsourced and established items. As it was in the interest of this study to evaluate the potential of using crowdsourcing platforms to create items, we calculated the percentage of the number of items that were retained after quantitative analysis was made.

Exploratory Factor Analysis

We briefly discuss some preliminaries on Exploratory Factor Analysis (EFA) before describing the results from our analysis. At the initial stage of scale construction, EFA is conducted to evaluate the item content and a corresponding factor structure which is a representative of the hypothesized latent construct (Thompson 2004). The procedure is exploratory in nature because the researcher presumably has no a priori expectations based on previous research or theory to support the composition of the latent variables (Cudeck 2000).

One of the foremost goals of EFA is to identify the factor structure and reveal the number of latent variables that best accounts for the common variance among the items (Finch and West 1997). In factor extraction, the shared variance of a variable is separated from its unique variance in order to investigate the structure of the common factor model (Cattell 1978).

To make sense of the initial factor extraction produced by the EFA, rotation algorithms are often used in an attempt to orient the factors into an interpretable position (Velicer and Jackson 1990). Simple structure rotation algorithms such as Varimax, Promax and Oblimin are used so that items are forced to load highly on one or two factors and have close to zero loadings on the remaining factors (Fabrigar et al. 1999).

There are several methods that are often used for investigating matrix dimensionality (i.e. the number of factors to retain). Over and under extraction of factors can have deleterious effects on the results (Zwick and Velicer 1986). According to Courtney and Gordon (2013), Horn's (1965) Parallel Analysis (PA) has emerged as the most recommended technique by several researchers (e.g. Ruscio and Roche 2012, Zwick and Velicer, 1986).

The PA Scree plot illustrates how researchers determine the number of latent variables of the underlying factor structure. It is a graphical representation of the incremental variance accounted for by each factor in the model using either principal component (PC) or factor analysis approaches (FA). The PA scree plot displays the scree for the observed and simulated data. The plot generates the number of dimensions/components on the x-axis and the eigenvalues on the y-axis. The PA method also generates a large number of data matrices from random data. Each matrix is generated parallel to the actual data, which means that the simulated matrix has the same number of cases and variables. The scree of the simulated matrix is often the slope with gentle gradient that is close to the horizontal line at the eigenvalue 1.

The PA plot has two objectives; one is to visually locate the number of factors that best represent the data, and the second is to compare the scree of factors of the observed data to that of a random data matrix of similar size to the original matrix. Generally, factors are retained as long as they are greater than the mean eigenvalues generated from the random data matrices (Ledesma and Valero-More 2007). Often, it can be assumed that components after the point where the retrieved factors form an approximate descending linear line do not significantly account to the percentage of the common variance (Reise, Waller, and Comrey 2000). If the predicted number of factors to extract is four and the scree test suggested five instead, it is advisable to run the analysis setting the number of factors at four, five, six and seven. After rotation, it is suggested to compare the item loadings and select the factor structure that best fits to the data by selecting the model that produces the highest number of item loadings above .30, has few or no item cross loadings, and no factors with fewer than three items (Costello 2009).

Participants

Participants were recruited from MTurk over a period of 2 days. Participants had to complete 125 personality items and they were only allowed to take the test once. All participants had to have a completion rate of 95% for previous assignments. The average time taken to complete the test was 18 minutes. The sample consisted of 52 males and 104 females with a mean age of 42.6 years old (SD = 13.4). 130 participants were Americans and 16 were from other countries. Participants completed both the crowdsourced items and established items as one personality scale. The items were displayed in a random order to reduce any order effects.

Result

Exploratory Factor Analysis performed on the crowdsourced items for each trait indicated that the majority of the items loaded significantly on the first latent factor. Parallel analysis and the scree plot for each trait is best represented by a single latent factor. That is, the EFA result for each trait showed that 24% of the variance is accounted for openness; 34% for conscientiousness; 31% for extraversion; 30% for agreeableness; and 43% for neuroticism.

Investigation of the factor loadings for each personality trait showed that 3 openness items; 3 conscientiousness items; 3 extraversion items, 7 agreeableness items, and 3 neuroticism items achieved a factor loading of less than .3 and were removed from the analysis. We subsequently imposed 5-factor solution of all data combined together and found that the data explained 41% of the total variance (10%, 9%, 9%, 8%, 6%, respectively). The Parallel Analysis Scree plot (Figure 2) inspection supported the extraction of a five-factor personality model from the data.

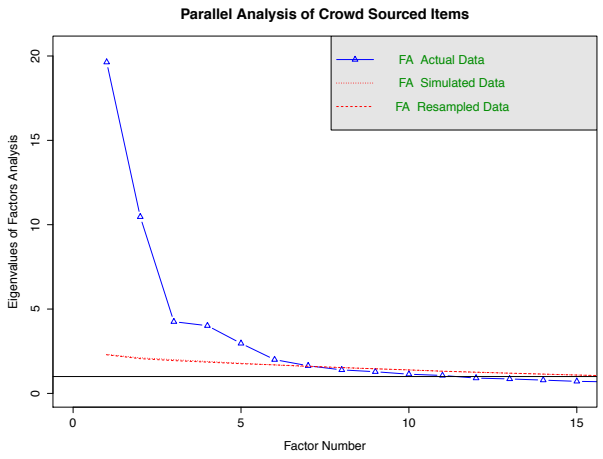

Figure 2. Illustrations of a Parallel Analysis Scree plot based on the observed and simulated data. The plot displays the scree of factor analysis. FA = Factor analysis.

However, evaluation of the item properties with an imposed 5-factor model showed evidence of significant item cross-loadings. This suggests that an ambiguous rather than a clear factor structure existed in the observed data.

After removing the crowdsourced items that had poor psychometric properties, we evaluated the number of items that were finally retained. 80% of the Openness items, 90% of the Conscientiousness items, 90% of the Extraversion items, 74% of the Agreeableness items, and 85% of the Neuroticism items were retained (Table 1). Taken together, 96 (84.8%) items out of the 125 reviewed crowdsourced items had adequate psychometric properties, which could be used as a measure of the Big Five personality model.

Scale	Revised Items	Validated items	Percentage
O	15	12	80%
C	33	30	90%
E	30	27	90%
A	27	20	74%
N	20	17	85%

Table 1. Percentage of items recovered after empirical validation.

The internal reliability (Table 2) of the scales consisting of crowdsourced items for each of the traits was compared to that of the established items and of previously reported scales in the literature (Costa and McCrae 1992). Conscientiousness had the highest Cronbach's alpha (α = 0.94), followed by Extraversion (α = 0.92), Openness (α = 0.89), Agreeableness (α = 0.85), and Neuroticism (α = 0.78).

	Cronbach's Alpha Index		
Scale	NEO-PI-R Published Scale	Established Items	Crowdsourced Items
O	0.85 (n=20)	0.86 (n=20)	0.89 (n=12)
C	0.90 (n=20)	0.94 (n=20)	0.94 (n=30)
E	0.91 (n=20)	0.92 (n=20)	0.92 (n=27)
A	0.91 (n=20)	0.94 (n=20)	0.85 (n=20)
N	0.89 (n=20)	0.88 (n=20)	0.78 (n=17)

Table 2. Internal consistency of scales reported for established items and crowdsourced items for each personality trait. The Cronbach's alpha that was reported in the literature (Costa and McCrae 1992) is also presented in the table. Cronbach's alpha is affected by the length of a scale; hence the number of items is also presented.

Finally, intercorrelations using factor scores and mean scores were computed between the crowdsourced and

established items for each trait respectively (Table 3). As expected, scores comparing the same personality traits were highly associated, with Neuroticism (r=0.91, p<0.01) being the highest, followed by Extraversion (r=0.90, p<0.01), Conscientiousness (r=0.89, p<0.01), Openness (r=0.82, p<0.01) and Agreeableness (r=0.82, p<0.01). Not accounting for same trait intercorrelation, absolute discriminant correlations ranged from 0.08 to 0.63. Overall, the five scales displayed clear evidence of discriminant validity.

		Established Items				
	Scale	O	C	E	A	N
Crowd-sourced Items	O	**0.82****	0.26**	0.30**	0.41**	-0.27**
	C	0.28**	**0.89****	0.24*	0.58**	-0.44**
	E	0.19*	0.22*	**0.90****	0.08	-0.48**
	A	0.38**	0.49**	0.17	**0.82****	-0.25**
	N	-0.20*	-0.63**	-0.54**	-0.39**	**0.91****

*Table 3. Intercorrelation between the factor scores of the crowdsourced items derived from an EFA and the scale scores from the established items. * Significant at the .05 level. ** Significant at the .01 level.*

Discussion

This research investigates the plausibility of employing crowdsourcing platforms to create personality items with adequate psychometric estimates. It took a period of two weeks to generate 164 items and another week for the group of three research psychologists to review the items. In the test phase, we spent a week to develop the personality scale and a further two days to collect data. After data collection, the analysis was completed in two weeks. In total, the item development, testing and analysis phase took between 6 to 7 weeks to complete.

Crowdsourcing platform has given us the opportunity to facility the collaborative work between expert and non-expert item writers. While non-experts as demonstrated in this paper can create new items, the expertise of professional item writers is still essential. Non-experts provide a source of diversity in the creation of the item content. However, the experience and knowledge of the expert is still critical to the evaluation of the item quality.

A five-factor structure was supported based on the observed data. However, item cross loadings were observed upon further inspection. Such findings in personality scale research are not unusual as EFA determines the factors based on empirical deduction approaches without accounting for theoretical concepts (e.g. Egan, Deary, and Austin 2000, Tierno et al. 1995). A follow-up study employing Confirmatory Factor Analysis (Harrington 2008) must be conducted with a larger sample size to investigate the factorial structure of the Big Five personality model with the current crowdsourced items. This will provide a more stringent approach at an item level in deciding which items should be removed from the model.

Personality traits	Statements
Agreeableness	Care about myself first and others second.
Agreeableness	Am happy if others are happy.
Conscientiousness	Have a good work ethic.
Conscientiousness	Do not see the point in trying hard.
Extraversion	Don't like to draw attention to myself.
Extraversion	Thrive in almost all social scenarios.
Neuroticism	Am afraid of a lot of things.
Neuroticism	Can process life objectively and positively.
Openness to experience	Like to get new information.
Openness to experience	Tend to be an explorer in all things I do.

Table 4. Sample items of Crowdsourced Big 5 personality items.

Nonetheless, correlational comparison between the crowdsourced items and the established items yield evidence of the two scales' convergent validity as measures of the Big Five personality model. Additional associations between personality traits are consistent with past research. For example, negative correlations between Neuroticism and Extraversion are commonly reported (Vitterso 2001). Similar associations were found between crowdsourced Neuroticism - established Extraversion (r= -.54) and vice versa (r= -.44) (Table 3). Therefore, the pattern of convergent and divergent correlation suggests that the scales for both the crowdsourced items and established items measure the Big Five personality traits.

Overall, of the 164 items that were crowdsourced, 39 items were removed during the reviewer phase, and 16 items were dropped due to poor psychometric properties. This resulted in a final set of 96 (59%) items. The result is similar to what was reported by Haladyna (2012), where 60% of the items developed by professional item writers remained after psychometric testing. This research supports the notion that once clear instructions are written and provided, crowdsourced non-experts have the potential to create high quality personality items. The payment for workers in this study was lower than the suggested rate on Dynamo Wiki[1], which could have led to selection bias and/or slower response time. Therefore, the quality of generated items could be affected by the payment rates,

[1]http://wiki.wearedynamo.org/index.php?title=Guidelines_for_Academic_Requesters

resulting in 24% of the crowdsourced items not passing the initial expert review phase.

While crowds can generate usable psychometric items, it would not be appropriate to use crowdsourcing for the highest stakes tests where item security is paramount, such as in government-run examinations. Given that the public is creating items, it would be possible for individuals to keep a record of the items they create and to make them available in future. We envisage that crowdsourcing of new items is most appropriate for tests where the stakes are low enough that item writers are not incentivized to find out how their items are going to be used in practice. For example, most personality tests are taken by a small enough group of job candidates that it is unlikely that crowd participants would become aware of which company would be using the new items. Additionally, it will be even more advantageous under CAT conditions for recruiters. Larger item banks reduce the exposure rate of the items and increase test security. This indirectly makes it more difficult for participants to cheat or fake in order to receive more favorable outcomes. Another situation where crowdsourcing items would be particularly valuable is for Massive Open Online Courses (MOOCs), which need to test large numbers of participants but where the stakes are relatively low. MOOCs have large numbers of course participants who could be incentivized to create new questions for the next cohort.

Our study is limited by the geography of its crowd: the majority of the item writers and participants were from the USA. Hence, the result of the psychometric analysis may only reflect items piloted on the American population. Future research should investigate any cultural biases of these items. A potential avenue of psycho-geographical research is to investigate the quality and diversity of the crowdsourced items developed by item-writers from different locations, both from within and between countries.

In future, the crowdsourcing process for psychological test items can be improved by perfecting the incentives for item writers and understanding which personality traits improve the quality of item writing. In our experiment, item writers were paid a flat rate regardless of the psychometric properties of the items they generated. Reviewers only evaluated the quality of the items after the item development phase. As such, none of the item writers received feedback from the work they had done nor did they receive any kind of extra financial reward that was dependent on the quality of the work they had done. Such an approach can be improved in the long run by increasing the motivation for future item writers to write high quality items (Yu et al. 2014). Future research should also evaluate the quality of crowd source psychometric items for a given standardized pay rate. This should be balanced against the finding that simply offering higher payment does not automatically increase the quality of work (Mason and Watts 2010). Online markets rely on reputation structure for giving trust to the workers. Thus, introducing a reputation system will allow us to increase the financial reward to item writers who write high quality items. We can offer to train these reputable item writers to create more complicated item content and provide them with feedback on their work. Moreover, previous research indicated that personality traits are significant predictors to job performance (Barrick and Mount 1991). The significance of the predictors are however, varied by occupational group and criterion type. Hence, it would be beneficial to evaluate potential item writers based on their personality as one of the criteria for selection. Finally, another potential avenue of research is to investigate the degree to which the crowd can be trained to conduct the quality check phase during the item development process. The optimal trade-off between high quality and diverse items is likely to be dependent on how the quality check phase is conducted. Reviewers need to consider the subtle differences in item content and meaning to be able to retain high quality items, which may be difficult for those with a lack of experience in item writing. These differences, if undetected, could lead to either overly high levels of homogeneous items being retained or instead, having many more items being discarded than expected.

Conclusion

The current research has shown that we can recruit item writers through crowdsourcing platforms to generate high quality personality test items in a fast and effective manner. This result means that new avenues of research in item development can be explored. An area that we envisage to be directly applicable is in health assessment, where the constructs are highly subjective in nature (Miller et al. 2009). Using the collective knowledge of the crowd may allow us to gain insights into ways in which more content can be created to measure a specific health-related construct.

Test translation is expensive due to the many additional steps and extensive review incorporated into the process. The cost of translating and adapting an existing test to a different population often costs as much as developing a new test (Stansfield 2003). However, it could be possible to take an alternative approach. Instead of translating an existing test, we could crowdsource item writers at a low cost to develop completely new items aimed at the new language and culture. Combining current approaches to designing psychological items with crowdsourcing methodology could be extremely beneficial to the future of item development.

Acknowledgements

We would like to thank the Cambridge Psychometric Centre team for their valuable comments. We also thank all the participants and item writers in our experiment.

References

Barreto, C., Fastovsky, D. E., and Sheehan, P. M. 2003. A model for integrating the public into scientific research. *Journal of Geoscience Education 51*(1):71-75.

Barrick, M. R., & Mount, M. K. 1991. The big five personality dimensions and job performance: a meta- analysis. *Personnel psychology*, *44*(1), 1-26.

Brabham, D. C. 2008. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: The International Journal of Research Into New Media Technologies 14*(1):75-90.

Cattell, R. B. 1978. *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Christoforaki, M., and Ipeirotis, P. 2014. STEP: A Scalable Testing and Evaluation Platform. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Clark, L. A., & Watson, D. 1995. Constructing validity: Basic issues in objective scale development. *Psychological assessment*, *7*(3), 309.

Costa, P. T., Jr., and McCrae, R. R. 1992. *Revised NEO Personality Inventory (Neo-PI-R) and NEO Five-Factor Inventory (NEO-FFI): Professional manual*. Odessa, FL: Psychological Assessment Resources.

Costello, A. B. 2009. Getting the most from your analysis. *Pan*, *12*(2):131-146.

Coughlan, S. Parents keep children off school in test protest. *http://www.bbc.com/news/education-36188634.* Retrieved 03 May 2016.

Courtney, M.G.R. and Gordon, M., 2013. Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2. 0 to make more judicious estimations. *Practical Assessment, Research & Evaluation*, *18*(8), pp.1-14.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika, 16(3)*:297-334.

Cudeck, R. 2000. Exploratory factor analysis. *Handbook of applied multivariate statistics and mathematical modeling*. 265–296.

Egan, V., Deary, I., and Austin, E. 2000. The NEO-FFI: Emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. *Personality and Individual Differences 29*(5):907-920.

Feldman, M., Papadimitriou, C., Chuang, J., and Stoica, I. 2004. Free-riding and whitewashing in peer-to-peer systems. In *Proceedings of the ACM SIGCOMM workshop on Practice and theory of incentives in networked systems*, 228-236. New York: ACM.

Finch, J. F., and West, S. G. 1997. The investigation of personality structure: Statistical models. *Journal of Research in Personality 31*(4):439-485.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C., 2006. The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Gibbons, R.D., Weiss, D.J., Kupfer, D.J., Frank, E., Fagiolini, A., Grochocinski, V.J., Bhaumik, D.K., Stover, A., Bock, R.D. and Immekus, J.C., 2008. Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*.

Haladyna, T. M., and Rodriguez, M. C. 2013. *Developing and validating test items*. Routledge.

Haladyna, T. M. 2012. *Developing and validating multiple-choice test items*. Abingdon-on-Thames: Routledge.

Hambleton, R. K., and Rogers, H. J. 1991. Advances in criterion-referenced measurement. In *Advances in educational and psychological testing: Theory and Applications*, 3-43. Dordrecht: Springer Netherlands.

Harrington, D., 2008. *Confirmatory factor analysis*. Oxford University Press, USA.

Horn, J.L., 1965. A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), pp.179-185.

International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences http://ipip.ori.org/newNEODomainsKey.htm. Retrieved 03 May 2016.

Kosinski, M., Bachrach, Y., Kasneci, G., Van-Gael, J. and Graepel, T., 2012, June. Crowd IQ: Measuring the intelligence of crowdsourcing platforms. In *Proceedings of the 4th Annual ACM Web Science Conference* (pp. 151-160). ACM.

Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., and Roullion, V. 2010. A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics. *Language Assessment Quarterly 7*(2):160-174.

Kleemann, F., Voß, G. G., and Rieder, K. 2008. Un (der) paid innovators: The commercial utilization of consumer work through crowdsourcing. *Science, Technology & Innovation Studies 4*(1):5-26.

Kline P. 1993. *The Handbook of Psychological Testing*. Routledge: London.

Ledesma, R. D., and Valero-Mora, P., 2007. Determining the number of factors to retain in EFA: An easy-to-use computer program for carrying out parallel analysis. *Practical Assessment, Research & Evaluation 12*(2):1-11.

Mason, W. and Watts, D. J. 2010. Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter 11*(2):100-108.

McLaughlin, E. "Image Overload: Help us sort it all out, NASA requests". *Cnn.com* (CNN). Retrieved 18 September 2014.

Miller, V. A., Reynolds, W. W., Ittenbach, R. F., Luce, M. F., Beauchamp, T. L., & Nelson, R. M. 2009. Challenges in measuring a new construct: Perception of voluntariness for research and treatment decision making. *Journal of Empirical Research on Human Research Ethics*, *4*(3), 21-31.

Nunnally, J. C. 1978. *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Reise, S. P., Waller, N. G., and Comrey, A. L. 2000. Factor analysis and scale revision. *Psychological Assessment 12*(3):287-297.

Revelle, W. and Revelle, M.W., 2007. The psych Package.

Roid, G. H., and Haladyna, T. M. 1982. *Toward a technology of test-item writing*. New York: Academic.

Ruscio, J. and Roche, B., 2012. Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological assessment*, *24*(2), p.282.

Rudner, L. M. 2009. Implementing the graduate management admission test computerized adaptive test. In *Elements of adaptive testing*, 151-165. New York: Springer.

Santos, J. R. A. 1999. Cronbach's alpha: A tool for assessing the reliability of scales. *Journal of Extension 37*(2):1-5.

Sharpley, C.F. and Rogers, H.J., 1985. Naive versus sophisticated item-writers for the assessment of anxiety. *Journal of Clinical Psychology*.

Stansfield, C.W., 2003. Test translation and adaptation in public education in the USA. *Language Testing*, *20*(2), pp.189-207.

Surowiecki, J. 2005. *The wisdom of crowds*. Anchor.

Thompson, B. 2004. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

Tierno, T., Zebrowski, S., Hofmann, V., Edelstein, W., and Hart, D. 1995. *The stability of the Five Factor Model personality dimensions from age seven to age nineteen*. Paper presented at the annual meeting of the American Psychological Society, New York.

van der Linden, W. J., and Glas, C. A. eds. 2000. *Computerized adaptive testing: Theory and practice*. Dordrecht: Kluwer Academic.

Vitterso, J. 2001. Personality traits and subjective well-being: Emotional stability, not extraversion, is probably the important predictor. *Personality and Individual Differences 31*(6):903-914.

Yu, L., André, P., Kittur, A., and Kraut, R. 2014. A comparison of social, learning, and financial strategies on crowd engagement and output quality. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 967-978. New York: ACM.

Zhang, X., Xue, G., Yu, R., Yang, D., and Tang, J. 2014. You better be honest: Discouraging free-riding and false-reporting in mobile crowdsourcing. In *Global Communications Conference (GLOBECOM), 2014 IEEE*, 4971-4976. IEEE.

Zhang, Y., and van der Schaar, M. 2012. Reputation-based incentive protocols in crowdsourcing applications. In *INFOCOM, 2012 Proceedings IEEE*, 2140-2148. IEEE.

Zwick, W. R., and Velicer, W. F. 1982. Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research 17*(2):253-269.