Philippine Journal of Psychology, 2016, 49 (1), 19-42

The Development and Validation of a Filipino Social Desirability Scale

Louie P. Cagasan Jr.
University of the Philippines, Diliman, Quezon City, Philippines

This paper is composed of two studies that describe the construction and validation of a Filipino Social Desirability (SD) Scale. Study 1 details the phases in developing the SD scale: item writing, item selection, and cross validation. In the item-selection phase, twelve of the twenty-six candidate items were selected based on a number of criteria. One is the correlation of items with self-criterion residuals, defined as the discrepancy between self-report scale scores and an objective criterion, in this case, peer-rating on the same scale. Residuals were generated from the five domain scores of the Mapa ng Loob. Other criteria were psychometric properties of the items and ratings of experts and target participants on the appropriateness of the items. On a sample with $n = 157$, the test-reliability of the scale was found to be .706. In the cross validation phase, SD scale scores (from 12 items) were found to be significantly correlated with self-peer discrepancies on Neuroticism, Agreeableness, and Conscientiousness. The scale reliability was computed at .731 ($n = 162$). In Study 2, convergent validity of the local SD scale was examined. Results showed that the Filipino Social Desirability Scale was significantly correlated with Paulhus' Balanced Inventory of Desirable Responding and the Marlowe-Crowne SD Scale.

Keywords: social desirability, indigenous measure, self-criterion residuals, personality traits

Getting the truth from a person may not be an easy task. It involves not only extraction of information but also identification of its validity. Having a way to determine faking and who is doing it would come in handy. In psychological research, quantifying bias or "faking" through a test is one of the ways of dealing with this threat to truth. This bias is commonly labeled as social desirability (SD), defined as the "tendency to give overly positive self-descriptions" (Paulhus, 2002, p. 50).

There exists a number of measures and approaches to social desirability, and Paulhus (2002) categorized these into three: minimalist, elaborate, and accuracy constructs. The minimalist constructs characterized social desirability in a "straightforward" manner, and theoretical explanations behind it are not that detailed. Among the common methodologies in this typology are deriving SD tests based on the ratings of expert judges and the use of contrasting groups (or role-playing) where one compares the criterion group (faking-good) with a control group (straight-take). The prominent social desirability questionnaires under the first approach are Edward's SD Scale and Wiggin's SD Scale. The elaborate constructs, on the other hand, pertain to SD measures that have theoretical underpinnings at the very start of the development phase, and establishment of the test validity has led to providing detailed feature of the construct. The known questionnaires under this approach are the Marlowe-Crowne SD Scale and the lie scale of Eysenck Personality Inventory. In accuracy constructs, social desirability is considered not as an embodiment of distortion but as an exact equivalence of the test results. That is, those scoring high in SD measures actually possess positive attributes. Among the prominent proponents of this approach are Block (as cited in Paulhus, 2002) and McCrae and Costa (1983).

As research in this area advances, there have been two contradicting conceptions about social desirability. In the 1950s, social desirability is commonly viewed as an error or nuisance in self-report assessment (Edwards, 1957). Later on, some researchers believe that social desirability has "substance" and correction for SD should be discouraged (McCrae & Costa, 1983; Ones, Viswesvaran, & Reiss, 1996; Piedmont, McCrae, Riemann, & Angleitner, 2000). To date, the issue of whether social desirability is an error or substantive construct is yet to be resolved. Aside from these issues, cultural robustness of social desirability tests needs to be established. In the study of Li and Reb (2009), he found that the Balanced Inventory of Desirable Responding (BIDR, a commonly used SD scale) may not function similarly for western and eastern subjects, and he explained it in terms of linguistic

and context-related factors. In particular, some items may be interpreted differently and/or irrelevant to people of Asian countries. The phrase *lose out on things* in the item, *I sometimes lose out on things because I can't make up my mind soon*, may not be easily understood by the Asian sample the same way westerners comprehend it. Also, the item, *I sometimes drive faster than the speed limit*, may be inappropriate because of the reliance on public transportation of most people in the sample representing Asian nations. The explanations of Li and Reb are quite valid even for Filipinos.

This paper aimed to develop a social desirability scale that is appropriate for Filipinos. With existing conceptual issues on whether social desirability is an error or substantive construct, the best way to go is through a minimalist approach, because working with an elaborate viewpoint may leave us hanging with SD's unresolved issues. It should be noted that a social desirability scale for Filipinos (Felipe, 1969) existed for more than 40 years. However, there have been no developments on that scale since then. This provides more reason to develop a new scale wherein significant advances in the field are integrated.

Criterion Problem

Psychological tests commonly are self-reports; that is, data come from the individual being assessed. Its validity is usually established through the judgment of others (e.g., peers, family). If the sources say the same thing, then we can be sure that self-report information is valid. This methodology is also known as the social consensus criterion in Robins and John's (1997) categorization. What makes research in social desirability different is that it is supposed to be a measure of bias of self-reports. By its very nature (bias on self-report questionnaire), agreement of self-report social desirability and ratings of knowledgeable people on the same scale would not make sense because it is out of social desirability's scope. A direct operationalization of social desirability was first introduced by Paulhus and John (as cited in Paulhus & John, 1998); this is known as the self-criterion residual. In this particular method, the rating of an individual is compared to a more objective criterion such as peer rating. Self-report scores are then regressed on the peer-rating scores. Discrepancy between the perception of an individual and his predicted score is the bias index, and this is specifically quantified by the residual obtained from the regression analysis. Here, the quantitative index and the definition (bias of self-reports) of social desirability directly

correspond with each other. In this research, self-criterion residual serves as the objective criterion of the social desirability test being developed.

Test Development Plan

Given that the nature of social desirability is yet to be resolved, it is logical to adhere to a simple working definition that can be quantified in some way. For this study, the definition of Paulhus (2002) is used; social desirability is "the tendency to give overly positive self-descriptions" (p. 50). He also believes that "no SDR [socially desirable responding] measure should be used without sufficient evidence that high scores indicate a departure from reality" (p. 50). Self-criterion residual is the quantitative representation of bias, and significant correlation of SD item and scale scores with this index can provide evidence of item and test validity.

This planned procedure is deemed important because when a person answers "Strongly Agree" to a sample SD item, *Mababuting bagay lang ang hangad ko sa aking kapwa,* it will not automatically follow that the person is indeed responding in a socially desirable way. Several factors can explain the responses to the item. Significant correlation of the social desirability items with the self-criterion residuals would ensure to some extent that what we are measuring is indeed social desirability. Even at the item level, there is a certain level of assurance that what is being tapped is the discrepancy between the self and peer (or bias) and not other person-related factors. This validity evidence closely matched the definition of SD in this study.

Moreover, with SD's nature as a bias in self-report questionnaire, experts cannot accurately validate the items through content inspection. Construct validity of social desirability can never be established through experts' opinion about the item, but one can check the appropriateness of a social desirability item for a specific group of people. This is considered as one of the criteria in test development.

To meet the objective of this paper, two studies were conducted. Study 1 aimed to develop a social desirability scale by executing the plan mentioned above and by validating the psychometric results with a different set of samples. This involves three phases: (1) item writing, (2) item selection for the social desirability test, and (3) cross validation of the social desirability test. Study 2 examines the convergence validity of the local social desirability test with existing foreign SD scale.

STUDY 1: Constructing the Social Desirability Test

The researcher had a preliminary version of the social desirability test consisting of eight items from his pilot studies (Cagasan, 2012). Study 1 focused on increasing the items to ensure that the scale would have an equal number of positively- and negatively-keyed items. Aside from that, several criteria were established to ensure the integrity of the items and test in general. Items satisfying the criteria were included in the final version of the social desirability test. The test developed was again evaluated using a different sample to ensure its reliability.

Phase 1: Item Writing

All items were written according to the criteria of Crowne and Marlowe (1960; 1964) and Paulhus (1998). That is, positively-keyed items are "culturally acceptable but improbable" or "desirable but rare", and negatively-stated items are characterized in the opposite direction. These should be "culturally unacceptable but probable" or "undesirable but common". The preliminary version of the test has eight items (or base items). These were again reviewed which led to having four test items or rewritten base items that are assumed to be simpler and easier to understand. Fourteen new items were written to ensure that the scale would have an equal number of positively- and negatively-keyed items. The assembled scale composed of 26 items.

Phase 2: Item Selection

In the next phase, the 26-item scale was subjected to psychometric and qualitative evaluation. Different criteria were established to ensure the integrity of the items and the test in general. Items that satisfy the criteria composed the preliminary version of the test.

Established Criteria. A "good" social desirability item would have to satisfy several criteria for inclusion in the final version of the scale. Firstly, items should show evidence that they indeed measure departure from reality. A correlation between an SD item and self-criterion residual would provide support that the item exemplifies attribution of positive traits and/or disclaiming of negative traits. Any measure can be used to represent self-criterion residuals as long as there exists self and peer data. For this study, discrepancy on personality test scores was used as the current trends of

research on social desirability links it to personality (Backstrom, 2007; Backstrom, Bjorklund, & Larsson, 2009) especially in trying to identify whether SD is bias or substantive construct (Pauls & Stemmler, 2003; Holden & Passey, 2010). Also, the initial data can be linked to research trends, and this provides a good basis in exploring and understanding social desirability later on.

Basing on the desirable pole of the personality domains, it is expected that high SD scorers would present themselves to be high on Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C) and low on Neuroticism (N). Social desirability items should manifest departure from reality by having at least one significant correlation (on the hypothesized direction) with the self-criterion residuals from the personality domains.

Secondly, social desirability items should show evidence of face validity. All the items were evaluated by two groups of raters, experts in Personality and Filipino Psychology (n = 4) and target respondents (a small group of college students; n = 28). Using a 5-point Likert scale ranging from Strongly Disagree (1) to Strongly Agree (5), raters would have to identify the degree to which the item exemplified the given criteria. That is, positively-keyed items should depict "culturally acceptable but improbable" or "desirable but rare", and negatively-keyed items should portray "culturally acceptable but improbable" or "desirable but rare". The mean rating value across two groups should be greater than 3.0 for an SD item to be included in the final form.

Thirdly, items should provide relative contribution to the overall scale. This criterion entailed computation of corrected item–total correlation and Cronbach's alpha if item deleted. Items with corrected item–total correlation value of 0.30 and above are considered for final-scale inclusion. Fourthly, items should provide unique information to the overall scale in terms of content. Although most of the items written were not entirely similar, items that were included in the final scale should at least provide different contexts. This was done to avoid having highly specific items which were incompatible with content validity. Lastly, there should be approximately the same number of positively- and negatively-keyed items to avoid the acquiescence response set.

METHODS

Materials

Masaklaw na Panukat ng Loob (Del Pilar, Sio, Cagasan, Siy, & Galang, 2015)**.** Masaklaw na Panukat ng Loob (Mapa ng Loob) is a Filipino personality inventory that was constructed based on the Five Factor Model. Each factor – Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism – is composed of four facets that are chosen based on their usefulness and applicability in industry and workplace, school, and even clinical settings. The personality domains have good reliability values ranging from .81 to .90 (M = .87). Reliability values of Mapa ng Loob were also generated for the current sample. The mean value for the self-report version was .82 with coefficients ranging from .74 to .87. A peer-report version of the scale was constructed for this study, and this was accomplished by changing the point of view of Mapa ng Loob from first person to third person. The reliability of the peer-report version ranged from .71 to .91 with a mean value of .83. Table 1 shows the reliability coefficients.

Table 1. Reliability coefficients of the Personality facets

Personality face/domain	Cronbach's Alpha			No. of items
	Manual	Self-report	Peer-report	
Neuroticms	90	85	82	32
Extraversion	90	87	86	32
Openness	81	74	71	32
Agreeableness	84	78	85	32
Conscientiousness	90	87	91	32

*Test Booklet and Answer Sheet***.** This research had a paper-and-pen administration format. In particular, a test booklet containing all the items was assembled, and it had a corresponding scannable answer sheet where all responses were recorded. For the self-report test booklet, the 26 social desirability items were interleaved in the 160-item Mapa ng Loob. The peer-report test booklet, on the other hand, had the 160 Mapa-ng-Loob items first, and the 26 social desirability items followed. The 186 statements were rated using a five-point Likert scale, ranging from Lubos na Di-Sumasang-

ayon (1) to Lubos na Sumasang-ayon (5). Questions describing the relationship of the self and peer raters followed the personality and social desirability items.

Data-gathering Procedures

The study was conducted in a classroom setting, and the researcher was given an hour to handle and facilitate the dynamics inside the class. The researcher introduced the activity as a study on personality and attitude. Participants were instructed to form pairs and choose a partner whom they are familiar with. They were also asked to sit beside their partners. The researcher acknowledged the possibility of having students with no partners, so while some of the participants moved to get to their target partners, the researcher mentioned that it would be okay to have no partners, and if anyone cannot find one, they were instructed to sit at the back towards the right side of the classroom.

Paired participants were asked to decide who among them would be part of the "ako" and "ikaw" groups and to come up with a five-digit number that would be used later to link paired data. After that, the paired respondents were asked to sit apart. The classroom was mainly divided between the right and the left sides. Those who identified themselves as "ako" were asked to sit on the right side and respondents in the "ikaw" group on the left side. After that, it was explained to them that individuals in the "ako" group would answer a personality test, and participants in the "ikaw" group would describe their partners in the "ako" group by answering the peer-report version of the personality test. After this brief explanation, the booklet and answer sheets were distributed to the participants. They were also encouraged to answer the questions as honestly as possible, and it was emphasized that the researcher has no way to trace their real identity given that the answer sheet does not ask for names or student numbers.

Some students inside the classrooms had no partners and were instructed to answer the self-report version of Mapa ng Loob and the SD scale. This is one of the reasons behind the different sample sizes reported later on. Either these people have no qualified partner (people who know them well) or the class size was not even leading to one student having no partner. After the participants completed the test, the researcher expressed his gratitude to the students and explained briefly the objectives of the study.

Appropriate Sample Size

The sample size needed to ensure the integrity of the correlation results between social desirability items and self-criterion residuals of personality domains was computed. Pilot-study data were used to calculate the effect size. In getting this statistic, the highest squared correlation values of social-desirability items and self-criterion residuals were identified, and the average of these values was used as the estimate of the effect size, which is 0.114. There is no consensus about the appropriate power; but what is definite is that it should be above .50, and .80 is commonly used in the literature (Murphy & Myors, 2004). Using G-power 3.1 with alpha set at .05 (2-tailed), power at .80, and effect size at 0.114, the computed minimum sample size is 80 for the correlation procedures.

Data Screening for Paired Data

A qualifying criterion was used for peer data to be considered valid; that is, peers should have "enough capacity" to rate their partners. This was measured through their response on the item, "do you think you have enough capacity to rate your partner?". Peer raters who answered "no" (n = 39) or did not answer this question (n = 2) were removed from the paired data. The excluded data together with the cases that cannot be matched because of missing, incorrect, or duplicating pairing numbers (n = 27) and respondents with no partners in the administration were combined and labeled as independent data. This procedure resulted to 194 valid pairs, 125 independent data, and 319 self-report data (194 + 125).

Data Pool

The data for this study were generated from two state universities in the province. The first school is a highly selective institution located in Southern Luzon, and the other is a vocational school in Central Luzon. The data collected were randomly divided into two. The first half was used in selecting the social desirability items against a set of criteria (item selection procedures), and the other half was used to cross-validate the psychometric properties of the developed scale (cross validation procedures). Table 2 shows the breakdown of participants according to data type.

Table 2. Sample size for Each Data Category

Result of Division	Paired Data	Independent Data*	Sef-report Data**
Item selection	96	61	157
Gross validation	98	64	162
Total	194	125	319

**Data constitutes the self-report data coming from cases who did not qualify in the paired data (because of set criteria or pairing number issues) and those with no partner in the administration.
**Data is composed of self-report data coming from the pired and independent data

Participants

Participants for Self-Report Data. There were about 157 students who completed the self-report test booklet. About sixty-four percent of them were from a state university in Central Luzon, while the remaining were from a state university in Southern Luzon. Most of them (63.1%) are females, and about 33.1% of them are males. A portion of them (3.8%) did not indicate their gender. Their age ranged from 16 to 36 years with a mean of 18.46 and standard deviation of 2.236. A huge percentage of the participants was in third year (67.5%) and second year (24.2%) college levels.

Participants for Paired Data. Ninety-six valid pairs of self and peer respondents composed the paired data for Study 1. Seventy-four percent of them were from a state university in Central Luzon. For the self-raters, there are about 66.7% females, 30.2% males, and 3.1% unspecified gender. Their age ranged from 16 to 36 years old (M = 18.49; SD = 2.323). Most of them were third (76%) and second year (18.8%) college students. A number of them were taking Education (35.4%), Forestry (17.7%), and Architecture (17.7%). Most of them indicated that their partner for this study was their "friend" (86.5%), and the length of their acquaintanceship ranged from 0.20 to 10 years with a mean of 2.59 years (SD = 1.572). They were also asked to estimate their partners' knowledge about them using a 10-point scale with 1 as not knowledgeable (*hindi kilala*) and 10 as very knowledgeable (*kilalang kilala)*. The mean of their ratings was 7.37 (SD = 1.582).

The other raters (or peers who rated their partners) were mostly females (64.6%). They have an age range of 16 to 22 years old with a mean of 18.16 years (*SD* = 1.123). Most of them indicated that the person they were rating was their "friend" (86.5%) and their reported length of acquaintanceship ranged from 0.17 to 12 years with a mean of 2.75 years (*SD* = 1.904). The degree of their reported knowledge about the person they were rating had a mean of 7.63 (*SD* = 1.297). Results from the two parties (the self and partner) somehow validated each other's data, as significant correlations were obtained for length of acquaintanceship ($r = .914$, $p < .01$) and degree of familiarity ($r = .322$, $p < .01$).

RESULTS

Item Correlation with Self-Criterion Residuals

A good social desirability item should be significantly and logically correlated with at least one of the self-criterion residuals on personality domains. Table 3 shows the item-level results for Phase 2. For this particular criterion, most of the items (22 out of 26 items) displayed acceptable correlation with a self-criterion residual. Absolute values of significant correlation coefficients ranged from .203 to .467 with a mean of .287 (*SD* = .068). There were three items that did not have any significant and logical relationship with the self-criterion residuals. These are new item 3 (*Lahat ng ginagawa ko ay tama*), new item 12 (*Minsan ay nakiki-ayon ako sa ginagawa ng mga kaibigan ko, mabuti man ito o masama*), and revised base item 8 (*Hindi ako nagsisinungaling kahit pa para ito sa ikabubuti ng iba*). One item (base item 1) had a significant and negative correlation with self-criterion residual on openness. This trend was found to be illogical, resulting for this item to be flagged as bad.

Ratings on Appropriateness and Understandability

An acceptable item should have a mean rating value greater than 3.0 for the two groups of raters (experts and students). Seven items did not satisfy this condition. It is interesting to note that raters had different perceptions on some items. Particularly, experts recognized new item 3 as inappropriate, but student raters found it as acceptable. In a similar manner, student raters considered new item 10 (*May panahon na iginigiit ko ang gusto ko kahit makasakit ako ng iba*) as unacceptable, but experts did not

Table 3. Phase 2 Item Characteristics (Item Statistics (n = 157)

Item	Item Keying	Item Statistics (*n* = 157)				Item Correlation with	
		Mean	Std.	Item	Alpha w/o the item	("r) N	(+r) E
Base 1	Negative	3.25	1.196	.39	.810	229*	.079
R-B1	Negative	3.40	1.031	.42	.810	-.149	.052
Base 3	Negative	2.66	1.249	.46	.807	-.203*	.114
R-B3	Negative	2.84	1.158	.44	.809	-.050	.013
Base 4	Negative	2.21	.974	.52	.806	-.272**	.052
R-B4	Negative	2.75	1.126	.56	.803	-.289**	.086
Base 8	Positive	3.37	1.231	.40	.810	.001	-.093
R-B8	Positive	2.67	1.021	**.22**	.817	**-.048**	**.032**
Base 2	Negative	3.20	1.260	.31	.814	-.084	.012
Base 5	Negative	2.18	1.028	.33	.813	-.020	.156
Base 6	Negative	2.90	1.133	.46	.808	-.115	.092
Base 7	Positive	2.84	1.212	.39	.811	-.320**	.286**
New 1	Positive	2.78	1.072	.31	.814	-.249*	.189
New 2	Positive	2.01	.974	.31	.814	-.173	.127
New 3	Positive	1.96	.933	**.22**	.817	**-.147**	**.040**
New 4	Positive	2.84	1.107	.32	.813	-.034	.121
New 5	Positive	3.40	1.005	**.28**	.815	-.288**	.203*
New 6	Positive	3.56	.887	.32	.814	-.191	.213*
New 7	Positive	2.66	1.114	**.19**	.819	-.347**	.208*
New 8	Positive	3.96	1.219	**.29**	.815	.158	-.131
New 9	Negative	3.87	.998	**.24**	.816	-.079	.040
New 10	Negative	3.32	1.122	.30[d]	.814	-.260*	.186
New 11	Negative	3.33	1.129	**.24**	.817	-.143	.301**
New 12	Negative	2.68	1.116	.39	.811	**-.160**	**.082**
New 13	Negative	2.70	1.232	.36	.812	-.212*	.023
New 14	Negative	3.56	1.179	.34	.813	-.215*	.136

Note:

***Correlation is significant at the 0.01 level (2-tailed).*

**Correlation is significant at the 0.05 level (2-tailed). R-B1 to R-B4 refers to revised based items (or test items).*

—' = Item is included. ' = Item is excluded. Statistics that failed to satisfy the established criteria are in bold face.

[a]For the face validation ratings, some of the items have a skewed distribution. Item median was computed for these items. Us the same.

[b]Mean student rating from one school is below 3.0. These items were excluded in the final form.

[c]Cronbach's alpha if item deleted is compared with the reliability coefficient of the 26-item scale which is .818.

[d]Corrected item total correlation is below 0.3 when the four base items were excluded in the analysis. This item was not inclu eR-B3 and base item 3 are couplets. One should be removed. Results show that base item 3 outperformed R-B3 in terms of ite

Alpha w/o the item	Item Correlation with Self-criterion Residuals (*n* = 96)					Face Validation Ratings[a]		Inclusion
	("r) N	(+r) E	(+r) O	+r) A	(+r) C	Expert (*n* = 4)	Student (*n* = 28)	
.810	229*	.079	**-.221***	.256*	.102	4.75	3.29	
.810	-.149	.052	.022	.339**	.150	4.75	3.46	—
.807	-.203*	.114	-.069	.358**	.113	4.25	3.54	—
.809	-.050	.013	-.005	.316**	.101	4.25	3.18	[e]
.806	-.272**	.052	.091	.195	.132	4.5	3.11[b]	
.803	-.289**	.086	-.012	.242*	.126	4.75	3.57	—
.810	.001	-.093	-.124	.233*	.167	4	3.68	—
.817	**-.048**	**.032**	**-.032**	**.097**	**-.079**	3.75	3.32	
.814	-.084	.012	-.082	.363**	.249*	5	3.32	—
.813	-.020	.156	-.105	.248*	.190	4.5	**2.82**	
.808	-.115	.092	-.068	.203*	.366**	4.25	3.39	—
.811	-.320**	.286**	.015	.290**	.024	4.25	3.21	—
.814	-.249*	.189	.109	.197	.181	3.5	**2.79**	
.814	-.173	.127	.041	.085	.243*	4	3.54	—
.817	**-.147**	**.040**	**-.015**	**-.087**	**.174**	**1.75**	3.21	
.813	-.034	.121	.128	-.037	.230*	4.5	3.54	—
.815	-.288**	.203*	.093	.044	.381**	5	3.50	
.814	-.191	.213*	.152	.048	.467**	4.5	3.61	—
.819	-.347**	.208*	-.026	-.072	.064	3.5	3.61	
.815	.158	-.131	-.052	.310**	.079	4.75	4.07	
.816	-.079	.040	-.137	.090	.354**	5	3.54	
.814	-.260*	.186	.015	.251*	.018	4.5	**2.86**	
.817	-.143	.301**	.101	.446**	.079	4.25	**2.96**	
.811	**-.160**	**.082**	**-.134**	**.151**	**.187**	5	3.18[b]	
.812	-.212*	.023	-.072	.339**	.088	4	3.39	—
.813	-.215*	.136	-.112	.254*	.079	4.75	3.46	—

revised based items (or test items).
established criteria are in bold face.
. Item median was computed for these items. Using the cutoff of 3.0, the decision of accepting or rejecting the item remained

d in the final form.
of the 26-item scale which is .818.
excluded in the analysis. This item was not included in the final form.
at base item 3 outperformed R-B3 in terms of item characteristics.

find it as such. For most of the items, student raters from two schools had fairly similar ratings in relation to the cutoff mean (greater than 3.0). Differences were only observed for base item 4 and new item 12, whereby student raters from School 1 did not find it acceptable but students from School 2 considered them adequate. These two items were also flagged to not satisfy the criterion.

Item Statistics

Another criterion was for items to have item discrimination of +0.30 and above. Results showed that seven out of the twenty-six items did not satisfy this criterion. Six of them were new items, and one was a revised base item. With the design of this study, it was possible to generate inflated corrected item–total correlations because of revised base items. These test items may be considered totally similar with their corresponding base items since only minor revisions were done. With very similar items present, most likely, one would generate higher correlations than the usual. To check for this possible bias, additional analyses were conducted. These were computation of item discrimination value without the four revised base items and another analysis without the four base items. Items should still have corrected item–total correlations of +0.30 and above from the analyses conducted. Obtained results were comparable with those of the first analysis except for one. New item 10 had an item discrimination value equal to .27 when all the four base items were not included in the analysis. This item was not included in the final form.

The Cronbach's-alpha-if-item-deleted was computed using the 26 items. Checking this statistic showed that the identified bad items based on item discrimination also had minor contribution to the test reliability. Particularly, the removal of new item 7 would only increase the reliability coefficient by 0.001. For the rest of the bad items, a decrease of about 0.001 to 0.004 is observed if these were removed.

Base vs Test Items

Each item should provide different contexts. This criterion necessitates comparing four base items with their corresponding test items and identifying the better one. Choosing revised base item 1 (R-B1), revised base item 3 (R-B4), and base item 8 was easy, because their counterpart items did not meet one of the criteria established. For the remaining pair, a detailed scrutiny on several psychometric properties and content differences was done. Results

showed that base item 3 had more decent statistics over revised base item 2 (R-B3).

Summary of the Item Selection Procedure

With the selection procedures enumerated, fourteen items were deemed inadequate. The twelve items that compose the final scale satisfied all the criteria established. Keying for the selected items was fairly distributed to do away with possible acquiescent responding; five items were positively-keyed, and the rest (n = 7) were in the opposite keying. The 12-item social desirability scale is composed of five base items, two revised base items, and five new items. The reliability of the 12-item SD scale is .706.

Phase 3: Cross Validation

Social desirability items that met specified criteria composed the final SD scale. In this phase, the psychometric property of the scale was examined; particularly, if the reliability coefficient would still be acceptable given a different set of samples. Also, this phase aimed to provide evidence of the scale's concurrent validity by correlating it with a bias index, the self-criterion residuals.

METHOD

Participants for Self-Report Data. One hundred sixty-two students responded in this phase. One hundred six (65.43%) students were from a state university in Central Luzon, while 56 (34.57%) students were from another state university in Southern Luzon. There were about 108 (66.7%) females, 47 (29%) males, and about 4.3% (n = 7) not indicating their gender. Their mean age was 18.48 years (SD = 1.475), ranging from 15 to 27 years. Most of the participants were third- (69.1%) and second-year (29.3%) college students. The fields of study they were in are mostly Education (33.55%), Forestry (29%), and Architecture (16.7%).

Participants for Paired Data. Most of them were from a state university in Central Luzon (72.45%). The self-raters were mostly females (70.4%). Their age spanned from 15 to 27 years, with a mean age of 18.51 years old (SD = 1.603). The majority of them were in their third (71.4%) and second year (19.4%) in college. Education (31.6%), Forestry (21.4%), and Architecture (15.3%) were the usual degree programs taken. Nearly everyone reported that the partner for this study is his or her friend (82.7%). Their

length of acquaintanceship between the self and peer rater ranged from .25 to 18 years, with a mean of 3.08 (*SD* = 2.942). The mean rating of the self-rater's familiarity with the peer is 7.72 (*SD* = 1.627).

Majority of the peers who rated their partners were females (70.4%), and their mean age was 18.22 years old (*SD* = 1.126). A bulk of them was in third (68.4 %) and second year (14.3%) of college, and they were enrolled in Education (25.5%), Forestry (20.4%), and Architecture (17.3%). A good number indicated that their partner was his or her friend (81.6%), and the years they had known each other ranged from .25 to 11 (*M* = 2.8; *SD* = 1.925). The mean rating in terms of familiarity was 7.76 (*SD* = 1.304). Responses from the dyad corresponded as significant correlation was observed for length of acquaintanceship ($r = .642$, $p < .01$) and degree of familiarity ($r = .310$, $p < .01$).

RESULTS

Reliability

Using the cross validation sample, the reliability coefficient of the 12-item SD scale went up to .731 from its previous value of .706. Item statistics were fairly comparable with the results in the item selection results. Generally, it can be said that the social desirability scale is relatively stable.

Concurrent Validity

Results showed that social desirability scale was significantly and positively correlated with Agreeableness ($r = .639$, $p < .01$) and Conscientiousness ($r = .464$, $p < .01$). Significant and negative relationship was observed between social desirability and Neuroticism ($r = ".592$, $p < .01$). Results were not statistically significant for domains of Openness to Experience ($r = ".029$, $p = .774$) and Extraversion ($r = .110$, $p = 280$). Correlation between the local social desirability scale and bias index on domains of A, C, and N provides evidence that the developed test can measure self-bias.

Table 4. Descriptive Statistics and Intercorrelations

	Items	Mean	*SD*	CITC	MEAN	1	2	3	4
1. Filipino SDS	12	36.56	6.427	0.37	(.730)				
2. MCSD	33	17.7	5.265	0.28	.644**	(0.788)			
3. BIDR (IM)	20	58.29	10.18	0.33	.692**	.662**	(0.762)		
4. BIDR (SDE)	20	64.15	8.779	0.26	.527**	.540**	.423**	(0.692)	

DISCUSSION

The primary objective of this research was to develop a social desirability scale that is reliable, valid, and appropriate for Filipinos. After going through all the criteria established to screen the pool of items, 12 items were identified to have good properties both on statistical and conceptual grounds. The newly developed 12-item social desirability scale was found to have adequate reliability, and this was replicated using a different set of samples. It can be inferred that the scale is reliable, but for now, this assertion can only be extended to college-student samples. Further research is warranted to see if it will still function the same way with samples like adults and job applicants.

The test also provided validity evidence on what it measures, by displaying a significant and appropriate relationship with self-criterion residuals on personality domains of Agreeableness, Conscientiousness, and Neuroticism. Given the correlational trends obtained, it can be inferred that high SD scorers would most likely have higher self-report ratings on the desirable traits compared to the evaluation of the peers. This was observed on Agreeableness and Conscientiousness. For undesirable traits, the opposite trend is expected; that is, high scorers on the social desirability test would tend to disown possession of such characteristic. This was observed in the current findings wherein high SD scorers had lower self-report ratings on Neuroticism compared to the appraisal of peers. The direction of the obtained correlation results reflects the simple working definition of social desirability in this study; that it is a tendency to present oneself in an exaggeratedly positive way. At the same time, it is assumed that individuals scoring low on the social desirability test would not present themselves in an overly positive manner. This means that their personality evaluation would be similar as with their peers.

Limitations

The self-criterion residual is the main validity criterion in Study 1. Positive values on self-criterion residuals directly correspond to the working definition of social desirability; that is, an exaggeratedly positive presentation in self-descriptions. On the other hand, self-criterion residuals with zero and negative values are viewed to present the other end of the working definition that self-raters are not presenting in a positive manner. Self-criterion residuals with zero values indicate that the self and peer evaluations are the same.

Negative values of self-criterion residuals denote that self-report ratings on desirable traits are lower than their peer-ratings and/or that self-ratings on undesirable traits are higher compared with the evaluation of peers. This study assumes that self-criterion residual with negative values is an indicator of not being socially desirable or not presenting oneself in an overly positive way. However, the possibility that this indicator could tap other things cannot be dismissed, and this necessitates further study on it. Validation studies using different criteria are also suggested.

STUDY 2: Convergent and Concurrent Validation

The second study aimed to provide additional evidence of the scale's validity by getting its correlation with well-known measures of social desirability — the Balanced Inventory of Desirability Responding of Paulhus and Marlowe-Crowne Social Desirability Scale. A positive and significant correlation with foreign-made scales would suggest that the Filipino Social Desirability Scale can indeed measure what the other scales can measure.

METHOD

Materials

Balanced Inventory of Desirable Responding (Paulhus, 1991). The Balanced Inventory of Desirable Responding is popular for its two domains that directly correspond to the Alpha and Gamma factors — the consistent factors that appeared in the early studies of social desirability (Wiggins, 1964; Block, 1965, both cited in Paulhus & John, 1998). The first domain of the BIDR is the self-deceptive enhancement which pertains to positively biased but honest self-reports. Impression management is the second domain and is defined as the conscious presentation of the self to an audience. The two are different in their psychological process; the former is said to be unconsciously done, while the latter is a conscious action.

Self-deceptive enhancement and impression management scales have twenty items each, and these were rated using a 5-point rating scale; this ranged from *Not True* (1) to *Very True* (5). The typical reliability coefficients of self-deceptive enhancement are from .67 to .77, and for impression management, these values ranged from .77 to .85 (D.L. Paulhus, personal communication, August 22, 2008).

Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960, 1964). The MCSD Scale intends to measure a person's tendency of handling evaluative situations, and the authors forwarded that the main motivation behind the said construct is the "need for approval". Scores on MCSD were also found to be significantly correlated to being conforming, cautious, and persuasible. The scale has 33 items. Each item has a dichotomous response option, true or false. The reliability coefficient (KR-20) of the test is .88.

Procedure

In a psychological-measurement class, students were asked to answer or recruit people to answer the online inventory in exchange for extra credit (or points). This was voluntary activity, and students may opt not to participate. The inventory contains the Filipino Social Desirability Scale, Balanced Inventory of Desirable Responding, and Marlowe-Crowne Social Desirability Scale. Most of the participants are females (n = 160, 67.2%), and their age ranged from 18 to 61 years old with a median of 20.

RESULTS

The social desirability scales have decent reliability coefficients. These ranged from .692 (BIDR-SDE) to .788 (MCSD). The SD measures are significantly and positively correlated with each other, with mean value of 0.58 (SD = 0.102). The minimum and maximum correlation values were 0.423 and 0.692, respectively. It can be inferred that the Filipino SD Scale can measure what the existing SD scales can measure. Table 4 shows the descriptive statistics and intercorrelations of the SD measures.

DISCUSSION

The reliability coefficient of the Filipino Social Desirability Scale is consistent with that of Study 1, which is around 0.7. Moreover, its performance is at par with existing foreign SD scales. Although MCSD and BIDR Impression Management appear to be higher in terms of reliability, the number of items is a factor to be considered. There is a direct relationship between the number of items and reliability; more items would yield to a higher reliability coefficient. The performance of the local scale is good, given that the Filipino Social Desirability Scale is just composed of 12 items

compared with MCSD and BIDR domains with 33 and 20 items, respectively. Moreover, the correlation results with the foreign SD scale add evidence to the validity of the Filipino Social Desirability Scale. That is, it can indeed measure social desirability, as with other scales.

Social Desirability as a Construct

This study developed a social desirability scale. Evidence in this research supports the idea that the construct exists; social desirability items go together, and as a scale, it is significantly related with discrepancies on personality evaluation between self and other raters. The line of inquiry that follows is how this construct behaves in the Philippine context. Looking at the trends of the correlation results, social desirability was significantly related with self-criterion residuals of Neuroticism, Agreeableness, and Conscientiousness but not for Openness to experience and Extraversion.

Following the theorizing of Paulhus and John (1998), the values of the individuals predispose them to certain bias. In particular, those who value agency will tend to have self-favoring bias on Extraversion and Openness to Experience, and preferential bias on Agreeableness and Conscientiousness is expected for those who value communion. The Philippines is considered a collectivist society (Hofstede, 2001). If Filipinos value communion more than agency, then it follows that Filipinos are inclined to have self-favoring bias on Agreeableness and Conscientiousness. The results provided concrete evidence to support the bias categorization of Paulhus and John (1998). Significant correlations were observed for self-criterion residuals of Agreeableness and Conscientiousness but not for Openness to Experience and Extraversion. On an interesting note, the direction of bias towards Neuroticism is not explicitly mentioned in their paper. It can be inferred though that the significant result obtained for Neuroticism was rational and can still be supported by the connections they made.

Viewing the results from a two-factor model (Digman, 1997) would also capture the observed trends in this study. Personal Growth and Socialization are Digman's two higher-order factors, and Paulhus and John (1998) classified these under agency and communion values, respectively. Socialization factor covers Agreeableness, Conscientiousness, and Neuroticism, and bias on this factor seems to be connected to social desirability for Filipinos. On the other hand, no relationship was found between bias on Personal Growth (where Extraversion and Openness are

under) and social desirability. The tendency to show self-favoring bias on Neuroticism is probably a trend that is unique to Filipinos, or a different lens should be used in interpreting the results. It is clear that research is needed to see how social desirability behaves among Filipinos.

The Filipino Social Desirability Scale

Any self-rating data would always be vulnerable to questions of one's social desirability. Measurement of social desirability adds integrity to any assessment that involves "self" as the rater. Whether we perceive SD as error or substantive construct, having an accessible scale and information about one's social desirability are deemed significant especially if a consensus has not been reached yet. In a pragmatic sense, it is better to have all possible data available, and it is up to the judgment of the data user on how to make sense or utilize all the information. If the data user views social desirability as an error, he or she is more inclined to correct self-report scores based on one's SD. If social desirability is viewed as a substantive variable, the data user would probably avoid doing any adjustment, because corrections will actually lower validity of a construct measure.

The use of the Filipino Social Desirability Scale is recommended for research that involves the use of self-report data. Studies that would assist the users in identifying criteria for extreme response biases and the cutoff scores for rejecting self-report data are also suggested. The production of this test is just the starting point. Research that would help us to understand self-report bias in the Philippine context is envisioned in the future.

REFERENCES

Backstrom, M. (2007). Higher-order factors in a five-factor personality inventory and its relation to social desirability. *European Journal of Psychological Assessment, 23*, 63–70.

Backstrom, M., Bjorklund, F., & Larsson, M. R. (2009). Five-factor inventories have a major general factor related to social desirability which can be reduced by framing items neutrally. *Journal of Research in Personality, 43*, 335–344.

Cagasan, L. (August, 2012). The development of the social desirability scale of the mapa ng loob. Paper presented at the 49th Annual Convention of Psychological Association of the Philippines, Cebu City, Philippines.

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

Crowne, D. P. & Marlowe, D. (1964). *The approval motive*. New York: Wiley.

Del Pilar, G., Sio, C., Cagasan, L., Siy, A. & Galang, A. J. (2015). Masaklaw na Panukat ng Loob (Mapa ng Loob). Quezon City, Philippines: University of the Philippines, OVCRD.

Digman, J. (1997). Higher-order factors of the big five. *Journal of Personality and Social Psychology, 73*, 1246–1256.

Edwards, A. L. (1957). *The social desirable variable in personality assessment and research.* New York: Holt, Rinehart & Winston.

Felipe, A. (1969). Social desirability tendency and endorsement of items in a forced-choice inventory. *Philippine Journal of Psychology, 2,* 12–18.

Hofstede, G. J. (2001). Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations (2nd ed.). Thousand Oaks, CA: SAGE.

Holden, R., & Passey, J. (2010). Socially desirable responding in personality assessment: Not necessarily faking and not necessarily substance. *Personality and Individual Differences, 49,* 446–450.

Li, A., & Reb, J. (2009). A cross-nations, cross-cultures, and cross-conditions analysis on the equivalence of the balanced inventory of desirable responding. *Journal of Cross-Cultural Psychology, 40*, 214–233.

McCrae, R., & Costa, P. (1983). Social Desirability Scales: More Substance Than Style. *Journal of Consulting and Clinical Psychology, 51,* 882–888.

Murphy, K. R. & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Ones, D., Viswesvaran, C., & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology, 81,* 660–679.

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, 46, 598–609.

Paulhus, D. L. (1991). Measurement and control of response bias. *In* J.P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.*), Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. *In* H.I. Braun & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49–69). Mahwah, NJ: Erlbaum

Paulhus, D. L., & John, O. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66,* 1024–1060.

Pauls, C., & Stemmler, G. (2003). Substance and bias in social desirability responding. *Personality and Individual Differences, 35,* 263–275.

Piedmont, R.L., McCrae, R.R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.

Robins, R. & John, O. (1997). The quest for self-insight: Theory and research on accuracy and bias in self-perception. *In* Hogan, R. (Ed), Johnson, J. (Ed), Briggs, S. (Ed), Handbook of Personality Psychology (pp. 649–679). San Diego, CA: Academic Press.