Wording effects and the dimensionality of the General Health Questionnaire (GHQ-12)

Lei Wang [*,1], Weipeng Lin

Department of Psychology, Peking University, Beijing, China

ARTICLE INFO

ABSTRACT

There has been a debate about the structure of the 12-item General Health Questionnaire (GHQ-12). While some researchers believe it taps two or three dimensions, others think its multidimensionality is merely an artifact of wording effects. To directly test these competing views, we had 855 Chinese employees take one of three versions of the GHQ-12: either the original version, which is comprised of six positively worded and six negatively worded items; or one of two alternative versions that are comprised of 12 positively worded or 12 negatively worded items. Analyses indicated that the GHQ-12 had a unidimensional factor structure after controlling wording effects. The substantive meaning of the factor analysis results is further discussed based on their correlations with a string of personality and organization behavior variables.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The General Health Questionnaire (GHQ) is a simple scale developed by Goldberg (1972), which is a widely used instrument for measuring general psychological health (e.g., Creed & Evans, 2002; Emery, Huppert, & Schein, 1996). There are multiple versions of the GHQ with different numbers of items, such as 12, 28, 30, and 60 items, all of them have been shown to have good reliability and validity in various samples (e.g., Cook, Young, Taylor, & Bedford, 1996; Ivkovic et al., 2007; Ploubidis et al., 2007). Among all these versions, the GHQ-12 (GHQ with 12 items) is the shortest and most commonly used version. Despite the brevity of the GHQ-12, previous studies have suggested that it is as accurate as longer versions of the GHQ for screening and case detection (Goldberg et al., 1997).

Although the GHQ-12 is intended to be used as a tool for a general evaluation of individual mental health status, the factor structure of GHQ-12 has been the target of debate since several researchers have suggested that the GHQ-12 reflects a multidimensional construct (Andrich & Van Schoubroeck, 1989; Graetz, 1991; Kalliath, ODriscoll, & Brough, 2004; Martin, 1999; Shevlin & Adamson, 2005; Werneke, Goldberg, Yalcin, & Uestuen, 2000; Worsley & Gribbin, 1977). For example, Andrich and Van Schoubroeck (1989) found that a two-factor model, in which both positively and negatively worded items respectively defined one factor, showed adequate fitness to the data. In addition, Graetz

(1991) advocated that a three-factor model fit the data better than the others. In contrast, Ye (2009) argued that the GHQ-12 is simply a single factor scale. Therefore, whether the multifactor models truly revealed the underlying construct or were just an artifact of item-wording is still open to debate.

The effects of how items are worded belong to one kind of method effects that may confound the factor structure of a rating scale. The manifestation of these effects is that, when psychological rating scales consist of positively and negatively worded items, factor analyses of responses to those items often reveal apparently distinct factors that reflect the positive and the negative items, respectively (Marsh, 1996). In recent decades, the effects of item-wording have been widely researched (Carmines & Zeller, 1994; Marsh, 1996; Ye, 2009), but the possibility that the presumed multifactor structure of GHQ-12 is an artifact of item-wording has got little attention from researchers. The exception was Ye's (2009) recent research, which examined whether item-wording effects play a role in the multidimensionality of the GHQ-12. The study used a confirmatory analysis to compare three alternative factor models, and found that after controlling the effects of item-wording, the GHQ-12 has a unidimensional construct. However, Ye's research has the limitation that it was not able to directly confirm the unidimensional construct of the measure because it only tested the wording effects within one version of the questionnaire, and did not compare whether the wordings from different versions would affect the dimensionality. Furthermore, the analysis found that the two- and three-factor models proposed by Andrich and Van Schoubroeck (1989) and Graetz (1991), respectively, also fit the data well. Namely, Ye's (2009) study can only indirectly test whether the GHQ-12 is a unidimensional measure.

* Corresponding author.
  E-mail address: leiwang@pku.edu.cn (L. Wang).

Table 1
The GHQ-12 and two re-worded versions.

The original GHQ-12	Revised-positive version	Revised-negative version
1. Lost sleep over worry[a]	1. Have a sound sleep	1. ~
2. Constantly under strain[a]	2. Feel no strain	2. ~
3. Able to concentrate	3. ~	3. Unable to concentrate[a]
4. Play useful part in things	4. ~	4. Play useless part in things[a]
5. Face up to problems	5. ~	5. Could not face up to problems[a]
6. Capable of making decisions	6. ~	6. Incapable of making decisions[a]
7. Could not overcome difficulties[a]	7. Able to overcome difficulties	7. ~
8. Reasonably happy	8. ~	8. Reasonably unhappy[a]
9. Enjoy day-to-day activities	9. ~	9. Could not enjoy day-to-day activities[a]
10. Unhappy and depressed[a]	10. Happy and inspired	10. ~
11. Loss of confidence in self[a]	11. Full of confidence in self	11. ~
12.Thinking of self as worthless[a]	12.Thinking of self as valuable	12. ~

Note: ~, Indicates that items were the same as that in the original version.
[a] Indicates that items were reverse-scored when calculating the total mental health score.

The aim of this study was to examine whether item-wording accounts for the multidimensionality of the GHQ-12 in the only direct way by controlling the wording effects across several versions using the method introduced by Greenberger, Chen, Dmitrieva, and Farruggia (2003). Specifically, based on the original 12-item GHQ (Goldberg & Williams, 1988), we created an all negatively worded version of the scale (Revised-negative version), in which previous positive items were re-written in the negative form, and an all positively worded version (Revised-positive version), in which previous negative items were re-written in the positive direction. The revised versions, therefore, are "uni-directional" in terms of their assessment of general mental health. If the multifactor model only appears in the Original GHQ-12, the artificial nature of the multidimensions would be supported. That is, item-wording would have some impacts on the original scale which confound the factor structure. However, if multifactor models fit equally well across all scale-versions, the contention that the GHQ-12 is multidimensional would be supported. Such a result would suggest that the two or three subsets of items, independent of the direction in which they are worded, tap different dimensions of mental health.

In addition, in order to compare the validity of the factors in assessing psychological well-being, we tested the relationships of the factors in the four models (i.e. three-factor model, two-factor model, one-factor model, and wording effects model) with several relevant measures including neuroticism, extraversion, and turnover intention. These measures were chosen for two reasons. First, neuroticism and extraversion have been found to be the two personality traits that are very robust (Zuckerman, Kuhlman, Joireman, Teta, & Kraft, 1993) and closely related to well-being (Diener, Oishi, & Lucas, 2003); turnover intention is a variable of great importance for enterprises, and is a commonly used predictive variable of well-being in organizational research (Spector et al., 2007). Second, these measures are all brief enough, making it feasible to conduct surveys in enterprises.

2. Methods

2.1. Participants and procedures

Data were collected among 1005 Chinese employees from a manufacturing company in southern China. The participants were recruited through their HR department and were randomly assigned to complete one of the three versions of the GHQ-12. All other measures were identical across the three conditions. Nevertheless, due to the length limit of the questionnaires, criterion-related measures of neuroticism and extraversion were not collected in the revised-negative condition. Participants completed the measure in meeting rooms after their working time. We finished our data collection with the assistance of the HR department.

Among the participants, 855 finished the survey (85.07% response rate) with 288 completed original GHQ-12 (12.9% were male, 79.7% were female and 7.3% did not report), 257 completed revised-negative GHQ-12 (18.7% were male, 75.5% were female and 5.8% did not report), and 310 completed revised-positive GHQ-12 (17.9% were male, 76.7% were female and 5.4% did not report). The questionnaires were completed anonymously and voluntarily in groups. Participants who completed the questionnaires received a small gift valued at 10CNY. Among these participants, the mean age was 23.0 (SD = 3.55); most of them were line-workers (90.3%) with high-school or lower education (98.0%).

2.2. Measures

2.2.1. GHQ-12

Three versions of the GHQ-12 were administered. The original GHQ-12 contains six positively worded items and six negatively worded items. The revised-negative version is an adaptation of GHQ-12 in which the six positively worded items in the original scale have been rephrased in a negative direction, resulting in a 12-item scale with all negatively worded items. The revised-positive version is an adaptation of GHQ-12 in which the six negatively worded items in the original scale have been rephrased in a positive direction, resulting in a 12-item scale with all positively worded statements (see Table 1). Adapting previously positively worded items to a negative version, and vice versa, was usually a matter of inserting or deleting the word "not" or changing a negative word to a positive one (e.g., "able" from Original GHQ-12, to "unable" for the Revised-negative version). However, in some instances modest additional changes were necessary. In the case of item 2, for example, the original GHQ-12 phrasing "Constantly under strain" (a negative item) was revised to read "Feel no strain" for the Revised-positive version. On all versions, respondents were asked to rate their agreement on a 4-point scale (1 = never, 4 = much more than usual). We reversed the score so that the higher score represents the better psychological health. Cronbach's alpha coefficients for the Original version, the Revised-negative version, and the Revised-positive version were .70, .87, and .85, respectively.

2.2.2. Turnover intention

Turnover intention was measured with a 4-item scale developed by Farh, Tsui, Xin, and Cheng (1998) and each item was measured on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). An example item is "I often would like to resign from my current work". The scale's alpha reliability in this study was .75.

2.2.3. Neuroticism and extraversion

Neuroticism and extraversion subscales from the brief version of the NEO Five-Factor Inventory (Saucier, 1994) were used in

Table 2
Model specifications for the alternative models of the GHQ-12.

Item number	Model 1 Single-factor model	Model 2 Two-factor model Andrich and Van Schoubroeck (1989)		Model 3 Three-factor model Graetz (1991)			Model 4 Wording effects model Ye (2009)	
	F1	F1	F2	F1	F2	F3	F1	F2
1	*		*		*		*	*
2	*		*		*		*	*
3	*	*		*			*	
4	*	*		*			*	
5	*	*		*			*	
6	*	*		*			*	
7	*		*		*		*	*
8	*	*		*			*	
9	*	*		*			*	
10	*		*		*		*	*
11	*		*			*	*	*
12	*		*			*	*	*

Note: F = factor. (−) Indicates items that were reverse-scored when calculating the total mental health score. * Indicates that items were free estimated in the factor analysis.
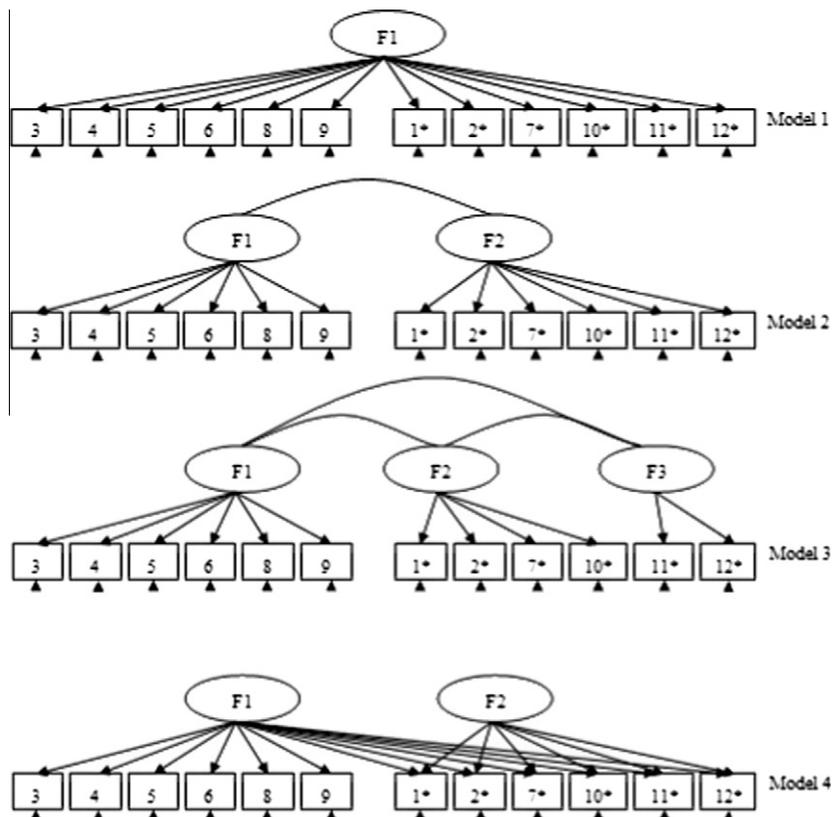
Fig. 1. Description of four tested models. *Note*: * Indicates that items were negatively-worded in the original version of GHQ-12. Model 1–4 are the same as those in Table 3.

the study. Both subscales were comprised of 8 items. Participants were asked to rate the extent to which each item is appropriate to describe themselves on a 5-point Likert scale (1 = strongly inappropriate, 5 = strongly appropriate). An example item for neuroticism is "Envious"; and that for extraversion is "Extroverted". Cronbach's alpha coefficients for neuroticism and extraversion subscales were .70 and .65, respectively.

Control variables: Participants' demographic characteristics, including gender, age, education, and tenure, comprised our control variables.

2.3. Plan of analysis

Four types of models (see Table 2) were evaluated for the three versions of the GHQ-12. The specification for each

model was taken directly from the results of the factor analysis reported by each author whose model is reported in the table. The detailed descriptions of all four tested models are presented in Fig. 1.

First, correlations between factors of all three versions were calculated to test whether wording effects had influenced factor correlations. Next, confirmatory factor analysis was conducted to examine the dimensionality of GHQ in all three versions, using LISREL 8.70 (Jöreskog & Sörbom, 2004). Maximum likelihood method of estimation was used. After that, mean comparisons were conducted to evaluate whether scale revision had resulted in different mean levels of self-reported mental health. Finally, we tested the criterion-related validity of the three versions of GHQ-12 and made comparisons based on the correlations between the factors of GHQ-12 and other relevant measures.

Table 3
Factor correlations for the alternative models of the GHQ-12.

		M1	M2		M3			M4	
		F1	F1	F2	F1	F2	F3	F1	F2
The original version	F1	–	–		–			–	
	F2		.06	–	.02	–		.73[**]	–
	F3				.14[**]	.41[**]	–		
Revised-positive version	F1	–	–		–			–	
	F2		.79[**]	–	.72[**]	–		.95[**]	–
	F3				.68[**]	.60[**]	–		
Revised-negative version	F1	–	–		–			–	
	F2		.83[**]	–	.77[**]	–		.96[**]	–
	F3				.63[**]	.43[**]	–		

Note: F = factor. M1: single-factor model; M2: two-factor model; M3: three-factor model; M4: wording effects model. Details about model specifications can be found in Table 2 and Fig. 1.
[**] $p < .01$.

Table 4
Summary of the fit statistics of four alternative models.

Index	Model	The original version				Revised-positive version				Revised-negative version			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
X^2		778.07	254.48	162.48	236.47	192.55	186.78	171.78	152.98	340.41	330.10	259.62	274.15
df		54	53	51	47	54	53	51	47	54	53	51	47
X^2 /df		14.41	4.80	3.19	5.03	3.57	3.52	3.37	3.25	6.30	6.23	5.09	5.83
CFI		0.51	0.83	0.91	0.84	0.94	0.94	0.95	0.96	0.89	0.89	0.91	0.91
GFI		0.69	0.87	0.91	0.88	0.91	0.91	0.92	0.92	0.82	0.82	0.86	0.85
TLI		0.40	0.79	0.88	0.78	0.93	0.93	0.93	0.94	0.87	0.87	0.89	0.88
IFI		0.51	0.84	0.91	0.85	0.94	0.94	0.95	0.96	0.89	0.89	0.91	0.91
RMSEA		0.22	0.12	0.09	0.12	0.09	0.09	0.09	0.09	0.14	0.14	0.13	0.14

Note: CFI = comparative fit index; GFI = goodness of fit index; TLI = Tucker–Lewis index; IFI = incremental fit index; RMSEA = root mean-square error of approximation. M1–M4 are the same as those in Table 3.

3. Results

3.1. Correlations between factors

The correlations between the factors are presented in Table 3. The results showed that for the original version, the correlation between factor 1 and factor 2 was not significant either in the two-factor model or in the three-factor model. However, for both revised versions, the correlations between factor 1 and factor 2 in the two-factor model and in the three-factor model were very high and significant. These results indicated that after controlling the wording effects, the factors in multi-dimensional models were highly correlated, suggesting that GHQ had a unidimensional structure.

3.2. Confirmatory factor analysis

Using LISREL 8.70 (Jöreskog & Sörbom, 2004), the four models were examined within all three versions of the GHQ-12. Table 4 reports the fit indices for the four models.

Results showed that the original version of the GHQ-12 had a significantly better fit for the three-factor model compared to the two-factor and one-factor model, with $\Delta X^2(2) = 92.00$, $p < 0.01$, and $\Delta X^2(3) = 615.49$, $p < 0.01$, respectively. Compared to the wording effects model, the three-factor model also fit better because of its lower X^2 score (162.48 and 236.47 for the three-factor model and wording effects model, respectively) but higher degree of freedom (51 and 47 for the three-factor model and wording effects model, respectively). The factor loadings for the three-factor model are presented in Table 5.

For the Revised-positive version, the three-factor model also fit significantly better than the two-factor and one-factor model, with

Table 5
Standardized factor loadings for the three-factor models of the original GHQ-12.

Item number	Model 3 Three-factor model Graetz (1991)		
	F1	F2	F3
1		0.44	
2		0.53	
3	0.42		
4	0.58		
5	0.52		
6	0.46		
7		0.31	
8	0.35		
9	0.39		
10		0.59	
11			0.88
12			0.48

Note: All factor loadings are statistically significant ($p < .01$).

$\Delta X^2(2) = 15.00$, $p < 0.01$ and ΔX^2 (3) = 20.77, $p < 0.01$, respectively. However, the wording effects model had a better fit than the three factor model not only on the ΔX^2 test [$\Delta X^2(4) = 18.80$, $p < 0.01$] but also on all other fit indices.

For the Revised-negative version, the three-factor model fit significantly better than the two-factor and one-factor model as well, with $\Delta X^2(2) = 70.48$, $p < 0.01$, and $\Delta X^2(3) = 80.79$, $p < 0.01$, respectively. Compared to the wording effects model, the three-factor model also fit better because of its lower X^2 score (259.62 and 274.15 for the three-factor model and wording effects model, respectively) but higher degree of freedom (51 and 47 for the three-factor model and wording effects model, respectively). Nevertheless, other fit indices of the three-factor model and wording effects model are very close.

Table 6
Correlations between factors and relevant measures.

GHQ version	Relevant measure	M1	M2		M3			M4	
		F1	F1	F2	F1	F2	F3	F1	F2
O	TI	−.23**	−.14*	−.18**	−.14*	−.14*	−.18**	−.23**	−.18**
	Neuroticism	−.31**	−.08	−.36**	−.08	−.31**	−.26**	−.31**	−.36**
	Extraversion	.31**	.13*	.32**	.13*	.27**	.25**	.13*	.32**
P	TI	−.22*	−.20**	−.18**	−.20**	−.13*	−.20**	−.22**	−.18**
	Neuroticism	−.25**	−.22**	−.24**	−.22**	−.23**	−.19**	−.25**	−.24**
	Extraversion	.13*	.11	.15*	.11	.11	.17**	.13*	.15*
N	TI	−.23**	−.21**	−.22**	−.21**	−.25**	−.07	−.23**	−.22**

Note: O = original version; P = revised-positive version; N = revised-negative version; F = factor; TI = turnover intention. M1–M4 are the same as those in Table 3.
* $p < .05$.
** $p < .01$.

Comparing the difference of model fit among three GHQ versions, we found that the three-factor model for the two revised versions had worse fit than that for the original GHQ-12. However, the one-factor model for the two revised versions had a better fit than that for original GHQ-12.

To sum up, after controlling the wording effects: (a) the model fit of multi-dimensional models became worse, while the fit indices of the single-factor model became better; (b) the wording effects model, which indicates that GHQ-12 is a uni-dimensional structure, fit better than any other model in the revised-positive version and had a close fit with the three-factor model in the revised-negative version.

3.3. Mean differences for the three versions of the GHQ-12

Mean scores of the global GHQ-12 scale on three version were calculated, and one-way ANOVA was used to test whether they were different from each other. Results showed that mean scores on the three versions of the GHQ-12 differed significantly: for the Original GHQ-12, $M = 2.71$ (SD = 0.38); for the Revised-positive version, $M = 1.53$ (SD = 0.50); and for the Revised-negative version, $M = 1.91$ (SD = 0.53), $F(780) = 436.20$, $p < 0.01$. The effect size (η^2) was 0.50. These results gave credence to the idea that item-wording may influence the mean scores of GHQ-12.

3.4. Correlates

Table 6 shows the correlations between the mean GHQ-12 scores and criterion-related measures. As expected, mean scores of global GHQ scale (i.e. mean scores of Model 1) derived from all three versions were correlated in the expected direction with turnover intention and neuroticism (negatively), and extraversion (positively). Pair-wise comparisons of the correlation coefficients were conducted, using Fisher's r-to-z transformations. Results showed that the wording of items did not seem to affect the magnitude of correlations between the mean GHQ-12 scores and other measures with one exception. The correlation between mean GHQ-12 score and extraversion was significantly lower when all GHQ-12 items were worded positively (Revised-positive version < Original GHQ-12, $z = 2.19$, $p < 0.05$).

To further examine whether the item-wording affects the validity of the GHQ-12, we correlated the factors of GHQ-12 in four models with the criterion-related measure (see Table 6). Results showed that the correlations were all in the same direction, but the magnitudes were somewhat different. For example, in the three-factor model, as for the correlations between the factor 1 of GHQ-12 and neuroticism, the correlations were significantly higher for the revised-positive version ($r = -.22$, $z = 2.44$, $p < .05$) when compared to the original version ($r = -.08$).

In addition, we examined whether the different GHQ-12 subscales had different associations with the criterion-related measures. Results showed that, in the three-factor model of the original GHQ-12, the correlation between factor 1 and neuroticism ($r = -.08$, ns) was lower than that between factor 2 and neuroticism ($r = -.31$) and that between factor 3 and neuroticism ($r = -.26$). In the two-factor model of the original GHQ-12, the correlation between factor 1 and extraversion ($r = .13$) was lower than that between factor 2 and extraversion ($r = .32$, $z = 3.52$, $p < .01$). However, after controlling the wording effects, the correlations between subscales of GHQ-12 and criterion measures were of similar magnitude. Although some exceptions existed, such as in the three-factor model of the revised-positive version, correlations between factor 3 and extraversion were significant ($r = .17$), while the correlation between factor 1 and extraversion ($r = .11$, $z = 0.74$, ns) and between factor 2 and extraversion ($r = .11$, $z = 0.74$, ns) were not statistically significant, the differences in the correlations were small in magnitude. Overall, there was little evidence that the GHQ-12 subscales had differential associations with the criterion-relevant measures, especially after excluding the wording effects.

4. Discussion

Although many researchers had investigated the factor structure of GHQ-12, the "real" structure is still in mist. The present study generated two revised versions of GHQ-12, which ruled out the wording effects, and tested a series of alternative factor structures in three versions of GHQ-12 using confirmatory factor analysis. The results showed that the three-factor model fit the GHQ-12 measure better than other models in the original version. However, the wording effects model, which indicates that GHQ-12 is a unidimensional structure, fit equally well in both revised versions (i.e., had the best fit in the revised-positive version, and had fit indices very close to the three-factor model in the revised-negative version). Besides, comparing the original version, fit indices of the one-factor model had great improvement in both revised versions. In addition, after controlling wording effects, correlations between subscales of GHQ-12 were very high, and the subscales had similar associations with the criterion-relevant measures.

All these results suggest that GHQ-12 indeed had a unidimensional structure after controlling for wording effects. Wording effects confounded the results of factor analysis, which had led to so many conclusions that the GHQ-12 was a multi-dimensional measure in previous research. It is not necessary to calculate scores for the different factors since the present research, as well as past research (e.g., Shevlin & Adamson, 2005), has found that the distinction of factors among the GHQ-12 did not help improve the predictive power of the tool. Therefore, it is recommended that the GHQ-12 be considered and used as a unidimensional instrument for evaluating general mental health.

Our research also found that the correlations between factors within three versions of GHQ-12 were also differential. The correlations were very low and not significant in the original version, but were very high and significant in both revised versions. Moreover, the correlations between criterion variables and GHQ-12 factors, namely, the criterion validity of the three versions of the GHQ-12 differed from each other to some extent. These results suggest that wording effects do affect the between-factor correlations and the criterion-related validity of the scales. Accordingly, item-wording effect should be taken into account when developing a new scale or using an existing one.

Here we should mention that the wording effects model was not the best fit for the revised-negative version. The possible explanation is that the revised-negative version of GHQ-12, which contained 12 negatively-worded items, required more cognitive resource for participants to handle. According to Marsh (1996), agreeing with or especially disagreeing with negatively-worded items adds a degree of cognitive complexity to the task of responding to a questionnaire.

Since the wording effects would confound the factor structure of a measure, some methods should be figured out to eliminate this problem. The easiest way is to use only positively worded items in rating scales. However, this method would not be good enough because it would lead to another problem. Our research found that the mean scores for the revised-positive version and revised-negative version were significantly lower than those for the original version, which confirms the arguments of most measurement experts (e.g., Spector, 1992) that items worded in one direction leads to acquiescence bias. In order to eliminate the wording effects without causing response bias, some strategies were recommended: (a) score positively-worded items only but retain negatively-worded items to reduce acquiescence bias (Marsh, 1996); (b) generate a method factor when conducting confirmatory factor analysis so that the wording effect could be partialled out (Ye, 2009).

Our research also has some implications for developing psychological rating scales. Because of the wording effects, when positively and negatively worded items are needed to include on a measure with multiple subscales, the proportion of positively and negatively worded items should be relatively constant for each subscale. Without this balance, it is hard to clarify the extent to which the distinction between different factors is due to differences in the underlying constructs being measured rather than wording effects (Marsh, 1996).

This study had several limitations. First, the present research that drew data from a large consulting project for a company lacks complete criterion-related validity. This limits our understanding of the comparison of the three versions of GHQ-12. Future studies should examine multiple kinds of validity of the three versions of the GHQ-12, for example, their associations with other measures of wellbeing, job stressors and strains. Second, it focused exclusively on a sample of Chinese employees. Thus, whether these findings can be generalized to other populations needs to be further examined.

References

Andrich, D., & Van Schoubroeck, L. (1989). The general health questionnaire: A psychometric analysis using latent trait theory. *Psychological Medicine, 19*, 469–485.

Carmines, E., & Zeller, R. (1994). *Reliability and validity assessment*. Sage Publications Inc.

Cook, M., Young, A., Taylor, D., & Bedford, A. (1996). Personality correlates of psychological distress. *Personality and Individual Differences, 20*, 313–319.

Creed, P., & Evans, B. (2002). Personality, well-being and deprivation theory. *Personality and Individual Differences, 33*, 1045–1054.

Diener, E., Oishi, S., & Lucas, R. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology, 54*, 403–425.

Emery, C., Huppert, F., & Schein, R. (1996). Health and personality predictors of psychological functioning in a 7-year longitudinal study. *Personality and Individual Differences, 20*, 567–573.

Farh, J., Tsui, A., Xin, K., & Cheng, B. (1998). The influence of relational demography and guanxi: The Chinese case. *Organization Science, 9*, 471–488.

Goldberg, D. (1972). *The detection of psychiatric illness by questionnaire*. London: Oxford University Press.

Goldberg, D., Gater, R., Sartorius, N., Ustun, T., Piccinelli, M., Gureje, O., et al. (1997). The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine, 27*, 191–197.

Goldberg, D., & Williams, P. (1988). *A user's guide to the GHQ*. Windsor: NFER-Nelson.

Graetz, B. (1991). Multidimensional properties of the general health questionnaire. *Social Psychiatry and Psychiatric Epidemiology, 26*, 132–138.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. (2003). Item-wording and the dimensionality of the Rosenberg self-esteem scale: Do they matter? *Personality and Individual Differences, 35*, 1241–1254.

Ivkovic, V., Vitart, V., Rudan, I., Janicijevic, B., Smolej-Narancic, N., Skaric-Juric, T., et al. (2007). The Eysenck personality factors: Psychometric structure, reliability, heritability and phenotypic and genetic correlations with psychological distress in an isolated Croatian population. *Personality and Individual Differences, 42*, 123–133.

Jöreskog, K., & Sörbom, D. (2004). *LISREL 8.70*. Chicago: Scientific Software International Inc.

Kalliath, T., ODriscoll, M., & Brough, P. (2004). A confirmatory factor analysis of the general health questionnaire-12. *Stress and Health, 20*, 11–20.

Marsh, H. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810–819.

Martin, A. (1999). Assessing the multidimensionality of the 12-item general health questionnaire. *Psychological Reports, 84*, 927–935.

Ploubidis, G., Abbott, R., Huppert, F., Kuh, D., Wadsworth, M., & Croudace, T. (2007). Improvements in social functioning reported by a birth cohort in mid-adult life: A person-centred analysis of GHQ-28 social dysfunction items using latent class analysis. *Personality and Individual Differences, 42*, 305–316.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63*, 506–516.

Shevlin, M., & Adamson, G. (2005). Alternative factor models and factorial invariance of the GHQ-12: A large sample analysis using confirmatory factor analysis. *Psychological Assessment, 17*, 231.

Spector, P. (1992). *Summated rating scale construction: An introduction*. Sage Publications Inc.

Spector, P., Allen, T., Poelmans, S., Lapierre, L., Cooper, C., Sanchez, J., et al. (2007). Cross-national differences in relationships of work demands, job satisfaction, and turnover intentions with work-family conflict. *Personnel Psychology, 60*, 805–835.

Werneke, U., Goldberg, D., Yalcin, I., & Uestuen, B. (2000). The stability of the factor structure of the general health questionnaire. *Psychological Medicine, 30*, 823–829.

Worsley, A., & Gribbin, C. (1977). A factor analytic study of the twelve item general health questionnaire. *Australasian Psychiatry, 11*, 269–272.

Ye, S. (2009). Factor structure of the general health questionnaire (GHQ-12): The role of wording effects. *Personality and Individual Differences, 46*, 197–201.

Zuckerman, M., Kuhlman, D., Joireman, J., Teta, P., & Kraft, M. (1993). A comparison of three structural models for personality: The Big Three, the Big Five, and the Alternative Five. *Journal of Personality and Social Psychology, 65*, 757–768.