A Study on the Item Development Strategies of a Self-Reported Personality Scale

Sheyu Chen and Zhiyong Li

***Abstract*—The objective of this paper is to discuss the item development of a self-reported personality scale that meets the psychometric standards, and provides some experiences for relevant studies. Results show that a psychometric sound personality scale items must possess the follow features: 1) considers the cultural background wherein the scale is applied; 2) should be develop based on a theoretical framework; 3) target a certain group of people considering their educational attainments, positions, and other background demographics; and 4) should fit appropriate psychometric standards. Furthermore, the items of a personality scale must be selected according to their effect towards the theoretical framework, social desirability factor, connotation and extension of the item, the presupposition of the question, and as well as the questions of privacy. In addition, a personality scale needs to contain a lie scale, so as to avoid the concealment when the participants answer the inventory.**

***Index Terms*—Characteristics of culture, item development, personality inventory, theoretical framework.**

I. Introduction

Currently, there're numerous personality questionnaires that are different with each other in the connotation of personality, dimensions of personality, and construction path of questionnaire. They have different ideas about the personality is cross-cultural consistency or not, and that how to detect lies of participants etc. This has led to the different levels of reliability and validity of personality scales, and we can't evaluate them in a general criterion. When the personality traits need to be taken into account in an academic research, many researchers are usually select one from them based on their own research purposes. Consequently, one personality research cannot compare with another directly, and the results of relationship between personality traits and another psych-traits or behavior traits are usually different from one to another because of different researchers using different personality inventory. Even more, the results of the researches are usually not powerful enough to receive a credible conclusion, so the ecological validity of personality researches is questioned universally.

This situation may be based primarily on the following causes: 1) personality itself, that there actually are not definite relationships between personality traits and other psychological traits; 2) methodology, that we do not use the correct methods to study personality; 3) understanding of the personality, that we have not yet revealed the nature of personality, or at least one of the personality theory framework is not available to accurately understand the nature of personality; 4) Technology, which we have used to access personality is not the accurate measurement of personality.

For personality itself, there is no doubt that it has a correlation or causal relationship with at least some of other psychological traits, and they must be examined by various researches; but this is beyond the scope of this article. For methodology, there're methods of self-report, project method, background analysis, observation, etc., while the method of self-report is more advantageous and more mature than others. Given the complexity of personality, it is acceptable that different scholars had different opinions and views on it, which is also in line with the law of human understanding of the world. As more and more depth of personality research, to form a unified view of personality is just a matter of time. For technology, though most of scholars generally assume that personality is neutral without good or bad, there is indeed a certain social desirability tendency when developing a self-report item, this leads to response bias of subjects. Moreover, even if the item itself does not have a social desirability, subjects will have a different choice on different representations of the same question.

II. Theoretical Framework

There is nothing more difficult to define than personality for psychologists. Since its birth of personality psychology, its theories are more and more with time elapsing; it is not exaggerating to say that the quantity of personality theories is equal to the quantity of personality psychologists. Some of the personality psychologists groped to establish a uniform framework for personality, but they failed and there's only beginning to take shape, personality theory is still in chaos. For example, Cattell considered intellectual characteristic as a personality trait, but some others didn't; though most psychologists thought of personality with no distinction between good and evil, and their scale excluded items that were evident in social desirability tendency, some others were the opposite. In recent years, as appealing to the unity of personality psychology is rising up, there has been a "big five" personality theories and "cognitive - affective system theory" that almost covers every personality dimensions and can be used in various situations, but they have not been recognized by all of the scholars [1].

Since this fragmented situation of personality cannot be concluded in a short time, it's more advantageous for us to research personality deeply within the theorists' own

Manuscript received May 14, 2014; revised August 12, 2014.

Sheyu Chen is with Nanjing Xiaozhuang University, China (e-mail: chensheyu@126.com).

Zhiyong Li is with Hubei Normal University, China (e-mail: yunwuji4@gmail.com).

framework than to develop a unified theory which is not accepted generally. And as more and more we understanding personality, we can establish a unified personality framework sooner or later. It is like "light" in physics, on the bases of various high validity researches of light's "wave" and "particle" by different scholars, "light" is ended conclusion of wave-particle duality ultimately. It's hard to imagine that Einstein and other scholars would have put those two together without thoroughly explored previous them.

If different scholars define personality and develop scale according to their own understanding of personality before reaching consensus, it will lead to a chaos state in personality research in short time, but it will eventually lead to the unity of the personality. In other words, the scholar's understanding of personality is a kind of theoretical framework; it provided guidelines to develop a scale. As long as the understanding is logical, and is actually descripted this controversial "personality", scholars can avoid confusions when developing the scale and can also guarantee that the scale has a rational explanation. Judging from the personality test development, any maturation scale such as 16PF, MMPI, CPI and MBTI, has a solid theoretical basis. In a word, only appearing large numbers of "fragmentation" studies can appear unified theory in the domain of personality psychology like the wave-particle duality in physics.

A. The Cultural Character of Personality Scale

Early in the personality study, scholars generally considered that the structure of personality traits was the same across races, and there were only differences on quantities, not on essential difference if there are differences. Consequently, the major work of trans-cultural studies is translating the instrument into another language when a personality scale was successfully developed. There will be some necessary changes when translating, but only replace some items that are not suitable obviously for local culture and customs, and even the scale was translated and applied for local people directly over a long period of time. Practice has shown that this approach is not appropriate [2].

With the rising of cultural psychology, there are more and more cross-cultural researches of personality, and have found that: although humankind in different area are facing similar natural geographical environments, different ethnic are influenced differently by the environments. While coming into civilized societies, different cultural traditions have been developed, and therefore have developed different cultural personality. Take David Ley for example, he contended that Chinese leaders are fond of "shi", and striking for a relative victory when competing with others, while American leaders are willing to seek an absolute victory [3]. Not only the users that translated personality scales to their own language observed the ethnic differences, but also the personality scale developers themselves paid more attention to such differences. This difference is not only in quantity but also in quality [4]-[6]. Moreover the differences of personality are evidence even in a similar cultural background [4]. Another problem we must take into account is that: almost every personality inventory is divided into large dimensions, and then organized into several factors in each dimension, but studies by Wang Dengfeng showed that while tested by 16PF, MMPI, EPQ inventory and re-extracted the factors using the data collected by these scales, the extracted factors or the items a dimension contained are very different from the original factors that the developer reported [4].

In a word, it's necessary to take cultural characteristics into account when developing a personality inventory; it will allow us to testing the traits of people in an appropriate cultural context more exactly.

B. Defining Dimension

Like other disciplines such as physics, it is difficult to achieve a balance between scientific and popular when a psychologist terming a concept. When the term is extremely scientific, the word that psychologist use will be very unfamiliar with people without psychological knowledge. The advantage is that it will force people to deliberate the meaning of the word, thereby avoiding a deviation of understanding bias. But the issue is, that scholars use uncommon words to term the concept of commonsense, will be not conducive for the public to accept a psychological glossary. Conversely, if terming a rigorous scientific concept with a colloquial words, it may be interpreted too literally by the reader and thus give a rise to understand the concept simply by their own knowledge regardless of scholars' definition. In either situation, it is detrimental to the development of scientific psychology. So it's necessary for psychologists to use scientific colloquial words to term an academic concept.

As has been discussed, it's necessary to develop a personality scale based on the theoretical framework, the next step is to prepare a two-way specification table, and the most important element of this work is to define each personality dimensions clearly. Due to the different theoretical frameworks, personality dimensions vary from one to another, even if they use the same words to term the dimensions in different theoretical framework, the implications are still different. At the same time, to take scientific and colloquial into account, because unless the subjects don't know anything about the term, they will understand the term with their own knowledge. That is to say, the subjects will have a typical meaning of the term by their self-awareness before comprehending developers' explanation. If the understanding of users is far away from the developer, and the developer does not give a reason for the particular meaning of the term, then the users will be very doubtful with the scale. Moreover, because of the stereotype effect, even though the users or subjects explain the score according to the developer's interpretation, they will misunderstand the meaning of the traits when retrieve the results subsequently. Take the responsibility dimension for example, if we define it as individual performance for seriously, carefully, and tendencies of firmness; subjects with a high score are high organized behavior and plan for future; whereas subjects with a low score are lack of directionality and self-discipline, impatient, more flexible when solving problems. In colloquial language, a high score of responsibility implies higher performance, and it is good; whereas a low score of responsibility implies lower performance, and it is bad. Obviously, the common

understanding of responsibility is far from the academic definition. It's better to term the concept as planning and flexibility, so that we can avoid from appraising a subject irrationally.

C. Participants' Characteristics

Sometimes, there is no choice but to develop different scales for deferent groups considering for their age, gender or culture context [7]. Also, we should explain the score differently. It is inevitable that there will be more and more particular scales for different group of people that divided by ages, races, or careers, etc., as the gradually more and more scientific and specified technique of developing a personality scale. According to the developmental psychology of personality theory, personality is not always the same since childhood to old; it also has a development process that is asymptotically and phases. And their response style or comprehension on the same item will be different from each other, so it is necessary to develop specified scale for different people.

III. Details of Developing an Item

A. Present with Appropriate Words

Use words easy to understand - The linguistic expression maybe vary according to their educational level and occupation status, so when developing items for a personality scale, we must: 1) the words used in the scale are easy to be understood; 2) considering the linguistic expression habits of most of the subjects that the scale maybe test; 3) the words used in the scale do not lead to different comprehension. In addition, the simpler words an item used, the shorter time when the scale is tested.

Each item covers only one concept - Each question can only refer to one concept, which can avoid confusing when subjects answering. Take the following question for example:

When disappointing, do you stick to it and try new methods to solve the problem?

The sentence covers two concepts, "stick to it", and "try new methods" to solve the problem. It will lead the subjects confusing when he or she agrees to only one of them. In this case, even though the participants make a choice, and we cannot know what the exact meaning of the response is. Therefore, in order to improve the accuracy of measurement, it's better to divide the question above into two independent items:

When disappointing, do you stick to solve the problem?
When disappointing, do you try new methods to solve the problem?

Avoid using subjective and emotive words - Objective and neutral words should be used in a questionnaire question, and words that can evoke emotional effects should not. Take the following question for example:

I usually refuse to learn from others when I make decisions.

The word "refuse" which implies negative meaning and maybe evoke emotional experience in this sentence is inappropriate. Thus, it's better to express the concept like the following sentence:

I usually want to learn from others when I make decisions.

Moreover, because there are not only neutral but also emotional words in our actually world, it is important to point out that if we develop a personality scale using words without emotional meanings, we cannot simple infer the results acquired by this type of scale into emotional situations. But even there are such adverse effects, using neutral words is appropriate because emotions maybe change from time to time, situation to situation, or event to event, and the more important is that the changes cannot be expected, this will lead to a unstable result from the emotional scale. In a word, using neutral words is necessary when developing personality scale.

Define options clearly - Options of each question in the questionnaire should be different enough to avoid overlapping in understanding and semantics. For example:

You have 6 weeks to complete an important task; you will take over the task:
A 5min later.
B 30min before the deadline.
C due to the concrete situation.
D immediately.

The choices A and D are semantic overlapping in this item, that is to say, compared to 6 weeks long; it is not distinctive for subjects to choose easily and consequently has negative effect to acquire the actual traits.

Not use too long sentences - Usually, the subjects are required to answer with their first reaction to the questions, but the accuracy of response depends on the accuracy of participants' understanding of the items. The issue is that it will force subjects to make trade-off between the accuracy of comprehension and the speed of reaction. The longer an item is, the less inaccuracy the subjects understanding it, and thus the lower reliability their responses are. So when developing a personality questionnaire, items should be short expression and easy to understand, then we can collect data more objective and reliable.

Privacy items - It is a troublesome problem to ask about privacy issues in personality scale, especially when it comes to aspects of sex, personal morals. But questions about these aspects sometimes are absolutely necessary, and even some personality scales simply set them as a dimension of personality study. How do we design items for this dimension? Item developing strategies for social desirability can be referred to. Of course, there are some other technologies such as projective technique that is very immature.

B. Take Psychological Effects into Account

Social desirability - The final purpose to developing a personality inventory is practice. Whether for clinical diagnosis and consultation, or for the selection and placement of personnel, we have to measure the personality

and the self-report personality scale is one of the most common survey tools. Generally in the domain of psychology, personality often is considered without good or bad, but it is social desirability in the view of public in daily life, and thus while measured by a self-report personality scale, people may cheat to present an impression which the interviewers are fond of, that is to say, subjects maybe manage their response in the scale to achieve their particular purpose. Consequently, it will challenge the reliability and validity of self-report personality scale. In order to reduce the impression management when answering a personality scale and to measure the participant's true personality, developers must transform the expression of items that are social desirability. Up to now, there are many strategies to control this effect. One of the important methods is generalize the social desirability item from direct question to indirect question. Take the following items for example:

Many people believe that "I'd betray all the people rather than letting them betray me", how do you think about it in your daily life?

In your daily life, do you believe "I'd betray the all the people rather than letting them betray me"

The former is better than the later, because of changing the expression of the idiom from definitely to generally and thus reducing the effect of ego defense. Not only should these negative tendencies of social desirability items be transform the expression, but also the positive items, then we can reduce subjects' vigilance. In addition to these methods, there are other ways to avoid occurrence of this effect [8].

The framing effect - The subjects' responses on the scale are deeply affected by the present way of questions. Changing the order of the options of an item in a scale, subjects may have different preferences. For example, Schuman and Scott (1981) [9] found that with two different expressions about divorce, there occurs evident recency effect. The two expressions are like the following:

Do you think in this country, divorce procedures should be easier, more difficult, or to maintain the status quo?

Do you think in this country, divorce procedures should be easier, to maintain the status quo, or more difficult?

One half of subjects answered the former question and the other half answered the later. For the former question, the selection ratio of every options are 23%, 36%, 41%, whereas for the later questions, the ratios are 26%, 29%, 46%. Moreover, when the questions or options are unfamiliar to subjects, this effect is particularly obvious. It is important to point out that when the item is forced-choice options or much more options settled, this effect does not exist [10]. Therefore, we'd better to develop scales that are forced-choice form or scales that are non-three options.

The responses of subjects not only affected by the order of options, but also affected by the formation of questions. Take the study by Tversky and Kahneman (1981) [11] for example, they asked their subjects essentially the same question in the following two forms:

If you want to see a movie. The ticket is 10$.When you arrive at the cinema door, found yourself dropped 10 dollars. Will you also spend 10 dollars to see the movie?

If you want to see a movie, and it took 10$ to buy tickets. When you arrive at the cinema door, found your ticket lost and impossible to find. Will you also spend 10 dollars for one ticket?

88% of respondents answered "Yes" in the former question, but the ratio is only 46% in the later form of the question. Actually, the two forms of the question mean the same, that is "will you spend 10$ more to see a movie". But participants' responses are different. Thus questions essentially the same may measure different psychological content. We should pay particular attention to such issues when developing personality scales.

Presupposition of questions - Presumption is a concept from qualitative research [12]. It refers to the pre-definition before a research or a question, and whether the researchers realize it or not, it indeed exists before the research or question. Its effect on the quality of a scale is vast, and almost every scale has its presumption. These presumptions often includes: 1) the scale actually measures the things that designed to measure; 2) each item in the scale is valid, and the understandings of the item among different subjects are the same, and the validity of each item is the same for all participants; 3) the reactions to the items reflect the actual trait one possesses; and so on. However, this is only an ideal state, and is researchers' wishful thinking; the fact is that the comprehension may vary from one to another. In order to develop a high-qualified scale, we must be extracted the common part of the comprehension of hundreds of thousands of users, and measure the differences between them, but it will undoubtedly increase the difficulty of developing a scale. It is truer for the development of the personality scale, thus personality developers usually pay special attention to the modification of the scale. It insures that every user has a similar understanding of each item, and thus we can measure the objective traits through deleting controversial items.

Another thing necessary to note is the arrangement of questions. Generally, at the beginning of a scale should be set some relatively neutral items, so that it can lead subjects to a good statement for testing. In the middle of the scale set items that are more social desirability and items related to personal privacy, so that it can reduce alertness and sensitivity of the subjects, and help to measure the actual personality of subjects.

C. Lie Detection

Lie detector is almost indispensable for personality, the reasons are: 1) items of personality always are social desirability; 2) subjects maybe disguise himself/herself as a good or a bad image because of his/her own interests; 3) subjects maybe driven to cheat on the scale by their ego defense mechanism, even when the researchers guaranteed that they are tested only for academic purpose and their privacy and results will be kept secret. Whereas up until now, there are not lie detectors that are reliable and validly enough, and validity of lie detectors are far from our true purpose.

Generally, lie detectors are often set whether two similar items appeared at different location of the scale, or contained absolutely good or bad items in the scale.

Actually, this is detected rather the subjects responding seriously or not, than lying itself. If subjects carefully answer the scale and take some strategies, they can deceive the interviewers easily, and thus losing the validity of lie detectors.

Moreover, "absolutely" is not "absolutely" but "relatively". Take the following question for example:

I have never lied.
A. yes B. no

We maybe not sure enough that the subjects lied when he/she responds "yes", this depends on how the subjects understanding the word "never". If they respond this question through remembering cases they experienced, they may not recall lying cases and they will answer "yes"; otherwise if they recall at least one lying case, they will answer "no". Therefore, it is difficult to detect lying or not when the subjects are required responding with their first reaction to the item. Consequently, in the former, though the option they choose implies lying, we should not believe that the subjects are lying. It is a paradox that we judge a subject lying when he/she answer the question actually. Similarly, if they answer this question through logical thinking, they will answer "no", because one will more or less lie in some cases. In this case, the items investigate rather logical thinking than lie. In a word, lie detectors that we use are not valid enough.

In addition, the assumption of these lie detectors is, bad behaviors are universal among people, and subjects usually are tending to admit [13]. But it is working well only on the premise that participants are not aware of this assumption. Once realizing the assumption, that the participants can manage their reaction to shape an honest individual, and consequently leading to malfunction of lie detectors.

To sum up, the validity of lie detectors depends on many factors, which make its effectiveness greatly reduced. Nevertheless, we have to use them and interpret the score carefully until more valid instruments are developed. Now some researchers [14] propose a new lie detector method, which uses "logical traps" to detect lying subjects, but it has not yet to be used in personality test.

IV. Conclusion

As a extensively studied field of personality, psychologist haven't developed a powerful instrument to explore the essence or traits of personality, which leads directly to the inconsistency of the findings in the study of personality that different researchers using different personality questionnaire have found different results even studying on the same issue. One of the most effective methods to resolve this situation is to develop powerful personality scales. Currently, it is almost impossible for researchers to develop a broadly accepted personality questionnaire; even the five-factor theory of more cross-cultural consistency also has cultural adaptive difficulties in Germany, the Netherland and the United States [15], Italy [16], Philippine [17], China [4], [18]. Moreover, because of the context effects of personality, it's also difficult to develop a scale that can be adaptive for any group of people in a particular culture. In addition, due to the understanding of the deviation, the participants' understandings of items may differentiate from the meaning that conveyed by the developer. All of these factors lead to unexpected reliability and validity of personality researches, and in turn influenced the further research on personality. So we should consider certain cultural backgrounds and subjects' characteristics such as ages, careers and educational contexts, and the response-bias should be considered at the same time, when developing a perfect personality scale. Only based on researches of concrete and effective that we can realize the essence of personality and prepare for the unity of personality theory.

References

[1] S. H. Luo, "The future of personality psychology: Waiting for 'big one'," *Journal of Developments in Psychology*, vol. 6, pp. 21-24, 1998.
[2] R. E. Nisbett, "Perception and cognition: West-East differences," in *Proc. ICP 2004*, Q. C. Jing, M. R. Rosenzweig, G. D. Ydewalle, H. C. Zhang, H. C. Chen, and K. Zhang, Eds., vol. 2, London: Psychology Press, 2006.
[3] D. Ley, "The essence of China's strategic thinking: 'Shi' compiled by H. F. Zhang," *Contemporary Military Digest*, vol. 2, pp. 16-18, 2005.
[4] D. F. Wang and H. Cui, "Is neuroticism an independent dimension of Chinese personality structure," *Journal of Southwest China Normal University*, vol. 31, pp. 25-30, 2005.
[5] D. F. Wang and H. Cui, "Theoretical analysis of the Chinese 'big seven' personality structure," *Essays of Personality and Social Psychology*, vol. 1, pp. 46-84, 2004.
[6] K. S. Yang and D. F. Wang, "Personality dimensions of Chinese people," presented at the Third Congress of Chinese Psychology, Beijing, 1999.
[7] S. A. Xu and J. J. zhang, "A study on personality structure of intellectuals," *Psychological Exploration*, vol. 29, pp. 46-51, 2009.
[8] Y. H. Li *et al.*, "Study and progress of socially desirable responding problem in personality measurement," *Chinese Journal of Clinical Rehabilitation*, vol. 9, no. 8, pp. 119-121, 2005.
[9] H. Schuman and J. Scott, "Problems in the use of survey questions to measure public opinion," *Science*, vol. 236, pp. 957-959, 1987.
[10] S. Plous, *The Psychology of Judgment and Decision Making*, Beijing: Posts & Telecom Press, 2004, p. 48.
[11] A. Tversky and D. Kahneman, "Belief in the law of small numbers," *Psychological Bulletin*, vol. 76, pp. 105-110, 1971.
[12] X. M. Chen, *Qualitative Research in Social Sciences*, Beijing: Educational science publishing house, 2000, ch. 1, pp. 3-24.
[13] G. M. Chen, "The adaptive analysis of Western lie detectors in Chinese cultural context," *Journal of Ningbo University (Educational Science Edition)*, vol. 20, no. 4, pp. 10-14, 1998.
[14] A. B. Dong, "The method of logical traps for lie detection and its mathematical testifying," *Acta Psychologica Sinica*, vol. 26, no. 2, pp. 176-183, 1994.
[15] W. K. B. Hofstee, H. Kiers, B. DeRaad, L. R. Goldberg, and F. Ostendorf, "A comparison of big five structures of personality traits in Dutch, English, and German," *European Journal of Personality*, vol. 11, pp. 15-31, 1997.
[16] L. D. Blas and M. Forzi, "An alternative taxonomic study of personality-descriptive adjectives in the Italian language," *European Journal of Personality*, vol. 12, pp. 75-101, 1998.
[17] A. T. Church, M. S. Katigbak, and J. A. S. Reyes, "Toward a taxonomy of trait adjectives in Filipino: Comparing personality lexicons across cultures," *European Journal of Personality*, vol. 10, pp. 3-24, 1996.
[18] F. Cheung, K. Leung, J. X. Zhang, H. F. Sun, Y. Q. Gan, W. Z. Song, and L. Xie, "Indigenous Chinese personality constructs: Is the five factor model complete?" *Journal of Cross-Cultural Psychology*, vol. 32, pp. 407-433, 2001.

Sheyu Chen was born in October 1963 in Jingjiang City, Jiangsu Province, China. He earned his doctor's degree of education majored in psychometrics from Nanjing Normal University China, and engaged in a two-year postdoctoral research at Nanjing University China.

He is currently a professor in Nanjing Xiaozhuang University, and worked in the Personnel Department of Jiangsu Province and then Yancheng Teaching

University. He is employed as an expert of interview or test by many departments of the central government. Currently he has published 8 pieces of scholarships and 40 papers. Representative works include a study of the validity of the administration occupational aptitude test, The assessment of reliability concepts of classical true score theory of generalizability theory. His major research interest is selection and examination of civil servants, is also involved in education science, performance management, talent development, etc.

Two research findings of Dr. Chen had awarded the Excellent Achievement Prize of Philosophy and Social Science which were granted by the People's Government of Jiangsu Province; one of his papers was awarded Essay Prize of the First China Human Resources Development and Management; several research findings were adopted by the provincial and municipal governments.

Zhiyong Li was born in February 1981 in Baoding City, Hebei Province, China. He earned his master's degree of Science majored in social psychology and personnel assessment from Nanjing Normal University in China.

He's currently a lecturer in the Department of Educational Science of Hubei Normal University. He has served as an interviewer and test-developer in the civil servants selection, as a trainer in psycho consultant and human resource manager training. He has published textbook of new elementary psychology (Nanjing: Nanjing Normal University Press, 2009), translated the psychology of interpersonal relationships into Chinese. His major research interests are psychometrics, selection and examination of civil servants, behaviors and decisions.