

Comparing frequency and Trueness Scale Descriptors in a Likert Scale Questionnaire on Language Learning Strategies

著者	Mizumoto Atsushi, Takeuchi Osamu
journal or publication title	日本語テスト学会研究紀要 = JLTA Journal
volume	12
page range	116-136
year	2009
URL	http://hdl.handle.net/10112/4497

**Comparing Frequency and Trueness Scale Descriptors
in a Likert Scale Questionnaire
on Language Learning Strategies**

**MIZUMOTO, Atsushi
TAKEUCHI, Osamu**

**日本語テスト学会 研究紀要 第12号
JLTA Journal No. 12
抜刷
2009**

Comparing Frequency and Trueness Scale Descriptors in a Likert Scale Questionnaire on Language Learning Strategies

MIZUMOTO, Atsushi (University of Marketing and Distribution Sciences)
TAKEUCHI, Osamu (Kansai University)

Abstract

This paper reports on the comparison of two types of scale descriptors in a questionnaire on language learning strategies. The main purpose of this study was to investigate which type of the two scale descriptors, frequency-based or trueness-based, is better for language learning strategy research. With a few weeks' interval, a questionnaire on learning strategies was administered to 408 EFL learners twice with frequency-based scale descriptors and trueness-based scale descriptors alternately. First, mean differences in the responses obtained from the two different scale descriptors were examined. Second, confirmatory factor analysis was applied to see which scale descriptor of the two shows better fit to the hypothesized model. Finally, equidistance between the categories was checked. Results show that trueness-based scale descriptors elicited slightly higher mean values of responses, which resulted in a better fit to the model. The distances between the categories were almost identical for both scale descriptors. The findings of the current study partially provide supportive evidence for the claim made by Dörnyei and his colleagues that trueness-based scale descriptors are preferable to frequency-based counterparts in a questionnaire on learning strategies. But they also show that the latter descriptors are not totally "flawed" as was claimed.

1. Introduction

Conventionally, researchers in the disciplines of social and behavioral sciences have utilized questionnaires in their research in order to investigate human behaviors and their underlying latent constructs. In the field of applied linguistics as well, especially after studies investigating individual differences came to the foreground of research interest in the late 20th century (See Dörnyei, 2005; Robinson, 2002 for a comprehensive review), researchers in the SLA field have increasingly employed questionnaires in their studies.

Among a number of rating scale questionnaires, the Likert-type questionnaire is "the most commonly used scaling technique" (Dörnyei, 2003, p.36). Although Likert scale questionnaires face severe criticisms regarding validity as a measurement of instrument, they have withstood the test of time, mainly because they are relatively easy to construct

and easy to administer to a large number of participants. Profoundly aware of the limitations of Likert scale questionnaires, however, many researchers (e.g., Gu, 2003; Takeuchi, 2003) have advocated the mixed use of quantitative and qualitative research methods. Since what can be clarified from a questionnaire study represents only a small fraction of a learner's true self, they suggest that qualitative research methods such as interviews, think-aloud procedures, stimulated recalls, or portfolios should be utilized to provide complementary evidence to the findings obtained from questionnaires. In spite of these criticisms and suggestions, due to its enormous practicality, studies using the Likert scale questionnaire as a main instrument of measurement will surely survive to shed light on aspects of learners' individual differences (as exemplified in the work by Vandergrift, 2005). As such, a Likert-type questionnaire has to be reliable and valid as an instrument of measurement. To this end, a study intending to further expand and elaborate on a technical aspect of a Likert-type questionnaire is undoubtedly necessary. Our current study has been undertaken to pursue this objective in the field of language learning strategies.

2. Literature Review

2.1 Definition of a Likert Scale

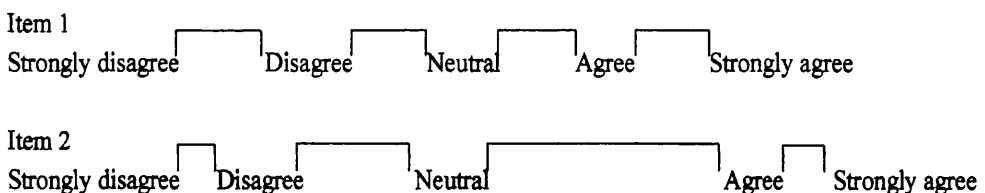
The Likert scale (Likert, 1932) was devised by Rensis Likert, an American psychologist. In Likert scales, researchers commonly ask respondents to mark their attitudes toward certain issues expressed in items. According to Aiken (1996, p.262), a Likert scale is an "(A)ttitude scale in which respondents indicate their degree of agreement or disagreement with a particular proposition concerning some object, person, or situation." While several variations on Likert scales exist (Dörnyei, 2003, p.38), the fundamental assumption is that the opinions expressed along a continuum from least to most (in the Likert scale, "strongly disagree" to "strongly agree") indicate the intensity or tendency of the respondents' characteristics to be measured. Sometimes the scale descriptors come with numbers, e.g., 1 through 5, to express judgments. Common practice after eliciting responses is to sum up the items belonging to the same category or subscale, and possibly divide by the number of the items in the category. A Likert-type questionnaire can be an inventory, or a mere list of items. However, a Likert scale questionnaire which is designed to be called "a psychometrical scale" will measure the same construct with more than two items (called multi-item scales, Dörnyei, 2003, p.32). A reliability coefficient for each subscale and the score of each subscale is calculated by adding up the items (or averaging them). These procedures are so straightforward that it has made the Likert-scale questionnaire the assessment instrument of choice for researchers around the world.

2.2 Considerations for Using Likert Scales in Research

2.2.1 General Concerns for the Likert Scale Use

The first problem with Likert scales, which has been pointed out by many, is the ambiguity of wordings. It is natural that individuals have different ideas toward a certain topic. As for the scale descriptors, it is often the case that the categories of “agree” and “strongly agree” practically mean the same for some respondents depending on the frame of reference they have in mind. In the learning strategy research, the scale descriptors in the questionnaire sometimes employ the degree of frequencies (from *Never* to *Always*) since strategic learning behaviors are the target of measurement. Related to this specific use of the Likert scale, Gu, Wen, and Wu (1995) investigated the research question “How often is often?” They then demonstrated that depending on the different reference systems given to the respondents, i.e., compared with friends, their past learning experiences, or other language skills, the results would be significantly different even with the same questionnaire. Also, some researchers (Oda, 1970; Spector, 1976; Wakita, 2004) have demonstrated how adverbs in the wordings of a questionnaire can cause respondents to have different semantic references.

Rather technical, but by far the most crucial problem in utilizing a Likert-scale questionnaire in research is the type of data obtained from a questionnaire. Clearly, the responses for a questionnaire are ordinal data. In fact, in this example below, choosing “Agree” in item 2 means the same as choosing “Strongly agree” in item 1 in their psychometrical distance. That is, the distances between “Agree” and “Strongly agree” in these two items are not the same and cannot be regarded as equidistant, that is, interval data (Bond & Fox, 2001).



This example depicts the difficulty of assuming a Likert-type questionnaire will naturally produce interval data. Another related problem is summing or averaging the data obtained from a questionnaire. Specifically, the practice of adding or dividing the scores is based on the assumption that each item contributes to the construct equally in multi-item scales. It would be obviously questionable, however, to regard two responses as the same when one respondent chooses 1 and 5 for two items and the other choosing 3 and 3, even though the summated or average scores are the same for these two

individuals, i.e., 6 points in total (or mean score of 3). In spite of these drawbacks, it is now common practice to treat data from more than four-point Likert scale as interval scales and can be used for statistical analyses such as factor analysis (e.g., Hagiuda & Shigemasu, 1996). Also, Hatch and Lazaraton (1991) suggest that the wider range of scales would yield data resembling interval scale.

2.2.2 Problems of Using Questionnaires in Learning Strategy Research

Dörnyei and his colleagues (Dörnyei, 2005; Tseng, Dörnyei, & Schmitt, 2006) have argued that using a questionnaire asking “specific strategic behaviors and the scale descriptors indicating frequencies of strategy use” is not psychometrically justifiable. They argue that this is because “we cannot assume a linear relationship between the individual items scores and the total scale scores” (Tseng et al., 2006, p.83). They took *Strategy Inventory for Language Learning* (SILL; Oxford, 1990) as an example of such a “flawed” (Dörnyei, 2005; Tseng et al., 2006) assessment instrument of learning strategy.

Dörnyei and his colleagues, however, contend two things at once in their discussion: the problem of stems (statements) and that of scale descriptors. These two issues in fact should be discussed separately in the development of a valid assessment instrument. As a starting point of the endeavor to elaborate and advance the questionnaire study on learning strategies to a higher level, we thus decided to focus on one aspect of these two problems—comparing scale descriptors using the expressions of frequency with those of trueness—to examine which is more appropriate for a learning strategy questionnaire with items asking specific strategic behaviors.

3. The Study

3.1 Research Questions

This study addresses the following research questions: By changing the scale descriptors, (1) is there any difference in responses between frequency-based descriptors and trueness-based descriptors, (2) of the two descriptors, which one is more preferable, and (3) of the two descriptors, which one is more close to an interval scale?

3.2 Participants

The participants of this study were Japanese university EFL learners in five private universities in western Japan (humanities or engineering majors, aged 18-22). With an interval of two to four weeks, all of them were required to respond to the questionnaire twice: one with frequency-based scale descriptors, the other with trueness-based scale descriptors. After two administrations, listwise deletion left 408 participants. Proficiency of the learners was measured with their TOEIC IP (the Test of English for

International Communication, Institutional Program) scores available from only those who had taken this test in the last one-year period ($n = 281$, $M = 397.91$, $SD = 126.0$). The TOEIC IP consists of the listening section (100 items) and the reading section (100 items). The full score for each section is 495, making 990 the total full score. According to Educational Testing Service (2006), the test developer, "TOEIC has been used to measure the English proficiency of nonnative English-speaking people." According to the TOEIC Steering Committee (2006), the mean scores of TOEIC for university humanities majors are 474 and for engineering 397. Therefore, the participants of the current study can be regarded as average or lower-level university students of EFL.

3.3 Instruments

We consulted a number of instruments in the field of language learning strategy which suit the purposes of the current study. That is, ones in which the items are clearly asking certain language learning behaviors, as are the case with most of the past learning strategy questionnaires. Also, the items have to be ones to which both frequency and trueness scale descriptors can be applied. After several rounds of discussion, we decided to use the learning strategy questionnaire developed by Hiromori (2004). In his study, he devised ten items on cognitive strategies and another set of ten items on metacognitive strategies by referring to such questionnaires as in Oxford (1990) and Purpura (1999). The authors modified the wordings of the items slightly so that it would be compatible with the purpose of the study. For the purpose of reducing the possibilities of giving the impression that the same questionnaire was administered twice to the participants, the item number was randomized in the questionnaire of trueness-based scale descriptors.

3.4 Procedures

The questionnaire was administered the first time in December, 2006 and the second time in January, 2007. The interval between these two seemed appropriate because there was a two-week winter break in between and it was unlikely that the participants' English proficiency or strategy use would change drastically during that period. To control the order effect (e.g., In'nami, 2006), the participants were randomly assigned to one of the two questionnaires, with one group (Group 1) responding to a questionnaire with frequency-based scale descriptors first ($n = 189$), the other (Group 2) with trueness-based scale descriptors first ($n = 219$). The second administration was conducted without any prior notice. In addition, feedback of any kind after administering the questionnaire was avoided so as not to make affective differences on the side of participants. Table 1 summarizes the procedures of the questionnaire administration.

Some empirical studies have shown that the validity of employing a questionnaire

can be further enhanced by actually presenting a task to the learner before responding to it (e.g., Ikeda & Takeuchi, 2000; Oxford, Cho, Leung, & Kim, 2004; Qian, 2004). However, we did not present any task in this study mainly because the questionnaire included more or less general questions asking the learner's strategy use and the purpose of the current study, i.e., comparing two types of descriptors, could be achieved without such a task at hand. A sample of the questionnaire items (See Appendix A for all the items) and two types of scale descriptors are as follows:

Example: Item 3 in Cognitive strategies (originally in Japanese)

I look for Japanese words that are similar to new words in English.

[Frequency-based descriptors]




1. I never do it
2. I rarely do it
3. I sometimes do it
4. I often do it
5. I almost always do it

[Trueness-based descriptors]

1. Not at all true of me
2. Not true of me
3. Somewhat true of me
4. True of me
5. Very true of me

Table 1

Overview of the Procedure

1st administration (December, 2006)	2nd administration (January, 2007)	Combined data
Group 1 (n = 189) Frequency-based - Cognitive strategies - Metacognitive strategies	 Trueness-based - Cognitive strategies - Metacognitive strategies	 Frequency-based - Cognitive strategy - Metacognitive strategy
Group 2 (n = 219) Trueness-based - Cognitive strategies - Metacognitive strategies	 Frequency-based - Cognitive strategies - Metacognitive strategies	
		Trueness-based - Cognitive strategy - Metacognitive strategy

3.5 Data Analyses

Since the authors modified the wordings of the original instrument by Hiromori (2004) and it was originally intended for Japanese high school EFL learners, we needed to put the instrument under scrutiny. In order to verify the unidimensionality of the instrument, the Rasch Rating Scale model (Andrich, 1978) was first applied using *WINSTEPS 3.63.2* (Linacre & Wright, 2000) on 10 items assessing cognitive strategies

and 10 items on metacognitive strategies in frequency-based and trueness-based descriptors respectively. The rationale behind utilizing the Rasch analysis lies in the fact that it can change ordinal measures (raw scores) to interval measures. With interval measures, we can say that a certain item in a psychological measurement is more (and how much more) difficult than others for the respondents to endorse. In addition, the Rasch model can detect misfit items, which show a departure, if any, from unidimensionality of the construct. A conventional rule of thumb for checking acceptable items using the Likert scale is the infit/outfit mean square ranging from 0.6 to 1.4 (Bond & Fox, 2001; Wright & Linacre, 1994). Following this criterion, the application of the Rasch model detected only one misfit items in the item No. 2 of Trueness-based scale (refer to Appendix B). Thus, this item was excluded in both Frequency-based and Trueness-based scale descriptors in the subsequent analyses. Exploratory factor analysis (maximum likelihood extraction with promax rotation) was then performed in order to double-check the unidimensionality of the data. Consequently, all the subscales proved to produce only one factor. The subscale scores were calculated using the mean of the items. Following the calculation of subscale scores, we computed internal consistency reliability estimates with Cronbach alpha.

After we generated the subscale scores for each scale, a paired *t*-test was conducted with the means of two scale descriptors for the purpose of investigating the first research question: Is there any difference between frequency-based descriptors and trueness-based descriptors? Also, correlation coefficients were checked to examine the extent of relationships between the two descriptors.

For answering which descriptor is more preferable of the two (research question 2), we employed confirmatory factor analysis to show which data with the two scale descriptors produce better fit indexes. Confirmatory factor analysis is used when “one derives a factor model or models a priori (i.e., reasoning deductively to hypothesize a structure beforehand) and then evaluates its goodness of fit to the data” (Bryant & Yarnold, 1995, p.109). When performing confirmatory factor analysis, it was assumed that in our present study the two subscales had high correlations and they could be considered two factors under the umbrella of more general factor, “learning strategies.” But we did not pursue this idea because the main thrust of the current study was to compare the two descriptors. In addition, such correlated model produced fit indexes which are almost the same as the model tested (i.e., four models specified and tested separately) in the current study. From these reasons, having two factors for the comparison of two descriptors was thought to be more desirable and informative. Figure 1 shows the model tested in confirmatory factor analysis. All of these analyses above were performed using *PASW Statistics 17* and *Amos 17.0*.

Finally, we investigated which type of the scale descriptors is close to an interval scale (research question 3) by examining the threshold estimates of the Rasch Rating

Scale model. The Rasch Rating Scale model assumes that all the items in the questionnaire share the same distance of threshold, whereas another model, the Rasch Partial Credit model (Wright & Masters, 1982) “allows item-by-item analysis of step structure with items involving rating scale steps of any kind” (McNamara, 1996, p.256). In the current study, since our purpose was to investigate which of the two types of scale descriptors were close to interval scale, not the properties of each item, we chose to apply the Rasch Rating Scale model over the Partial Credit model.

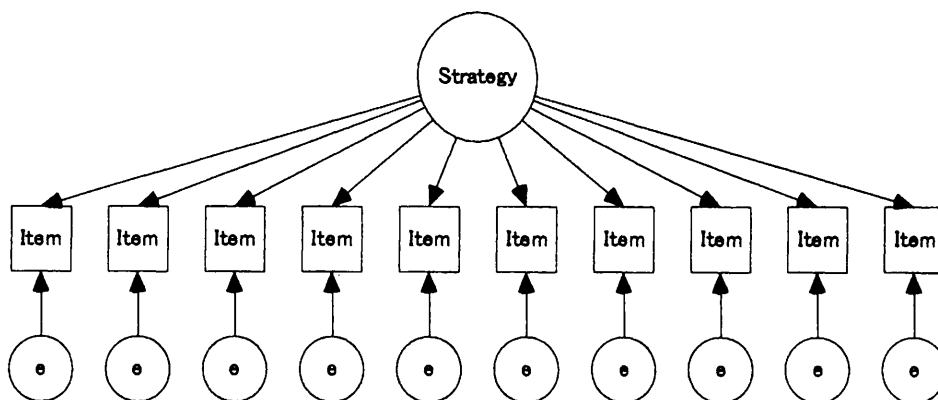


Figure 1. The model tested in the confirmatory factor analysis. The same model was tested for the four types of scales: (a) Cognitive strategies with frequency-based scale descriptors, (b) Metacognitive strategies with frequency-based scale descriptors, (c) Cognitive strategies with trueness-based scale descriptors, and (d) Metacognitive strategies with trueness-based scale descriptors. Nine items were used for (a) Cognitive strategies with frequency-based scale descriptors and (c) Cognitive strategies with trueness-based scale descriptors because one item was excluded after diagnosed as misfit.

4. Results and Discussion

The descriptive statistics, Rasch item difficulty estimates and fit statistics for each item are presented in Appendix B. Table 2 shows the means of the composite scales (subscale scores), standard deviations, Cronbach alphas, and the Pearson correlation coefficients of two types of descriptors. The means of trueness-based descriptors were higher in both strategies. Cronbach alphas were moderately high. Pearson correlation coefficients of the two descriptors were not so high (Cognitive strategies, $r = .61$; Metacognitive strategies, $r = .68$), given the fact that the statement of the questionnaire

in the first and the second administration was the same. However, Cronbach alphas of the two types of scale descriptors, as a matter of course, did not produce perfect scale reliability. We thus corrected the correlation for attenuation. When a correction formula is applied, the corrected correlation yielded much higher figures (Cognitive strategies, $r = .89$; Metacognitive strategies, $r = .83$).

A paired t -test (two-tailed) revealed that the mean scores for two types of scale descriptors were significantly different with, according to Cohen (1988), the small to medium degree of effect sizes (Cognitive strategies, $t(407) = -11.09$, $p < .05$, $r = .48$; Metacognitive strategies, $t(407) = -4.63$, $p < .05$, $r = .22$). A test result could be statistically significant even when there is no true effect, due to a large number of samples used (Field & Hole, 2003). However, it was not the case with the current results because effect size is a measure that is not influenced by sample size. A 95% confidence interval on the difference between the two population means for Cognitive strategy (frequency minus trueness) is (-0.28, -0.19) and for Metacognitive strategy (frequency minus trueness) is (-0.15, -0.06). Even though the figures are small, the interval for these two strategies does not include 0. This indicates that we can be sure that, at the 0.05 level of significance, there is a statistically significant difference in the population means.

From these results, we might as well conclude that the two different types of descriptors, with the standard practice of using the mean scores of the subscales, produced a different result even when the same items were given to the same respondents.

Table 2
Descriptions of the Two Scale Descriptors in the Two Subscales

Subscale	Type of descriptors	No. of items	M	SD	α	r	$t(407)$	95% confidence interval of the difference	
								Lower	Upper
Cognitive strategy	Frequency	9	2.71	0.58	.70	.61	-11.09*	-0.28	-0.19
	Trueness	9	2.91	0.47	.67				
Metacognitive strategy	Frequency	10	2.82	0.59	.83	.68	-4.63*	-0.15	-0.06
	Trueness	10	2.93	0.52	.80				

Note. $N = 408$, *adjusted $p < .05$, which means the α -level is set a priori at $0.025(0.05/2)$.

As for the second research question—of the two descriptors, which one is more preferable?—fit indices generated as a result of confirmatory factor analysis using structural equation modeling (SEM) were examined for each of the strategies and scale descriptors. Since it is common practice to compare several fit indices to prove the fit of the model to the observed data in SEM (See Tabachnick & Fidell, 2006 for details), we compared several measures of indices. Table 3 summarizes these indices.

From Table 3, it is clear that trueness-based scale descriptors yielded better fit indexes compared to frequency-based scale descriptors in all combinations. The rule of thumb for acceptable fit index of the ratio of the χ^2 to the degrees of freedom is 2 or less (In'nami, 2006). Although the scale descriptors in both Cognitive and Metacognitive strategies showed that χ^2/df is slightly over this criterion, it is still better than the figures produced with frequency-based descriptors. The acceptable index figures for GFI, AGFI, and CFI are above .90, with “higher values indicating better fit” (Bryant & Yarnold, 1995, p.112). Based on this criterion, trueness-based scale descriptors in all pairs showed better fit indices. As for RMSEA, Arbuckle and Wothke (1999) suggest that “RMSEA of about .05 or less would indicate a close fit to the model,” while the maximum acceptable value is .08. The smaller the better is the case for this index. In this instance as well, trueness-based scale descriptors were better compared to frequency-based scale descriptors. AIC is used in comparing competing models and smaller values indicate a better fitting (Tabachnick & Fidell, 2006, p.719). Trueness-based scale descriptors obviously resulted in smaller AIC values, meaning they are better than frequency-based scale descriptors in this model.

Table 3
Summary of the Fit Index in Confirmatory Factor Analysis

Fit index	Cognitive strategies		Metacognitive strategies		Criteria for acceptable fit
	Frequency	Trueness	Frequency	Trueness	
Chi-square (χ^2)	184.09	77.63	144.73	72.96	$p > .05$
χ^2/df (degrees of freedom)	$p < .05$	$p < .05$	$p < .05$	$P < .05$	< 2.0
GFI	.911	.963	.929	.966	$> .90$
AGFI	.860	.942	.888	.946	$> .90$
CFI	.748	.904	.894	.951	$> .90$
RMSEA	.102	.055	.088	.052	$< .05$ $< .08$
AIC	224.09	117.63	184.73	112.96	-

Note. GFI = goodness of fit index, AGFI = adjusted goodness of fit index, CFI = the comparative fit index, RMSEA = root mean square error of approximation, AIC = Akaike information criterion

In sum, all the goodness of fit indexes corroborated that the model with trueness-based scale descriptors is showing a better fit to the observed data. It should be noted that when used for large samples in confirmatory factor analysis, the Chi-square statistic (χ^2) is not a good indicator of overall goodness of fit because it is “extremely sensitive to sample size” (Bryant & Yarnold, 1995, p.112). Therefore, the fact that the *p*-values of all the Chi-square values were significant does not necessarily indicate the hypothesized model misfits the data.

In order to examine the third research question—of the two descriptors, which one is more close to interval scale?—the Rasch Rating Scale model was applied. Table 4 shows the threshold for each scale descriptor in two types of the questionnaires.

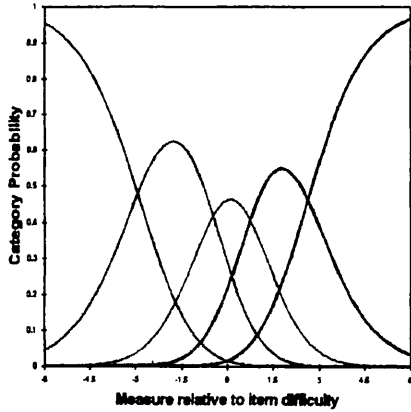
Thresholds are the steps that divide each category, e.g., the point between 2 and 3 on a 5-point Likert scale. In other words, the threshold is “the level at which the likelihood of failure at a given response category (below the threshold) turns to the likelihood of success at that category (above the threshold)” (Bond & Fox, 2001, p.70). By investigating the distances of thresholds between the categories when ordinary scale is converted into interval scale with the help of the Rasch model, we will be able confirm which of the two scale descriptors have a property which is close to equidistance.

We illustrated the thresholds for frequency-based and trueness-based scale descriptors—Cognitive strategies and Metacognitive strategies respectively (Figure 2). In Figure 2, thresholds are places where each curve is crossed. Each curve in the figure represents the probability of one person with certain ability, strategy use in this case, choosing one category over the others. If the distance for each category is equal, it shows that the responses in the questionnaire are interval data. From the figure, it seems that the distances between categories are almost the same for each pair (Cognitive and Metacognitive strategies) in frequency-based and trueness-based scale descriptors. Therefore, it is difficult from the result to distinguish which type of scale descriptors can yield scales more close to interval scale. In terms of equidistance between categories, the two types of scale descriptors could be regarded as the same in their functions.

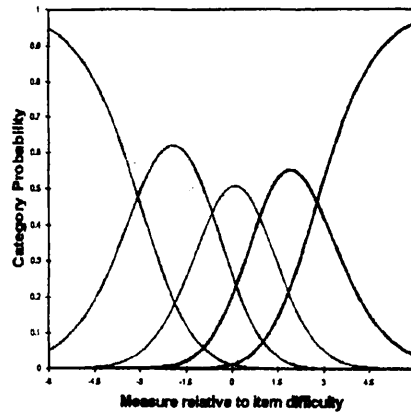
Table 4
Thresholds for Each Scale Descriptor

	Cognitive strategies		Metacognitive strategies	
	(1) Frequency	(2) Trueness	(3) Frequency	(4) Trueness
Threshold 1	-2.94	-3.07	-3.12	-3.56
Threshold 2	-0.46	-0.64	-0.61	-0.62
Threshold 3	0.76	0.91	0.95	1.37
Threshold 4	2.63	2.80	2.78	2.80

Cognitive strategies

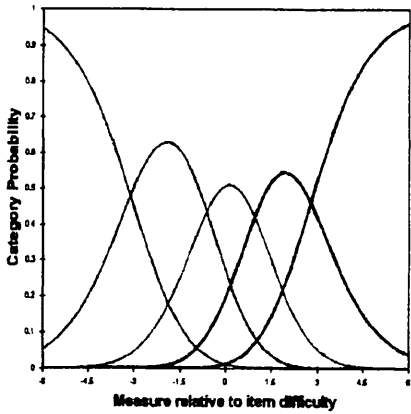


(1) Frequency-based descriptors

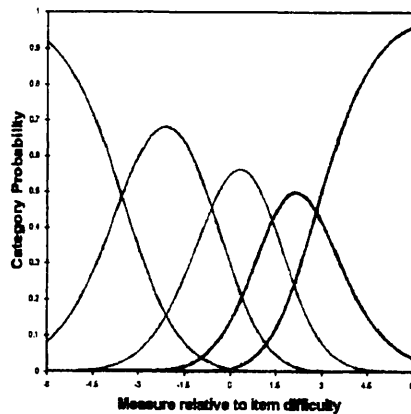


(2) Trueness-based descriptors

Metacognitive strategies



(3) Frequency-based descriptors



(4) Trueness-based descriptors

Figure 2. Visual representation of the thresholds for each scale descriptor. Item difficulty is expressed in logits.

The findings of the current study can be summarized as follows: (1) there were significant differences between the responses for two descriptors in the two scales of learning strategies (i.e., Cognitive strategies and Metacognitive strategies) investigated, (2) trueness-based scale descriptors produced better fit indexes to the model; thereby, making it superior to frequency-based scale descriptors, and (3) both were almost identical when converted into interval scale. By synthesizing these findings, we can argue that two scale descriptors in a learning strategy questionnaire elicit differing responses from the learners. If we are to choose one of these two descriptors, the trueness-based scale descriptor is better because it is more likely to fit the data in a hypothesized model. However, we do not mean to claim that the studies on learning strategies employing frequency-based scale descriptors in the past have produced misleading findings. As can be seen in the result of the Rasch Rating Scale model, frequency-based scale descriptors are no less accurate as an instrument of measurement than trueness-based scale descriptors. Trueness-based scale descriptors are superior to frequency-based descriptors only when the question is posed: Which of the two types of descriptors is better suited for the studies of learning strategies?

At this point, we need to contemplate what caused the differences in the results between the two scale descriptors. In order to seek an explanation regarding this matter, a follow-up inquiry for randomly sampled participants ($n = 14$) via email was conducted in order to explore how the learners regarded the two types of questionnaires. In the response, some wrote that they noticed the differences of the scale descriptors, while others noted that they were not aware of the changes in the wordings. One particular participant described in the response for the question “why do you think your responses were different in the two questionnaires?”:

I think the statement “I do something” asks what I do “NOW,” but the statement “true of me” can be interpreted as “I try to, but actually I don’t do it” or “I used to, but not now.” These differences in interpretation might have caused my responses to change. (A. K. 500T; translation ours).

With regards to self-report questionnaires used in learning strategy research, Tseng et al. (2006) suggest that “questionnaires in this area are based on the assumption that strategy use and strategic learning are related to an underlying *trait* because items ask respondents to generalize their actions across situations rather than referencing singular and specific learning events” (p.82). Interestingly enough, what this participant has expressed in the comment above reflects the basic assumption of self-report questionnaire in learning strategy research (Tseng et al., 2006). In other words, with trueness-based scale descriptors, respondents of the questionnaire can generalize their

actions, showing something latent behind the action, which results in a better fit to the model. This finding at the same time implies that, with frequency-based scale descriptors, it might be difficult for respondents to generalize their learning behaviors. In this regard, if the purpose of a study is to measure true learning behaviors, not “generalized actions” which can be presumably measured with a self-reported questionnaire, we strongly recommend using other alternative assessment methodology reviewed by Cohen and Scott (1996) or Macaro (2006) in order to accurately depict the actual strategy use.

The findings of the current study “partly” support the argument made by Dörnyei (2005) and Tseng et al. (2006) that measuring specific strategic behaviors might lead to psychometrically unjustifiable results. We have used the word “partly” because, although we reached the conclusion that trueness-based scale descriptors are better than frequency-based ones as far as the scale descriptors are concerned, it does not necessarily mean that frequency-based scales produce dramatically wrong results. Indeed, the interval nature of the scales proves to be almost the same for both scale descriptors. The other reason we used “partly” is that we investigated only the scale descriptors without changing the wordings in the statement, or items per se. Further studies focusing on the wordings of items will be necessary to truly verify that measuring specific learning strategies with a frequency-based questionnaire is in fact psychometrically unjustifiable.

While this study has revealed some interesting findings, a few limitations should be pointed out here. First, we based our analyses on the premises that it would not lead to considerable differences when each type of scale (frequency and trueness) was administered in test-retest manners (Brown, 2005). In fact, a study by Sakui and Gaies (1999) discovered that test-retest reliability for the same questionnaire was not as high as they had expected. From this point of view, measurement error, if any, might have played some role in the responses of the participants. In addition, the assignment of numbers (from 1 through 5) in front of the scale descriptors might have had a marginal influence on the responses by providing the participants an inadvertent frame of reference. From these grounds, a study including test-retest reliability and scales without any number assigned for the scale descriptors should be further implemented. Finally, the questionnaire was administered only in Japanese to Japanese university EFL students, leaving a possibility that, if given in another language or in other settings, we might obtain different results. It is thus recommended that the same type of research be carried out in other languages or in other learning environments. If it is confirmed that the results obtained from the current study can be replicated, we will then be in a better position to argue that using trueness-based descriptors is preferable to frequency descriptors in learning strategy research.

5. Conclusion

The current study attempted to provide empirical evidence showing how different scale descriptors may yield different results in questionnaires of learning strategies. As a result, it came to light that trueness-based scale descriptors elicited slightly higher mean values of responses, which exhibited better fit to the hypothesized model. However, both scale descriptors showed similar patterns of distances from one scale category to another. Based on these findings, we suggest using trueness-based scale descriptors over frequency-based scale descriptors in studies of learning strategies because the former may reflect the hypothesized model more validly.

One important implication of the current research is that what can be measured with the self-reported questionnaire in the studies of learning strategies might not be the frequencies of strategies but rather the underlying *trait* manifested in strategic learning behaviors. This may be one reason why the current study found that trueness-based scale descriptors fit better to the model than frequency-based scale descriptors. At the same time, however, there is still a possibility that, as long as the claims between frequency or agreement or trueness are related to performance in a coherent way—represented by an underlying trait structure, it may not really matter.

Because the research field of language learning strategies has recently seen drastic changes in the theory, the current study could be served as part of a reappraisal of research into language learning strategies. For example, Dörnyei and his colleagues (Dörnyei, 2005; Tseng et al., 2006) have pointed out the problematic nature of definition and measurement in learning strategy research, and they have suggested that we shift the focus of research attention to the self-regulation of the language learner. Also, Macaro (2006) has proposed a theoretical framework which posits “learner strategies occur only in working memory and that they become other constructs elsewhere” (p. 327). These new agendas for future language learning strategies research are theoretically very promising in that it is always the construct of “strategy” that has come under criticism (Cohen & Macaro, 2007). However, before completely shifting our efforts to a new approach, we believe that we should continue to explore the possibilities of improving the conventional methods of learning strategy research. That is to say, if we can demonstrate that such methods do not contain fatal design flaws, it is prudent to make use of findings this research field has produced for over 30 years of practice.

References

- Aiken, L. (1996). *Rating scales and checklists: Evaluating behavior, personality and attitude*. New York: Wiley.
- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.
- Arbuckle, J. L., & Wothke, W. (1999). *Amos 4.0 User's Guide*. SmallWaters, Chicago, IL.
- Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Brown, J. D. (2005). *Testing in Language Programs: A Comprehensive Guide to English Language Assessment* (New ed.). New York: McGraw-Hill.
- Bryant, F., & Yarnold, P. (1995). Principal-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 99–136). Washington, DC: American Psychological Association.
- Cohen, A. D., & Macaro, E. (Eds.). (2007). *Language learning strategies: Thirty years of research and practice*. Oxford: Oxford University Press.
- Cohen, A. D., & Scott, K. (1996). A synthesis of approaches to assessing language learning strategies. In R. L. Oxford (Ed.), *Language learning strategies around the world: Cross-cultural perspectives* (pp. 89–106). Manoa, HI: University of Hawai'i Press, National Foreign Language Center at Manoa.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Dörnyei, Z. (2003). *Questionnaires in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Educational Testing Service. (2006). *Test of English for International Communication (TOEIC)*. Retrieved October 18, 2006, from <http://www.ets.org/toeic>
- Field, A., & Hole, G. (2003). *How to design and report experiments*. London: Sage publications.
- Gu, Y. (2003). Fine brush and freehand: The vocabulary-learning art of two successful Chinese EFL learners. *TESOL Quarterly*, 37, 73–104.
- Gu, Y., Wen, Q., & Wu, D. (1995). How often is often? Reference ambiguities of the Likert-scale in language learning strategy research. *Occasional Papers in English Language Teaching*, 5, 19–35. (ERIC Document Reproduction Service No. ED391358)
- Hagiuda, N., & Shigemasu, K. (1996). Jyunjyo tsuki categorical data eno inshibunseki no tekiouni kansuru ikutsukano chuuiten [Some remarks on the application of

- factor analysis to ordered categorical data]. *Japanese Journal of Psychology*, 67, 1–8.
- Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Boston, MA: Heinle & Heinle.
- Hiromori, T. (2004). Motivation and language learning strategies of EFL high school students: A preliminary study through the use of panel data. *JACET Bulletin*, 39, 31–41.
- Ikeda, M., & Takeuchi, O. (2000). Tasks and strategy use: Empirical implications for questionnaire studies. *JACET Bulletin*, 31, 21–32.
- In'nami, Y. (2006). The effects of test anxiety on listening test performance. *System*, 34, 317–340.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 5–53.
- Linacre, J. M., & Wright, B. D. (2000). *WINSTEPS. Multiple-choice, rating scale, and partial credit Rasch analysis [Computer software]*. Chicago, IL: MESA Press.
- Macaro, E. (2006). Strategies for language learning and for language use: Revising the theoretical framework. *Modern Language Journal*, 90, 320–337.
- McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.
- Oda, K. (1970). Nihongo no teido ryou hyougen yougo ni kansuru kenkyuu. [A psychological study on the Japanese qualitative and quantitative words]. *Japanese Journal of Educational Psychology*, 18, 166–176.
- Oxford, R. L., Cho, Y., Leung, S., & Kim, H-J. (2004). Effect of the presence and difficulty of task on strategy use: An exploratory study. *International Review of Applied Linguistics*, 42, 1–47.
- Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House.
- Purpura, J. E. (1999). *Learner Strategy Use and Performance on Language Tests: A Structural Equation Modeling Approach*. Cambridge: Cambridge University Press.
- Qian, D. D. (2004). Second language lexical inferencing: preferences, perceptions, and practices. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language* (pp. 155–169). Amsterdam: John Benjamins Publishing Company.
- Robinson, P. (2002). *Individual differences and instructed language learning*. Amsterdam/Philadelphia: John Benjamins.
- Sakui, K., & Gaies, S. J. (1999). Investigating Japanese learner's beliefs about language learning. *System*, 27, 473–492.
- Smith, R.M., Schumacker, R. E., & Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement*, 2, 66–78.
- Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology*, 61, 374–375.

- TOEIC Steering Committee. (2006). *TOEIC TEST 2005 Data & Analysis*. Retrieved July 1, 2006, from <http://www.toeic.or.jp/toeic/data/pdf/DAA2005.pdf>
- Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics* (5th international ed.). Boston, MA: Pearson/Allyn & Bacon.
- Takeuchi, O. (2003). What can we learn from good language learners?: A qualitative study in the Japanese foreign language context. *System*, *31*, 313–432.
- Tseng, W. T., Dörnyei, Z., & Schmitt, N. (2006). A new approach to assessing strategic learning: The case of self-regulation in vocabulary acquisition. *Applied Linguistics*, *27*, 78–102.
- Vandergrift, L. (2005). Relationships among motivation orientations, metacognitive awareness and proficiency in L2 listening. *Applied Linguistics*, *26*, 70–89.
- Wakita, T. (2004). Hyoutei shakudo ho ni okeru category kan no kankaku ni tsuite [Assessment of the distance between categories in rating scales by using item response model]. *Japanese Journal of Psychology*, *75*, 331–338.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. Retrieved September 2, 2009, from <http://www.rasch.org/rmt/rmt83b.htm>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Appendix A

Questionnaire on Language Learning Strategies (Originally in Japanese; adopted from Hiromori 2004)

(Frequency)

1. I never do it
2. I rarely do it
3. I sometimes do it
4. I often do it
5. I almost always do it

(Trueness)

1. Not at all true of me
2. Not true of me
3. Somewhat true of me
4. True of me
5. Very true of me

Cognitive strategies

1. I say or write new English words several times.
2. I watch TV programs or movies in English.
3. I look for Japanese words that are similar to new words in English.
4. I try to connect what I am learning with what I already know.
5. I look for examples in which new words or phrases in English are used.
6. When I am learning new material in English, I translate it into Japanese.
7. When I am learning new material in English, I try to somehow organize it in my mind.
8. I look for differences between English and Japanese.
9. I try to use new words, phrases, or basic sentences in new situations.
10. I ask other people to check if I correctly understand what I have newly learned.

Metacognitive strategies

11. Before studying English, I plan what I am going to do in order that I can use my time well.
12. I try to think of ways to improve my English skills and to prepare materials and a learning environment for my study.
13. I think about my progress in learning English on a regular basis.
14. I try to have clear goals for improving my English skills.
15. I try to learn English from the mistakes I make.
16. When I learn something new, I make sure if I have learned it completely by using it.
17. I think about the best ways for me to learn English.
18. After I take a test, I think of the ways to get a higher score next time.
19. I notice my mistakes in grammar or usage when speaking English.
20. I consciously listen to the parts I do not understand when listening to English.

Appendix B

(a) Descriptive Statistics of the Frequency-based Descriptors ($N = 408$)

Scale	Item No.	M	SD	Skewness	Kurtosis	Difficulty Estimate	Infit Mean Square	Outfit Mean Square
Cognitive strategies	01	3.13	1.05	0.02	-0.65	-0.61	1.10	1.12
	02	2.97	1.04	-0.17	-0.55	-0.37	1.31	1.36
	03	2.42	0.96	0.49	-0.15	0.40	0.94	0.94
	04	2.88	0.89	0.02	-0.51	-0.28	0.70	0.69
	05	2.41	0.95	0.56	-0.01	0.42	0.90	0.91
	06	3.50	0.99	-0.49	-0.20	-1.14	1.20	1.24
	07	3.22	0.91	0.07	-0.66	-0.76	0.80	0.81
	08	2.32	0.94	0.75	0.40	0.56	1.04	1.03
	09	2.30	0.88	0.67	0.56	0.58	0.77	0.76
	10	1.95	0.93	0.92	0.54	1.19	1.26	1.23
Metacognitive strategies	11	2.52	0.99	0.44	-0.23	0.54	1.00	0.98
	12	2.59	0.92	0.40	-0.06	0.41	0.84	0.91
	13	2.07	0.85	0.66	0.21	1.47	0.92	0.88
	14	2.87	1.06	0.21	-0.68	-0.10	1.08	1.08
	15	3.07	0.92	-0.04	-0.25	-0.43	0.98	0.99
	16	2.42	0.79	0.47	0.38	0.74	0.77	0.77
	17	3.31	0.96	-0.19	-0.37	-0.86	1.00	0.99
	18	3.17	0.95	-0.02	-0.30	-0.63	0.95	0.96
	19	2.83	1.04	0.16	-0.52	-0.04	1.25	1.24
	20	3.46	0.97	-0.30	-0.26	-1.11	1.15	1.20

(b) Descriptive Statistics of the Trueness-based Descriptors ($N = 408$)

Scale	Item	M	SD	Skewness	Kurtosis	Difficulty Estimate	Infit Mean Square	Outfit Mean Square
Cognitive strategies	01	3.64	1.02	-0.29	-0.71	-0.82	1.22	1.23
	02	3.23	1.19	-0.17	-0.85	-0.4	1.47	1.49
	03	2.79	0.93	0.06	-0.24	0.31	0.94	0.94
	04	3.20	0.83	0.08	-0.25	-0.33	0.74	0.74
	05	2.75	0.93	0.45	-0.24	0.35	0.85	0.86
	06	3.43	0.88	-0.17	-0.15	-0.94	1.14	1.18
	07	3.13	0.81	0.12	-0.35	-0.53	0.76	0.77
	08	2.56	0.89	0.38	-0.01	0.57	0.96	0.95
	09	2.67	0.87	0.43	0.01	0.48	0.73	0.72
	10	2.06	0.88	0.65	0.17	1.31	1.18	1.16
Metacognitive strategies	11	2.69	0.89	0.49	-0.07	0.46	1.08	1.07
	12	2.87	0.93	0.30	-0.26	0.1	0.88	0.88
	13	2.17	0.76	0.56	0.48	1.63	0.93	0.93
	14	3.06	0.92	0.20	-0.46	-0.28	0.9	0.91
	15	3.12	0.89	0.12	-0.04	-0.38	1.06	1.06
	16	2.75	0.85	0.53	0.16	0.34	0.97	0.95
	17	3.36	0.86	0.19	-0.39	-0.84	0.85	0.85
	18	3.17	0.88	0.01	-0.07	-0.48	1.04	1.02
	19	2.78	0.90	0.13	-0.24	0.28	1.11	1.11
	20	3.35	0.95	-0.05	-0.52	-0.82	1.18	1.19