Analyzing and Interpreting Rating Scale Data from Questionnaires

Richard Watson Todd
King Mongkut's University of Technology Thonburi

Abstract

Questionnaires comprising Likert rating scale items are probably the most commonly used instrument for data collection in education and educational research, yet there is much less guidance available concerning how to analyze and interpret such rating scales than there is on how to design them. Likert scale questionnaire data is most commonly analyzed using means, although this assumes that the Likert scale descriptors are equally spaced points on a continuum. Alternative ways to analyze rating scale data include percentage frequencies, medians and modes. If means are used, there are several different approaches to interpretation depending on how concrete, intuitive or unbiased the interpretations should be. This paper presents the options available in analyzing and interpreting Likert scale data and shows how the choice between options depends on the purposes of the research.

1. Introduction

One of the most commonly used instrument for data collection in education and educational research is the questionnaire (Gass and Mackey, 2007), and probably the most commonly used question format on questionnaires is the Likert rating scale. Likert scales consist of a statement or question to which respondents are asked to select a response (usually from 4, 5 or 6 choices) that best matches their opinion. For example:

	Disagree **1**	**Neutral** **2**	**Agree** **3**	**Agree strongly** **4**
rEFLections is a valuable resource for English language teachers				

Likert scale questionnaires have a wide range of applications. They are frequently used in needs analyses (e.g. Kormos, Kontra and Csölle, 2002), course evaluations,

students' evaluations of teachers (e.g. Block, 1998), self-assessments (e.g. Brantmeier, 2006), and research. In English language teaching research, Likert scales have been used to collect data on a wide range of issues including use of learning strategies (Grainger, 2005; Griffiths, 2003; Hsiao and Oxford, 2002; Sheorey and Mokhtari, 2001), attitudes towards language (e.g. Barkhuizen, 2002) and towards learning activities (e.g. North and Pillay, 2002), student beliefs (e.g. Diab, 2006), and students' reactions to educational innovations (e.g. Yoon and Hirvela, 2004).

With such high levels of use and such a wide diversity of potential applications, it might be expected that the literature contains extensive guidance on designing, using, analyzing and reporting Likert scale items. While there is very useful advice on design and administration of Likert questionnaires (e.g. Brown, 2001; Dörnyei, 2003), the issues of analyzing and reporting Likert-based data collection are overlooked in most of the literature. In this paper, I intend to examine and discuss how to analyze and report data collected through Likert scales.

2. Likert scale descriptors

A key issue in analyzing Likert scale data is the nature of the descriptors used on the scale, so before we can investigate how the results from such scales can be analyzed, we need to examine the types of descriptors used. In the example above, the descriptors concern levels of agreement with the prompting statement, and this type of scale is very common (e.g. Barkhuizen, 2002; Diab, 2006; Yoon and Hirvela, 2004). Also common are descriptors of frequency (e.g. both Oxford's (1990) Strategy Inventory for Language Learning and the Survey of Reading Strategies focus on frequency of use of strategies), although there are two ways of indicating frequency. Frequency can be described through adverbs (e.g. *never*, *occasionally*, *sometimes*, *frequently*, *very frequently*) or through amount of use in a given time period (e.g. *never*, *once or a few times a year*, *once or twice a month*, *once or twice a week*, *on a daily basis*), and these may also be combined (the examples given here both come from Kormos, Kontra and Csölle (2002) who define a '5' on their scale as *very frequently, on a daily basis*). While agreement and frequency are the most commonly used bases for Likert scales, descriptors for nearly any adjective, such as from *not at all interesting* to *very interesting*, or from *very easy* to *very difficult*, can be created easily. Most Likert questionnaires use the same descriptors for a string of prompts, but it is also possible, though more laborious in both design and responding, to write specific descriptors for each question. For instance, Brantmeier's (2006) questionnaire for self-assessment of reading asks respondents

to rate themselves as a reader of Spanish where the 5 choices range from *I am not a good reader of Spanish* to *I am an excellent reader of Spanish*.

3. Statistics for analyzing Likert scale data

The nature of the descriptor used in a Likert scale is crucial in deciding what statistics to use in analyzing the data. In this paper, I will consider only descriptive statistics that provide a summary of the data. The most basic way (and a prerequisite for other analyses) of presenting Likert data is to provide a table showing the frequency of responses. For instance, let us suppose that the question at the beginning of this article had been given to 20 respondents who rated the prompting statement at the frequencies shown in Table 1.

Table 1: Frequencies of response to a Likert scale item

	Disagree **1**	**Neutral** **2**	**Agree** **3**	**Agree strongly** **4**
rEFLections is a valuable resource for English language teachers	0	2	10	8

Table 1 shows the raw data, and, in this case, the table is fairly clear and easy to understand. Where there are 20 or more prompts and a much larger sample size (especially where different total numbers of respondents answered different questions), tables of frequency of response can be very difficult to interpret. These issues can be partially overcome by converting the raw frequencies into percentages, as shown in Table 2.

Table 2: Percentages of frequencies of response to a Likert scale item

	Disagree **1**	**Neutral** **2**	**Agree** **3**	**Agree strongly** **4**
rEFLections is a valuable resource for English language teachers	0%	10%	50%	40%

Even with the clearer percentages, however, it is easy to drown in the volume of data for questionnaires with large numbers of items, and it is difficult both to see patterns emerging from the questionnaire responses and to compare the responses

to different questions. A single number summarizing the responses for each question would overcome these problems.

The most obvious single number to use is the mean, which for the data in Table 1 is 3.30. For a questionnaire with many questions, listing the mean ratings for each question provides a clearer picture of responses (especially if we also give the standard deviation to show how widely spread out the responses are). However, using a mean for summarizing data from a Likert scale assumes that the four discrete categories in the scale form a continuum based on an underlying continuous variable (Clason and Dormody, 1994). After all, there is no category for 3.30 making it difficult to know what 3.30 means, so we must assume that a mean of 3.30 indicates a level of agreement lying somewhere between *Agree* and *Strongly agree*. A key problem here is that "averaging rating scale responses is potentially hazardous, as it requires the assumption that the intervals between points on the rating scale are equal" (Brown and Daniel, 1990, p. 7). For the scale in our example, is *neutral* the same semantic distance away from *agree* as *agree* is from *strongly agree*? If we believe that it is, we can use a mean as a summary statistic; if not, we should avoid the mean. For other categories, we might ask whether *occasionally* is the same interval away from *sometimes* as *sometimes* is from *frequently*. While we may think that these adverbs are at equal intervals on a continuum, the second set of frequency descriptors used by Kormos, Kontra and Csölle (2002) quoted above, namely, *once or a few times a year*, *once or twice a month*, and *once or twice a week*, which they take as being equivalent to the adverbs of frequency, are not equidistant. *Once a year* is twelve times less frequent than *once a month*, which is only four times less frequent than *once a week*. For this scale, the unequal intervals between ratings mean that we should avoid using the mean (in fact, the full scale here is almost a logarithmic scale with possible responses being once every 300 days, every 30 days and every 3 days, in which case we should use the geometric mean, i.e. the nth root of the product, rather than the more familiar arithmetic mean).

If we cannot use a mean value but still wish to summarize the data into a single number, we still have two further options: the median and the mode. The median is the middle number when all responses are sequenced from lowest to highest. In the data from Table 1, with an even number of respondents, the middle value is the value between the 10th and the 11th respondents, in this case, 3 or *agree*, since both these respondents answered 3. The mode is the most frequent response category – for our data, again this is 3 or *agree* with 10 respondents choosing this. A

key advantage with using the median and the mode is that, generally, the number summarising the data is a whole number matching one of the categories (although there is a small chance that, with an even number of respondents, the median might be midway between two categories), and is therefore straightforward to interpret.

Generally, the default statistic for summarizing Likert scale data appears to be the mean. Of the studies cited above, 10 out of 11 report the mean in presenting the data (only Barkhuizen, 2002 reports percentage frequencies but no means), including studies such as Kormos, Kontra and Csölle (2002) where assuming a continuous underlying variable is dubious. In analyzing Likert scale data, rather than automatically using means to summarize data, researchers, teachers and administrators need to consider the nature of the data collected before making a decision on how to deal with the data.

4. Interpreting Likert scale data

Summarizing the data is generally only the first step in data analysis. The findings then need to be interpreted to give them meaning. If percentage frequencies are used to summarize the data, often the highest percentage (i.e. the mode) is then highlighted for interpreting the meaning of the findings. The category descriptor for the rating where the mode is located (and similarly for the median) then becomes the basis for interpreting the data. Thus, the interpretation for the data in Table 2 is that most commonly the respondents agree that *rEFLections* is a valuable resource. While straightforward and uncontroversial, this approach may overlook key issues in the data (for instance, we might have collected data where 11 of the respondents strongly agree with the statement and 9 disagree; in this case a summary stating that respondents generally strongly agree appears to obscure key aspects of the findings) and does not allow easy comparison between the responses to different questions.

Using the mean overcomes these shortcomings (although the issues of whether the mean is valid discussed above are more important), and also allows several different approaches to interpretation. In the example we have been using in this paper, assuming that the scale from *disagree* to *strongly agree* consists of equal intervals, we find a mean of 3.30. How should this be interpreted?

The most straightforward approach is simply to present the mean value without making any attempt to link it with a descriptor on the rating scale. Stating that the mean is 3.30 without interpreting this avoids the need to decide whether 3.30

should be categorized as *agree* or *strongly agree* (and thus really treats the rating scale as a continuum), while still allowing easy comparison between the summarizing statistics for different questions on a questionnaire. This may be the most appropriate approach in formal research (where readers might be expected to tolerate the ambiguity of the meaning of 3.30), but in many other situations where rating scales are used, a clearer interpretation is expected.

Often needs analyses, course evaluations and teacher evaluations using rating scales are conducted with the end audience being administrators, and busy administrators often expect clear unambiguous interpretations, preferring a conclusion of, say, *agree* to an ambiguous mean of 3.30. Similarly, students using rating scales for self-assessment may prefer clear conclusions. For instance, in the presentation of the Strategy Inventory for Language Learning in Oxford (1990) the questionnaire is designed for self-administration by students. To help students gain practical benefits from completing the questionnaire, they are asked to calculate means for different groups of strategies, and meanings are assigned to various ranges of mean values (for instance, mean values from 2.5 to 3.4 are categorized as "sometimes used" and interpreted as "medium" use of the strategy group (p. 291)). In this way, abstract number values are made concrete to enable students to understand the meaning of their self-ratings.

There are two main ways of converting mean values into concrete interpretations reflecting the descriptions in the Likert scale. The most intuitive method is to categorize mean values into the descriptor for the number on the rating scale closest to that value. Using this approach, the mean of 3.30 would be interpreted as *agree*, since 3.30 is closer to 3 than to 4. Overall, mean values from 1.00 to 1.50 would be categorized as *disagree*, 1.51 to 2.50 as *neutral*, 2.51 to 3.50 as *agree*, and 3.51 to 4.00 as *strongly agree*. This approach is used in interpreting the Strategy Inventory for Language Learning and in several other research articles (e.g. Grainger, 2005). Although intuitively satisfactory, this approach is biased against end-values. The likelihood of a mean value being interpreted as *strongly agree* is only half of that for *agree*, since the 'length' of the *strongly agree* category is only 0.50, while the 'length' for *agree* is 1.00. Together with a common aversion by respondents to extreme categories, this bias means that this approach to interpreting mean values often results in over-interpretation of central points on a rating scale.

An approach which overcomes the problems of biased categories is to ensure that the 'length' of each category is equal. In our example, the overall 'length' of the

rating scale is 3 (4 minus 1) and there are 4 possible ratings. To generate categories of equal 'length', each category should cover a 'length' of 0.75 (3 divided by 4). Thus 1.00 to 1.75 would be interpreted as *disagree*, 1.76 to 2.50 as *neutral*, 2.51 to 3.25 as *agree*, and 3.26 to 4.00 as *strongly agree*. Our mean value of 3.30 would therefore be identified as *strongly agree* under this approach. Although none of the research studies cited above interpret mean values following this approach, it is fairly widely used. For example, Australian government departments recommend its use in interpreting rating scale data. While avoiding biases, this approach can seem counter-intuitive, since some mean values, such as our mean of 3.30, are not categorized into the rating descriptor closest to the value.

Since all of the options available in interpreting rating scale data have advantages and disadvantages, there is no single method of interpretation that fits all studies. Rather, the researcher's choice depends on the answers to questions such as "Is it important that a concrete meaning be assigned to a mean value?" and "Is intuitive understanding of the results more important than avoiding biases, or vice versa?" The answers, in turn, depend on the purposes of the research and how the research findings will be used. For instance, if the target audience is students (as in self-administered strategy use questionnaires) or administrators (as in many teacher evaluations), interpreting the data as descriptors is probably preferable to providing numbers as summaries; and if, in a research study summarizing data as means, it is seen as important that potential biases against interpretation of extreme categories be avoided, either raw uninterpreted means or the second way of interpreting means as categories should be used.

5. Conclusion

While questionnaires employing Likert rating scales are common, perhaps not enough thought is given to their analysis and interpretation. In this paper, I hope that I have shown that there is a wide range of choices available in analyzing and interpreting Likert scale data and that the nature of the Likert scale and how the results will be used should influence which choice is most appropriate. As with most issues in research, some careful consideration of the nature and purposes of questionnaire data collection and analysis leads to stronger and more persuasive results.

References:

Barkhuizen, G. P. (2002). Language-in-education policy: students' perceptions of the status and role of Xhosa and English. *System, 30*(4), 499-516.

Block, D. (1998). Exploring interpretations of questionnaire items. *System, 26*(3), 403-425.

Brantmeier, C. (2006). Advanced L2 learners and reading placement: Self-assessment, CBT, and subsequent performance. *System, 34*(1), 15-35.

Brown, J. D. (2001). *Using surveys in language programs*. Cambridge: Cambridge University Press.

Brown, T. C. & Daniel, T. C. (1990). *Scaling of ratings: Concepts and methods* (USDA Forest Service Research Paper RM-293). Fort Collins, CO: USDA Forest Service.

Clason, D. L. & Dormody, T. J. (1994). Analyzing data measured by individual Likert-type items. *Journal of Agricultural Education 35*(4), 31-35.

Diab, R. L. (2006). University students' beliefs about learning English and French in Lebanon. *System, 34*(1), 80-96.

Dörnyei, Z. (2003). *Questionnaires in second language research*. Mahwah, NJ: Lawrence Erlbaum.

Gass, S. M. and Mackey, A. (2007) *Data Elicitation for Second and Foreign Language Research*. London: Routledge.

Grainger, P. (2005). Second language learning strategies and Japanese: Does orthography make a difference? *System, 33*(2), 327-340.

Griffiths, C. (2003). Patterns of language learning strategy use. *System, 31*(3), 367-384.

Hsiao, T.-Y. & Oxford, R. L. (2002). Comparing theories of language learning strategies: A confirmatory factor analysis. *The Modern Language Journal, 86*(3), 368-383.

Kormos, J., Kontra, E. H. & Csölle, A. (2002). Language wants of English majors in a non-native context. *System, 30*(4), 517-542.

North, S. & Pillay, H. (2002). Homework: Re-examining the routine. *ELT Journal, 56*(2), 137-145.

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. Boston: Heinle & Heinle.

Sheorey, R. & Mokhtari, K. (2001). Differences in the metacognitive awareness of reading strategies among native and non-native readers. *System, 29*(4), 431-450.

Yoon, H. & Hirvela, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing, 13*(4), 257-284.

Biodata:
Richard Watson Todd is Associate Professor in Applied Linguistics at KMUTT. He holds a PhD from the University of Liverpool and is the author of numerous articles and books, most recently, *Much Ado about English* (Nicholas Brealey Publishing).
E-mail: irictodd@kmutt.ac.th