Developing Likert-Scale Questionnaires

Tomoko Nemoto
Temple University, Japan Campus

David Beglar
Temple University, Japan Campus

Reference Data

Nemoto, T., & Beglar, D. (2014). Developing Likert-scale questionnaires. In N. Sonda & A. Krause (Eds.), *JALT2013 Conference Proceedings*. Tokyo: JALT.

Likert-scale questionnaires are the most commonly used type of instrument for measuring affective variables such as motivation and self-efficacy, given that they allow researchers to gather large amounts of data with relative ease. However, despite their widespread use, there is relatively little information in the second language literature concerning the development and analysis of Likert-scale questionnaires. The purpose of this paper is to present a set of five guidelines for constructing Likert-scale instruments: understanding the construct, developing items, determining the outcome space, specifying the measurement model, and gathering feedback and piloting the questionnaire. The use of these guidelines can lead to the development of Likert-scale questionnaires that are more likely to yield reliable data that lead to valid interpretations.

リッカート尺度によるアンケートは、比較的容易に極めて多くのデータ収集を可能とするため、動機づけや自己効力感のような情意変数の測定としてもっとも一般的に利用されている手段のひとつである。しかしこうした状況下においても、リッカート尺度によるアンケートの開発や分析に関しての情報は第二言語習得学分野では比較的限られている。この論文の目的は、リッカート尺度による測定手段構築のための5つのガイドライン（構成概念の理解、アンケート項目の作成、標本結果空間の決定、測定モデルの特定、そしてフィードバックの収集とアンケートの試験調査の実施）を提示することである。これらのガイドラインを用いることが、より信頼性の高い、妥当なデータを生み出すリッカート尺度によるアンケート作成へとつながる。

Educational research has three primary purposes. The first is to contribute to the development of theory, the second is to investigate phenomena believed to play an important role in the educational process, and the third is to develop more effective pedagogy. Attempts to pursue any of these purposes require the use of data-gathering tools, and these tools take diverse forms, such as tests, interview protocols, classroom observations, and questionnaires. Regardless of the tool used, the questions that researchers hope to answer nearly always concern abstract issues that are not directly observable. For this reason, ensuring that the data gathered can be used to make particular inferences is extremely important, as this is at the heart of modern conceptions of validity. Messick (1989) stated that validity is "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores and other modes of assessment" (p. 13). Thus, it is essential that any interpretation of a test score must be defended using both theory and empirical evidence. The primary purpose of this paper is to present guidelines for developing Likert-scale questionnaires. We believe that

following these guidelines can strengthen the validity argument that can be made for data gathered using these questionnaires.

A Likert scale is a psychometric scale that has multiple categories from which respondents choose to indicate their opinions, attitudes, or feelings about a particular issue. In the field of SLA, Likert-scale questionnaires have most frequently been used in investigations of individual difference variables, such as motivation, anxiety, and self-confidence. Some advantages of Likert-scale questionnaires are that (a) data can be gathered relatively quickly from large numbers of respondents, (b) they can provide highly reliable person ability estimates, (c) the validity of the interpretations made from the data they provide can be established through a variety of means, and (d) the data they provide can be profitably compared, contrasted, and combined with qualitative data-gathering techniques, such as open-ended questions, participant observation, and interviews.

In this paper, we divide the Likert-scale development process into five main sections, primarily based on Wilson's (2005) approach to psychological measurement. The five sections are (a) understanding the construct, (b) developing the items, (c) determining the outcome space, (c) specifying the measurement model, and (e) gathering feedback and piloting the questionnaire.

Understanding the Construct

Likert-scale instruments are most frequently used to measure psychological constructs (see Messick, 1989, for a detailed discussion of the notion of construct), which is one aspect of a person's affect or cognition that can be operationalized and measured. Constructs in the field of SLA are typically linguistic (e.g., syntactic knowledge), affective (e.g., listening anxiety), or personality based (e.g., extraversion), and they are conceptualized as extending from one extreme to another—low to high, small to large, negative to positive, or weak to strong. In other words, they form a continuum. Regardless of the type of construct being measured, the starting point for questionnaire development is to arrive at a thorough understanding of the target construct, primarily by reading academic literature on the topic. This reading should be focused on both understanding the theory associated with the construct and on analyzing items from previous questionnaires designed to measure that construct. In addition to reading, it is useful to engage in a critical discussion of the content of the reading with persons also familiar with the construct, as this strategy can result in a more well-developed, accurate understanding of the construct.

Beginning the measurement process by focusing on a psychological construct is a part of what has been termed a *construct-centered approach* (Messick, 1994). In the construct-centered approach, we begin "by asking what complex of knowledge, skills, or other attributes should be assessed" (p. 17) and then we consider "what behaviors or performances should reveal those constructs, and what tasks or situations should elicit those behaviors. Thus, the nature of the construct guides the selection or construction of relevant tasks" (p. 17). These quotations emphasize the close relationship between the hypothesized construct and the tasks or items on an instrument used to provide evidence of the degree to which the construct is present in each individual. They show that item development must begin with a well-developed understanding of what it is we wish to measure.

Developing the Items

After gaining an understanding of the hypothesized construct, it is possible to consider item design. Items are the concrete realization of the abstract construct, and as such, represent the theoretical understanding of the construct. Each item should be designed to measure a specific aspect of the construct. There are two main advantages to this approach. First, item design is no

longer arbitrary because it is based on a detailed understanding of the construct. For this reason, data gathered using the items can be related back to the construct with greater confidence. Second, the statistical results that flow from the data produced by the items can be used to improve our understanding of the construct. For instance, statistical analyses can show which items adequately measure the construct and which appear to measure a different construct. The use of this sort of information can be likened to a conversation between theory and item performance. The theory initially informs item development, but then the data gathered using those items inform further theory development, as they potentially indicate where the theory is, and is not, supported.

Each item should be designed to measure one idea and should be written in straightforward, easy-to-understand language so that the meaning of the item is unambiguous to respondents (Wolfe & Smith, 2007). For instance, high-frequency, nontechnical vocabulary should be used, and complex grammatical constructions should be avoided. Moreover, conjunctions, such as *and*, *or*, and *but* should not be used, as they generally indicate the presence of two ideas (i.e., a so-called "double-barreled" question). The problem with doubled-barreled questions is that they invite respondents literally to answer different questions. For instance, if the item says *I can understand written and spoken academic texts*, some respondents might respond to the word *written*, but others might respond to the word *spoken*. If the item writer wishes to include both ideas on the questionnaire, they should be presented as separate items (i.e., *I can understand written academic texts* and *I can understand spoken academic texts*).

Another issue to consider in item development is the number of items needed and the difficulties of those items. These are important considerations because achieving a sufficiently high level of reliability and measurement precision depends primarily on these two points. First, six to eight good-performing items are generally sufficient for measuring a single construct reliably, but this means that it is best to initially write and pilot 10-12 items in order to be able to select the best performing items. Second, the items need to differ in terms of their difficulty (in a Likert-scale questionnaire, difficulty is often termed *endorseability*) so that the scale's entire response range (e.g., 1-6) is used. This can be accomplished at the item development stage by dividing the items into at least three groups (e.g., easy to endorse, moderately difficult to endorse, and difficult to endorse) and making sure that the number of items in each group is similar (e.g., four items in each group). Items produced by previous researchers can be helpful at this stage, but they should be analyzed critically before adopting or adapting them to ensure that they are well aligned with the understanding of the target construct.

Two further issues should be considered when constructing items. First, positively and negatively worded items should not be used to measure a single construct, as this approach negatively affects *unidimensionality* (Quilty, Oakman, & Risko, 2006; Yamaguchi, 1997). Unidimensionality, which is the idea that a set of items measures a single construct, is important because it is difficult to interpret the results of items measuring multiple constructs. Items should be written in a single direction with a preference for positively worded items (Wolfe & Smith, 2007). Second, when possible, the items should be written in a language the respondents understand well or that is their native language. This is in order to reduce the contamination of construct irrelevant variance, which is produced by contextual variables that affect the responses made by the respondents (e.g., excessive heat, loud noises, or in this case, a poor understanding of the meaning of the item). The goal should be to produce items that the respondents immediately and accurately comprehend.

Example: Writing Items to Measure Classroom Speaking Self-Confidence

In this concrete example, we wish to measure a construct we call *Classroom Speaking Self-Confidence*. Notice how (a) each item concerns a single idea, (b) each item is worded in a straightforward way, (c) a sufficient number of items (i.e., 12) has been written for this first draft, and (d) all items are worded positively.

Let us say that previous research has indicated that confidence is influenced by (a) the number of interlocutors, (b) the identity of the interlocutor(s), (c) the conversation topic, and (d) the length of the interaction. This results in the following hypotheses:

- Speaking self-confidence is higher when speaking with fewer rather than more interlocutors.
- Speaking self-confidence is higher when speaking with familiar rather than unfamiliar persons.
- Speaking self-confidence is higher when speaking about common, everyday topics rather than academic or technical topics.
- Speaking self-confidence is higher when engaging in shorter rather than longer speaking tasks.

With these hypotheses in mind, we can divide the following items into three hypothesized difficulty groupings. It is also important to note that it is ideal if the students have experience performing the tasks described in the items. This is preferable to asking the students to rely only on their imaginations to determine their degree of confidence.

Group 1: Difficult for Respondents to Endorse

- I can discuss an academic topic (e.g., an environmental issue) for 30 minutes with three to four other students.
- I can discuss an academic topic (e.g., an environmental issue) for 15 minutes with my teacher.
- I can discuss an academic topic (e.g., an environmental issues) for 15 minutes with a classmate.
- I can make a 10-minute presentation on an academic topic (e.g., an environmental issue) to the entire class.

Group 2: Moderately Difficult for Respondents to Endorse

- I can discuss an academic topic (e.g., an environmental issue) for 15 minutes with three to four other students.
- I can discuss an academic topic (e.g., an environmental issue) for 10 minutes with my teacher.
- I can discuss a common topic (e.g., summer vacation) for 10 minutes with a classmate.
- I can make a 5-minute presentation on a common topic (e.g., summer vacation) to the entire class.

Group 3: Easy for Respondents to Endorse

- I can discuss a common topic (e.g., summer vacation) for 10 minutes with three to four other students.
- I can discuss a common topic (e.g., summer vacation) for 5 minutes with my teacher.
- I can discuss a common topic (e.g., summer vacation) for 5 minutes with a classmate.
- I can make a 3-minute presentation on a common topic (e.g., summer vacation) to three to four other students.

Determining the Outcome Space

The items are only one part of a Likert-scale questionnaire. An outcome space, which concerns how responses to the items are

categorized and scored, is also needed. The outcome space for Likert scales is made up of a limited range of possible responses on continua such as *Disagree/Agree*, *I am not like this/I am like this*, *I am not willing/I am willing*, or *Not useful/Useful*. Most Likert scales should be made up of four or six points. Analyses have shown that scales with more than six categories are rarely tenable, possibly because of limitations in working memory capacity (see Smith, Wakely, Kruif, & Swartz, 2003, for a detailed study concerning this issue). Four points are desirable for young respondents and for respondents with low motivation to complete the questionnaire because 4-point scales are easy to understand and they require less effort to answer. When possible, however, 6-point scales should be used as they permit the possibility of increased measurement precision. See Figure 1 for an example.

1	2	3	4	5	6
Strongly disagree	Disagree	Slightly disagree	Slightly agree	Agree	Strongly agree

Figure 1. A sample 6-point Likert scale for items requiring agreement or disagreement.

Figure 1 displays a number of desirable characteristics for a Likert scale. First, the scale moves from a weaker endorsement (meaning less agreement with or approval) of the item (i.e., *Strongly disagree*) to a stronger endorsement of the item (i.e., *Strongly agree*). In this sense, the scale mimics the best form of measurement available, the ratio scale that exists with physical measurement systems such as the metric system. All physical measurement scales run from smaller to larger amounts of the construct (e.g., 1 cm. → 2 cm., → 3 cm.), and that is also the approach that should be taken with psychological measurement.

A second desirable characteristic is that the scale has no *Neutral* or middle category. Neutral categories should not be used for three reasons. First, as noted above, Likert-scale categories should be conceptualized in the same way as physical measurement. If we look at a ruler, we find that there is no neutral category (i.e., no point on the ruler is labeled "no length" or "neutral length"). Second, middle categories cause statistical problems in that analyses of rating scales often show that neutral categories disturb measurement in the sense that they do not fit statistical models well or they are disordered. For example, *Neutral* is designed to be more difficult to endorse than *Disagree*, but statistical analyses of rating scales using Rasch software such as Winsteps (Linacre, 2013) sometimes show that it is easier to endorse. This sort of finding is reasonable given that neutral categories are inherently illogical in that they do not conform to the fundamental continuum of the scale (i.e., neutral ≠ (dis)agreement). Third, a neutral category is unnecessary because researchers should only include items on a questionnaire that respondents can answer, and this should be confirmed through piloting. In the rare event that some respondents cannot respond to some items, they should not answer the item, as reasonable amounts of missing data present no problems for modern approaches to psychological measurement (see Wolfe & Smith, 2007, for a discussion of how neutral categories produce construct-irrelevant variance).

Specifying the Measurement Model

The measurement model, which is also known as a psychometric model or interpretational model (National Research Council, 2001), allows researchers to evaluate and interpret the responses to the items. The first commonly used formal measurement model is the true score model of classical test theory. The true score model is based on the following equation: Observed score = True score + Error. The second commonly used measurement model is the family of item response models, one of which is the Rasch model (Rasch, 1960; see Bond & Fox, 2007 for an excellent

introduction to the most commonly used Rasch models). Two Rasch models are available for analyzing Likert-scale data: the partial-credit model (Masters, 1982) and the rating scale model (Andrich, 1978), both of which are part of the Winsteps software package (Linacre, 2013). Researchers trained in educational measurement moved to using latent trait models (i.e., statistical models that relate items to the construct they are designed to measure), such as the Rasch model, decades ago because of the many advantages they provide over classical test theory.

There are a number of important reasons for using a formal measurement model, such as one of the Rasch models, rather than using the raw scores provided by the Likert scale:

- First, it is possible that the scale categories have not performed as intended. Three common problems can be identified through the use of Rasch software: (a) few respondents chose a particular category (e.g., *Strongly disagree*), so the category is statistically unstable; (b) two categories (e.g., *Strongly disagree* and *Disagree*) were conceptualized in nearly the same way by the respondents, so they are not clearly separated and are best seen as one category; or (c) two categories are misordered (e.g., *Agree* is more difficult to endorse than *Strongly Agree*). All of these problems can be detected and dealt with using an appropriate Rasch model.
- Second, Rasch models provide statistical indices that indicate the degree to which persons and items fit the probabilistic predictions made by the model. Misfitting items can be deleted from the analysis, or, if they are detected during the pilot phase, revised or replaced.
- Third, the dimensionality of the items can be investigated using the Rasch principle components analysis of item residuals, which shows items that appear to measure something different from the targeted construct. In more extreme cases, the analysis might show that the hypothesized construct has divided into two or more parts (e.g., *Speaking Self-Confidence* divides into *Speaking Self-Confidence Inside the Classroom* and *Speaking Self-Confidence Outside the Classroom*).
- Fourth, the relationship between person ability and item difficulty can be determined. This relationship is important because the best measurement occurs when the range of item difficulties matches the range of person ability. If the bulk of the items are either too easy or too difficult for the respondents to endorse, measurement precision is degraded.
- Finally, the Rasch model produces interval measures from ordinal Likert-scale data. This is a key issue in educational and psychological measurement, as it means that the distances among different points on the scale are equal (after taking measurement error into account). Interval measures are far superior to the ordinal data produced by the Likert scale.

Gathering Feedback and Piloting the Questionnaire

Although the above four steps in the questionnaire development process are extremely important, they are of limited effectiveness without piloting because the actual performance of the items is unknown until they are piloted. Piloting involves gathering both quantitative and qualitative feedback about the construct and questionnaire items at multiple points in the development process:

- First, as noted above, persons developing the questionnaire should consult with others who are familiar with the construct in order to refine their understanding of the construct.
- Second, once a first draft of the items is written, it should be shown to at least three reviewers, who should independently review the items for their relationship to the construct, clarity of expression, and probable difficulty.

- After revisions have been made based on the reviewers' feedback, the items should be shown to the reviewers once again for further feedback. This process should continue until there is general agreement that the items are well written.
- If the questionnaire items must be translated, feedback must be gathered again after translation, given that arriving at an accurate translation is a complex and crucial part of questionnaire development. One or two qualified, bilingual persons (e.g., professional translators or bilingual teachers) should produce the initial translation. The translated items should then be reviewed by at least three other bilinguals, and their feedback should be communicated to the original translator(s) who should make the necessary revisions. This process should continue until there is general agreement that the translations are accurate and easily comprehensible. A back translation can also be produced to ensure that the original meaning has not been altered during the translation process.
- At this point, the questionnaire items should be shown to at least four persons from the same population as the targeted respondents (e.g., university students) to gather their feedback concerning wording and clarity of expression. Their feedback should be communicated to the original translators who can make any necessary adjustments. Review by persons from the targeted population should continue until the target respondents agree that the meaning of the items is clear and unambiguous.

At this point in the item development process, the questionnaire items are frequently quite good because the opinions of a wide range of persons have been gathered; however, their actual functioning cannot be determined without piloting them with at least 30 persons. The resulting data should be analyzed by (a) inspecting rating scale structure, (b) inspecting item fit to a measurement model (e.g., the Rasch rating scale model), (c) comparing actual item difficulty estimates with the hypothesized difficulties (e.g., the three hypothesized difficulty groups described earlier), (d) checking the relationship between item difficulties and person abilities in the statistical output, (e) inspecting item dimensionality to ensure that each item is contributing to the measurement of the same construct, and (f) checking item reliability. Items that appear to be performing poorly should be revised or replaced unless there is a compelling reason not to do so, and questionnaire developers should make every effort to understand the reason(s) for poor performing items. The results should also be considered in light of what is known about the construct. For instance, questionnaire developers might ask what misfitting items or items that appear to measure a different construct tell them about the construct. It is at this point that the two-way conversation between theory and practice occurs: The theory initially informs our approach to item development, and item performance potentially reveals where our understanding of the construct is probably correct as well as where it is possibly incorrect and in need of further thought. Once final revisions are made, the questionnaire is ready for use.

Final Comments

Questionnaire development is a challenging enterprise because it involves the measurement of abstract psychological constructs. Inferences about the respondents are made based on the data elicited by the items, and for this reason, item development must be conducted with great care. The five steps outlined in this paper are (a) understanding the construct, (b) developing the items, (c) determining the outcome space, (d) specifying the measurement model, and (e) gathering feedback and piloting the questionnaire. These steps can lead to the development of Likert-scale questionnaires that are more likely to yield data that

are reliable and that lead to more valid interpretations. This is because (a) an effort has been made to understand the construct, (b) considerable care has been taken with item development, (c) an appropriate Likert scale has been selected, (c) the questionnaire items have been reviewed by various persons, (d) the items have been piloted, and (e) careful statistical analyses have been conducted to ensure item quality.

A final suggestion that we feel is quite important is that Likert-scale questionnaires should ideally be administered in conjunction with other data-gathering approaches in order to produce a more well-rounded understanding of the construct under investigation and to overcome the inherent limitations of numerical Likert-scale data, namely that numerical data cannot provide a complete picture of educational phenomena. To arrive at a more complete understanding of the phenomena, data-gathering options such as open-ended questions, participant observations, interviews, and objective tests should also be used. By investigating a construct from multiple angles, there is a higher probability of accurately understanding that construct and arriving at more defensible interpretations and conclusions.

Bio Data

David Beglar teaches in the MSEd and PhD programs at Temple University, Japan Campus. His primary research interests are in the areas of foreign language assessment and reading fluency development.

Tomoko Nemoto is a faculty member in the Graduate Education Program at Temple University, Japan Campus. Her research interests are program evaluation and research methodology.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Linacre, M. (2013). Winsteps software version 3.80.1. Chicago, IL: www.winsteps.com

Masters, G. N. (1982). A Rasch model for partial-credit scoring. *Psychometrika*, *47*, 149-174.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan/American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13-23.

National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press,

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosenberg Self-Esteem Scale method effects. *Structural Equation Modeling*, *13*(1), 99-117.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

Smith, Jr., E. V., Wakely, M. B., DeKruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement*, *63*(3), 369-391.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith, Jr., E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement*, *8*, 97-123.

Yamaguchi, J. (1997). Positive versus negative wording. *Rasch Measurement Transactions*, *11*(2), 567.