SURVEYS IN TEST & EVALUATION

REBECCA A. GRIER, PH.D.

Surveys are a common tool in operational test and evaluation (OT&E). Yet, their full value as a tool in assessing effectiveness and suitability is often not realized. This article describes the best practices for incorporating surveys in OT&E, with an emphasis on survey design. The steps to designing defensible surveys are discussed. Additionally, the article highlights best practices for drafting items, selecting response options, and formatting.

INTRODUCTION

A survey is a systematic collection and analysis of data relating to the thoughts of a population. Surveys measure the opinions, attitudes, knowledge, anticipated/perceived behaviors, observations, feelings, experiences, or beliefs of a group of individuals. Because surveys can capture the thoughts of operators on military systems, surveys are useful in assessing the effectiveness and suitability of a system beyond its technical performance. Specifically, surveys of system users can serve as response variables for many aspects of suitability such as, compatibility, transportability, interoperability, wartime usage rates, safety, human factors, manpower supportability, and documentation. In addition, surveys of subject matter expert (SME) observers can provide a quantitative measure of effectiveness of the system to be used by military personnel. Surveys of both system users and SMEs can also be utilized as diagnostic measures (i.e., measures explaining why response variables related to effectiveness and suitability were observed in the test). Diagnostic measures provide insight into changes in tactics, techniques and procedures (TTPs), training, or system design that could improve effectiveness and suitability.

The value of survey data is often not realized in OT&E, because some consider survey measurement as less objective than other measures used in OT&E. It is true that surveys measure subjective experiences (i.e., thoughts) and they are not accurate measures of anything other than thoughts. However, when designed, administered, and analyzed correctly, surveys provide objective, reliable, and valid measurement of these subjective experiences. There is a substantial body of research on how to collect accurate survey data (e.g., Likert, 1932; Babbitt & Nystrom, 1989). Table 1 presents a list of common errors identified in this research. Although there are errors related to all phases of OT&E, from test design to analysis, this article focuses on prevention of errors made during the design of the survey, and it is the first step in conducting an objective survey. All guidelines are based on research.

SURVEY DESIGN

A well-designed survey obtains accurate data from the respondents, in a form usable to analysts, so that the data can be used to answer a question or make a decision. Thus, there are three groups who must be considered in the survey design process: the commissioners (the persons making the decision), the respondents (the persons providing the data), and the analysts (the persons translating the data into information for the commissioner). Each of these groups can introduce error into the measurement as a result of a poorly designed survey.

The commissioner can make the wrong decision because the information that he/she has is incorrect. Thus, it is the survey designer's responsibility to make sure that the specific goal of the commissioner is understood, that a survey is appropriate to that goal, and that the survey, as written, will support that goal. Designing the survey to support the commissioner's goal requires understanding his/her goals and needs and the types of errors that the respondent and analyst can make. The following sections provide detail on the steps in survey design: gathering requirements, drafting questions, and formatting the survey.

GATHERING REQUIREMENTS

The first step in designing a good survey is to gather the requirements of the stakeholders (commissioner, analyst, respondent) and the constraints of the testing environment. The first set of requirements comes from the commissioner because he/she determines the ultimate goal of the survey. The goals for OT&E surveys are typically related to the definitions of effectiveness and suitability. They are derived by understanding the system under test—specifically, the limitations observed in previous test events, the systems requirements, and how similar systems have been used. Whatever the goals for the survey, they must be known before the survey design. If not, the survey will not collect the proper data.

The second set of stakeholder requirements comes from the analyst. The time allowed for analysis is a constraint. If not much time is available for the analyst, then the number of analyses to be performed must be kept to a minimum. It is also best to speak to the analyst regarding the commissioner's goals to determine the appropriate statistics to be calculated. The design of the survey may or may not support the statistical analyses that the analyst needs to conduct to answer the commissioner's goals.

The third set of requirements comes from the respondents. The survey must be written considering their motivation for completing the survey. Respondents who are unmotivated are more likely to provide inaccurate data. That is, they will attempt to complete the survey as quickly as possible rather than as accurately as possible. The motivations of the respondents is affected by what the respondents were asked to do before the survey, their beliefs as to what will happen with the survey data, how interested they are in the topic of the survey, and how frequently they are surveyed. Motivation can be increased or decreased by the design of the test (e.g., completing the same survey multiple times), the design of the survey (e.g., length, question wording), and the administration of the survey (e.g., when administered). Thus, the designer must consider the inherent motivations of the respondents before designing the test and the survey.

Similarly, the survey designer must consider the education level and knowledge level of the respondents. It is impossible for respondents to provide accurate data to a question they have no knowledge of or cannot understand. Further, when respondents do not understand a question easily or feel that they cannot provide an accurate answer, their motivation for completing the remainder of the survey is diminished.

Administration constraints, which come from things such as the test design and the survey delivery method (paper, electronic, verbal), are the last set of requirements to gather. The test design can limit the amount of time that respondents have to complete the survey regardless of motivation. It is important to make sure that the number of questions asked can be answered reasonably in the time allotted.

Finally, each of the delivery methods has pros and cons as well as formatting considerations. Verbal surveys should be avoided. Data collected via verbal administration can be affected by the administrator’s rank, tone of voice, body language, and characteristics (e.g., Zechmeister, Zechmeister, & Shaughnessy, 2001). This is particularly true of lower ranking respondents (both officers and enlisted). In addition, the chance of error is increased when verbal surveys are used because the administrator—not the respondent—must record the response. The one advantage of verbal surveys is the opportunity to ask individuals to expand upon answers. However, this is also possible with paper-based surveys.

Electronic surveys have two advantages over paper-based surveys: there is less error when the data are transferred and when branching questions are used (i.e., when the answer to one question alters the questions the respondent will see at a later point in the survey). Fewer errors occur because the software controls the processes. However, there are also problems with using electronic surveys. Most commonly, survey designers use off-the-shelf survey software rather than hiring a computer programmer to build the survey and database. When using off-the-shelf software, the survey designer is constrained by the question formats, survey layouts, and database format in the software. Thus, the survey designer has very little flexibility in designing the survey beyond drafting the questions. Furthermore, many off-the-shelf survey packages do not present questions in a format that follows the best practices within this guide.

Paper-based surveys allow the designer much more flexibility. Paper-based surveys also have the following advantages over electronic surveys: the administrator can more easily check data for inconsistencies, can follow-up with on-the-spot interviews to get more information about responses, and can modify the survey based on occurrences within the test. These tasks are much more difficult to accomplish with electronic surveys than with paper- based surveys.

Drafting Questions

After gathering requirements, the next step is to begin drafting questions. Every question has three parts:

1. Identifier – A unique symbol labeling each question.
2. Item – The actual words the respondents are addressing in their responses.
3. Response – The data provided by the respondent.

Identifiers are very important to use throughout the survey process. They make it easier to discuss changes to the survey during the evaluation phase, they provide information to the respondent as to how much of the survey has been completed, and they make it easier for the analyst to transfer data and communicate with the test team. However, they are typically not added until the formatting stage. The first step is to draft the items.

5 Golden Rules for Items

When drafting items, there are five golden rules to follow to prevent error in the collected data:

1. User Friendly – do not require a lot of thought or interpretation by the respondent.
2. Singularity – there is only one idea per item.
3. Knowledge Liability – only ask questions that the respondent has enough information to answer.
4. Neutrality – the items do not imply value judgments nor are they emotionally charged.
5. Independence – the response to each item does not affect the responses to other items.

User friendly items should be short, clear, and specific. A good rule of thumb is that they have less than 20 words, few commas, and are written for a sixth grade reading level. One can check the reading level of all documents written in Microsoft Word when completing a spell check. When items are not user friendly, the respondents may misinterpret the question. Even if the respondents interpret the question correctly, items that are difficult to read result in a survey that is more challenging to complete, which affects the respondents' motivations to provide accurate data (Babbie, 2007).

Questions that violate the singularity rule are also known as double-barreled questions. These questions address multiple concepts but only permit one answer. For example, "Was the timeliness and accuracy of the system acceptable – yes or no?" These questions increase the error from the respondent and the analyst. If the accuracy was acceptable but the timeliness was not, how could the respondent answer accurately? When it is not possible to answer accurately, the respondent's motivation is reduced. Furthermore, the analyst can interpret any answer to the question in multiple ways.

Knowledge liability refers to the fact that the survey only obtains accurate information if the respondent has knowledge of the topic. Therefore, only questions that the respondents have knowledge on should be asked. Moreover, if questions cover topics that the respondents do not feel confident about, then their motivations for completing the survey thoughtfully are diminished.

Preserving the neutrality rule is essential for reducing inaccuracies in survey data. This is because humans like to conform. Respondents are very good at identifying "correct" responses, and they will tend to choose those responses even when these response are not representative of their own thoughts. When considering this golden rule, the designer must consider the context of the question and the item wording. If the number of items worded positively and negatively is equal, bias for any one item is not a concern.

As with neutrality, the final golden rule, independence, depends on the wording and the context of the item. Independence leads to inaccurate data as the result of the lack of motivation or the misinterpretation of the questions. The motivation of the respondent to answer truthfully is often lower if branching questions are used, which violate independence. If many branching questions are included, then the respondents will realize certain responses require the completion of more questions. As such, they may choose not to answer truthfully to minimize the number of questions. Therefore, survey designers should minimize the number of branching questions (Babbitt & Nystrom, 1989).

If independence is violated by specifically asking respondents to consider previously answered questions, the responses may be unreliable (e.g., considering your previous responses, rate the system effectiveness.) Some respondents will consider the question as written and base the response on previous responses. Other respondents may treat this question as if it is a general question and ignore the caveat. The discrepancy in how respondents read the question makes it challenging to interpret the data.

Finally, independence is necessary when the planned analyses involve aggregating the data (i.e., combining several questions into a single score). Aggregating responses into a scale improves the statistical sensitivity and specificity of a measure. However, the statistical improvements to the measurement only occur when the responses are independent. When they are not independent, there is a greater risk of data inaccuracy (see Nunnally & Bernstein, 1994 for more information).

Selecting Response Types

The final part of any question is the response, and many different response types can be used. The selection of response type for any particular question must consider the goals of the test, the analyses to be performed, and how the respondent will want to answer the question. All response types fall into two categories: closed responses and open responses. Closed responses are those in which the possible responses are limited to a finite set determined by the survey designer.

Open responses are not limited by the survey designer but require more work on the part of the analyst. Moreover, the analyses of open responses are limited to a count of respondents whose statements can be grouped together. In OT&E, closed-response questions may be response variables or diagnostic measures. Open responses should only be used as diagnostic measures.

Closed Responses

There are several types of closed-response questions. When selecting the type of closed response, one must consider which response type is appropriate in terms of the planned analyses and how the user will want to respond. Closed-response types can provide data that are nominal, ordinal, or interval. Nominal data are categorical only. As such, the statistical analyses for these data are quite limited. Ordinal data are similar to the medals awarded at the Olympics. There are categories, and the categories are ranked. Ordinal data allows for more informative statistical analyses options than nominal data but fewer than with interval data. Interval data provide the most flexibility in statistical analyses. For more information on the types of data, see Nunnally and Bernstein (1994). In the following paragraphs, the various response types (multiple choice, dichotomous, Likert/Likert-like, and behaviorally anchored) are described. For each response type, best practices for writing options and when to use are also described.

The first type of response option is multiple choice. When a question is multiple choice, the respondent selects a designated number of options from more than two options. In OT&E, multiple-choice questions are most typically used for obtaining demographic data. However, multiple-choice questions can be useful for other purposes. Data from multiple-choice questions are typically nominal, which limits the analysis to non-parametric statistical methods.

When writing multiple-choice questions, it is important to ensure that all possible responses are included as options. It is tempting to include an “other” response. However, including an “other” response converts the question into an open-response question, resulting in increased analysis complications. When writing multiple-choice response options, it is also important to remember that most respondents assume that only one response is allowed. Therefore, the options must be mutually exclusive. If they are not, then the question must indicate that the respondent is supposed to select (1) only the best option, (2) all that apply, or (3) up to a certain number.

The second class of closed responses is dichotomous. Dichotomous questions are a special case of multiple-choice questions that have two mutually exclusive options (e.g., True/False). They are a special case of multiple choice because they provide nominal data that are binary. Binary data are the least powerful for statistical analyses. Therefore, a very large sample is required to determine whether there are significant differences in categories.

Dichotomous questions are tempting in OT&E surveys because the decisions being made are often binary (i.e., does it meet the requirement or not?). However, dichotomous questions are rarely appropriate in OT&E. Beyond the statistical limitations of dichotomous data, it is unusual for thoughts or observations to fall neatly into dichotomous categories. Rather, humans tend to think in terms of continuum. If the respondent is asked to put the response into only one of two buckets, the respondent is being asked to draw a line on the continuum. This line may or may not be where the analyst would put the line or where other respondents would put the line. Thus, a response continuum will obtain data that are more accurate. Response continuums include Likert, Likert-Like, or a behaviorally anchored response set. Similarly, one must consider the response options to be included in the question. What might be perceived as a biased question with a dichotomous response option is not perceived as such with a properly formatted Likert or Likert-like response option set.

Likert, Likert-like, and behaviorally anchored response sets are called response continuum because the respondent selects from options that vary by degree. The most well-known continua are the those that range from "strongly agree" to "strongly disagree," which are called Likert (pronounced Lick-ert, not Like-ert) questions. All other attitude-based response options (e.g., acceptable, good, adequate) are referred to as Likert-like.

Behaviorally anchored is a third kind of response continuum. Behaviorally anchored questions are those in which the continuum represents specific actions on a scale of amount, frequency, or goodness. For example, how often do you get up from your desk: once per hour, once every two to three hours, once every four to five hours, once a day or less? Behaviorally anchored response options are especially good for observer surveys.

Response continua are generally preferred in OT&E because when properly written, they have the potential of providing interval data. If not properly written, they provide ordinal data. A response continuum that is properly written will have the following characteristics: (1) four, five, six, or seven response options, (2) equidistant response options, and (3) parallel response options (i.e., an equal number of positive and negative options) (Babbitt & Nystrom, 1989). To meet these criteria, one can use Likert's (1932) original response set (i.e., strongly (dis)agree, somewhat (dis)agree, slightly (dis)agree, neither agree or disagree), the response options proven by Babbitt & Nystrom (1989) to be equal interval, or only end points as in Figures 1 and 2. Likert's (1932) original response set and response sets by Babbitt & Nystrom (1989) were developed through extensive research and proven statistically to have all of the characteristics of interval level data.

Note that it is very difficult to write behaviorally anchored response options that meet these criteria. As such, most behaviorally anchored questions provide ordinal data. In addition, behaviorally anchored response sets are more reliable when each option is labeled with a specific behavior. Thus, using just endpoints is not recommended with behaviorally anchored continuum.

To be clear, the use of response continuum does not guarantee interval data. Rather, there is the potential for interval data. Other conditions also have to be met. These conditions include having an adequate sample size and a relatively normal frequency distribution (for more on this topic, see Zechmeister, Zechmeister, & Shaughnessy, 2001). It is through aggregation of appropriately written Likert or Likert-like response continua that surveys most often achieve interval data.

One of the most challenging questions with response continuum questions is whether to include a neutral response (Babbit & Nystrom, 1989). Note that this is not about having an odd or even number of response options. It is possible, particularly with behaviorally anchored response continua, to have an odd number of response options without a neutral option. Examples of neutral options are neither agree nor disagree, average, and no difference. The reason this question is difficult is that it depends on the requirements of the survey (Babbit & Nystrom, 1989). Specifically, one must consider the motivations of the respondents, the knowledge of the respondents, and the commissioner's purpose. In populations that have low motivation or who are not sure how the survey will be used, neutral responses are a way of answering surveys without considering the question. In these cases, it is best not to have a neutral option (Babbit & Nystrom, 1989). Another consideration related to the respondents is how they will think about the question. Specifically, how likely are the respondents to consider a neutral response as the only acceptable response, which is particularly the case when the users are asked to compare two things. In this case, "no difference" is a valid response. The third consideration is that of the commissioner's goals. If a neutral response will assist the commissioner in making a decision, then it is an appropriate response option. However, this is rare in OT&E.

Formatting the Survey

Once a list of questions has been drafted, the questions need to be combined into a survey. This is the process of formatting and is an important step in the survey design process because it sets the context of the questions. It is well documented that human's thoughts and behaviors are affected by context, and testers must be cognizant of the choices they make that affect the context (Babbie, 2007). As it relates to surveys, the respondent's motivation to complete the survey accurately is greatly affected by the format of the survey. In addition, the format affects the independence of the responses and the neutrality of the questions.

One of the most important aspects of formatting is the introduction. The introduction sets the tone for the survey by informing the respondent of the survey's topic and purpose. To ensure that people read and understand the introduction, it should be brief and clear. If for whatever reason the introduction cannot be brief, the administrator should read it to the respondents. In addition to the topic and purpose of the survey, the introduction should inform the respondents that their responses are confidential (Babbie, 2007). Respondents are more motivated to complete surveys accurately when they have an interest in the topic and they believe that their responses will have an impact (Babbie, 2007). Furthermore, when the respondents understand the purpose of the survey, they do not make assumptions about the purpose, which can lead to misinterpreting questions. Motivation for completing a survey also increases when the respondents know that their responses will not be attributed to them. Not only are these best practices for collecting accurate data, they are also required by the ethics standards for all social science professional organizations (e.g., Association for Psychological Science, Human Factors and Ergonomic Society) and the National Research Act of 1974.

When respondents are completing the survey, their motivation to complete the survey accurately is affected by the effort that they perceive is needed to complete the survey. The more effort the respondents feel they need to put forth to complete the survey, the more errors there are likely to be (Babbie, 2007). Even before the respondents begin to complete the survey, they make an assessment of how much effort will be required to complete the survey. This assessment is based on formatting. Specifically, the number of questions, the number of pages, the number of questions per page, and the amount of white space on a page affect the respondents' assessments of perceived effort. Surveys should have the fewest number of questions possible to meet the goals of the commissioner. When there are more than 10 questions, additional formatting is required to reduce the perceived effort.

Grouping questions into sections is a good way to reduce the perceived effort. Grouping questions changes the length of the survey from 30 questions to 3 sections. Questions should be grouped by topic and response format (e.g., Likert, multiple choice, and so forth). Grouping questions by topic allows respondents to focus their attention on that topic for that set of questions, which reduces the mental effort of the respondent. Similarly, using one response format reduces the effort required by the respondents. When formats are switched frequently, the respondent thinks less about the topic of the question and more about the mechanics of responding, which increases the likelihood of error in the response.

Minimizing the number of open-response questions and altering their format also reduces the perceived effort. Open responses are more arduous for respondents than closed responses (Babbitt & Nystrom, 1989). Instead of selecting a response, the respondent must generate a response. At most, there should be one open response at the end of a group of closed-response questions (e.g., please explain any negative answers) and a group of four or fewer open-ended questions at the end of the survey. When open-ended questions are included, there should be white space—not lines—provided for the answer since lines reduce the amount of whitespace, which increases the perceived effort for completing the survey.

Both the order of the questions within the groups and the order of the groups must be considered. The impact of the order of questions on responses is well documented (cf. Becker, 1954). As such, there are several guidelines regarding the order of the questions. The first questions on the survey should be interesting items that are clearly connected to the goals of the survey. When the first questions on the survey are not interesting, the motivation of the respondent to complete the rest of the questions is reduced. When the first questions are not clearly related to the goals of the survey as defined in the introduction, the respondents begin to think about why the question is being asked rather than the response. This increases the likelihood that they misinterpret the question.

Another guideline for question order is that the respondents should see the order of questions as logical. For example, questions about events in an exercise should be in the order that the events occurred. This again keeps the respondents focused on their responses to the questions rather than the survey construction.

The third order guideline is that the questions should flow from the most general to the most specific. Although this guideline is somewhat controversial, all who construct surveys agree that every question is affected by the questions that preceded it (Babbie, 2007). Since the overall goal of OT&E is to determine the suitability of the system, this general question is the one for which the test team wants an unbiased answer. All questions about the specifics are diagnostic of this response. If the specifics precede the general question, then the general assessment will be affected by the previous questions.

These formatting guidelines are important for paper-based and electronic surveys, and there are additional formatting guidelines for electronic surveys. Specifically, it is important to follow best practices for electronic forms (e.g., U.S. Department of Health and Human Services, 2006). Also, it is important to include in the instructions about how long it will take to complete the survey and show a progress bar to indicate where they are in the survey. This is because in electronic surveys the respondents will not have the visual cues of survey length that exist in paper surveys.

Conclusion

Evaluating, the effectiveness and suitability is about more than just the systems performance during a test event. Assessing the effort and difficulty experienced by the military personnel who are operating and maintaining the system to achieve that performance is just as important. Thus, surveys are a valuable tool in OT&E. Designing a good survey is more than just asking questions. One must begin by understanding the goals of the survey and the common mistakes made by analysts and survey respondents. This article detailed the best practices for designing surveys for OT&E. These best practices were gleaned from the extensive research by social sciences on how to collect the most accurate survey data.

References

Babbie, E. R. (2007). *The practice of social research* (11th ed.). Belmont, CA: Wadsworth Publishing Company.

Babbitt, B. A., & Nystrom, C. O. (1989). *Questionnaire construction manual annex. Questionnaires: Literature survey and bibliography.* Westlake Village, CA: Essex Corporation

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly*, *18*(3), 271–278.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.) New York: McGraw-Hill.

U.S. Department of Health and Human Services. (2006). *The research-based web design & usability guidelines,* enlarged/expanded edition. Washington, DC: U.S. Government Printing Office.

Zechmeister, J. S., Zechmeister, E. B., & Shaughnessy, J. J. (2001). *Essentials of research methods in psychology*. McGraw-Hill Higher Education.

BIO

Dr. Rebecca Grier is an applied experimental psychologist with over ten years experience working to improve Human Systems Integration (HSI). Currently she is a Research Staff Member at the Institute for Defense Analyses, conducting social science research on issues of importance to the U.S. government and supporting the test and evaluation of DoD systems. Previously, at Naval Sea Systems Command Dr. Grier was the acting Technical Warrant Holder for Displays and Controls and the HSI lead for Aegis Modernization. She has also worked at Aptima, Inc. & SBC Technology Resources Inc. (now AT&T Laboratories). Dr. Grier holds a Ph.D. & M.A. in Human Factors/ Experimental Psychology from the University of Cincinnati & a B.S. Honors in Psychology from Loyola University, Chicago.