Practical Recommendations for Validating Survey Apparatuses in Coalition Training Environments

Emilie Reitz
Alion S&T
Norfolk, VA
emilie.a.reitz.ctr@mail.mil

ABSTRACT

If a training event happens and no one builds a record of its gains and outcomes, does it matter? How do you know that the gains and outcomes you recorded, or the tools you used to make that record, are even valid and generalizable to other situations? Are you really improving human performance, or just inferring that you improved it? It's a challenge faced by all communities of research (Teijlingen & Hundley, 2002), whether attempting to solicit survey data in support of human factors assessments or training effectiveness analyses. This challenge is increased in multi-national events, where results contribute to a shared end state for the coalition. To create a valid new measurement apparatus, reliability and validity must be established, and correlations should be built between subscales. Nonetheless, that takes time, results measured from a comparable apparatus or repeated tests, and access to audiences that many researchers lack. During Bold Quest 15.1, two apparatuses were run for precisely this testing and validation purpose and presented to the multinational training audience under one of two circumstances: uncommented testing of the apparatuses or careful explanation of the validation and verification purpose. Two-hundred and seven participants provided over 1600 free text responses which were taken as indicators of their engagement with each apparatus, compared against a non-pilot-tested survey. The pilot-tested apparatuses that were actively administered, elicited significantly more productive responses from the participants than the passive administration groups. Recommendations focus on optimizing apparatuses that cannot be translated into a native language due to constraints, and provide suggestions to bolster both pilot tested and non-pilot tested apparatuses.

ABOUT THE AUTHOR

Emilie Reitz, M.A., is a Research Analyst at Alion Science and Technology. She is currently supporting the Joint Fires Division of Joint Staff J6, Deputy Director for Command, Control, Communications and Computers and Cyber Integration (C5I). In this capacity, she is the data collection and analytical working group lead for Bold Quest, a series of demonstrations and associated analysis focused on interoperable fires applications in a multi-national context. Her research focuses on integrating joint capabilities into modeling, simulation, and training, as a performance enabler.

Practical Recommendations for Validating Survey Apparatus in Coalition Training Environments

Emilie Reitz
Alion S&T
Norfolk, VA
emilie.a.reitz.ctr@mail.mil

INTRODUCTION

If a training event happens and no one builds a record of its gains and outcomes, does it matter? If you conduct training but fail to effectively measure it, then are you really improving human performance or just guessing that it improved?

Creating valid, reliable human performance measures is a challenge faced by all communities of research (Teijlingen & Hundley, 2002), whether attempting to solicit survey data in support of human factors assessments or training effectiveness analyses. This challenge is increased in multi-national military training events. In these events, participants use the same equipment, train to the same standards, or test the same tactics, techniques and procedures (TTPs) in order to create some shared end state for the coalition. During these events, multiple nations interact —and provide survey responses—using a determined common language (typically English), which is often participants' second or third language. Missing the opportunity to collect this data (e.g., due to surveying being too difficult) is unacceptable because the results of these efforts might influence NATO agreements or directly impact the interoperability of systems.

Creating reliable and valid measurement surveys takes time, application of related apparatuses as comparison units or repeated tests, and participant audiences that many researchers cannot access. During Bold Quest (BQ) 15.1, two surveys, or assessment apparatuses, were run for the purpose of checking validity, and they were presented to the training audience under two circumstances: (1) uncommented testing of the apparatuses and (2) careful explanation of the validation, verification, and purpose of the collection. This paper discusses the challenges of building valid surveys in training and testing environments, the outcomes of the data collected in BQ 15.1 through both use cases, and the risks and gains associated with leveraging similar participant groups as the eventual targeted audience for pilot testing.

CHALLENGES OF SURVEY VALIDATION FOR COALITION ENVIRONMENTS

The DoD (2014) and Office of the Secretary of Defense (2014) recently released best practices for surveying and data collection. These publications stress the importance of not just collecting systems information and ground truth from training and interoperability exercises but also pairing that information with warfighters' perceptions, workload data (e.g., from training or interfacing with gear), and other qualitative data. Collecting this information during a large military exercise is a challenge, made more so when the participants come from multiple nations and different services—each with their own operating concepts and internal military jargons that also differ between services.

Some of the other problems inherent in this data collection environment include:

> **Idioms and cross-cultural translations**. In a multi-national context, using an English-language only apparatus may limit the feedback received as compared to an equally valid survey administered in the participants' native languages, particularly if the survey topic evokes an emotional response or is particularly technical (Marshall and While, 1994). Does one translate the apparatus after validating it in English, or is it better to run a side-by-side validation of two separate tools, a translated version and a non-translated version? This, of course, assumes that the developing organization can afford a translation capability and that the schedule allocates time for translation (Sperber, Devellis, and Boehlecke, 1994). Adding to this challenge, many nations' translation services lack knowledge of military concepts and might make unhelpful leaps of logic when creating a translation, leading to confused responses from participants, thereby, adversely impacting the survey results.

SME Validations. Validation of an apparatus by subject-matter experts (SMEs) offers a solution, which is well supported by the literature, but this can cause problems for tools employed across a diverse audience. As an example, a US Army SME validating an apparatus on computer network use might use the word “tier” when asking about which level of a network a participant interacts with (i.e., this is asking whether the participant is a service provider or a service user). From the SME’s expert knowledge perspective, “tier” is the correct doctrinal phrase; for a coalition recipient of the survey, the use of “tier” implies cutting participants off from data due to releasability and creating ‘second class citizens’ from a network perspective. This will create vastly different answers than the SME had intended when defining the question. The same specialization that allows SMEs to perform at a high level of military operations might inadvertently cause them to build gaps and assumption into questions they validate.

Creating a sense of insider trust for multiple groups. Surveys have the potential to trigger a sense of social desirability in respondents, and military respondents may be unwilling to provide negative feedback that they perceive to be “on the record.” These tendencies are also affected by cultural and national norms and participants’ subconscious desire to provide the “right” answer. For instance, most researchers in the US have encountered a participant who provides a warm “on the record” statement on the formal feedback survey but then describes a much more expressive, negative opinion under less formal circumstances. A survey should elicit those less-formal answers on paper and free respondents from a dutiful sense of social desirability or fear of reprisal as a result of their feedback.

These are just some of the many risk factors in building a new assessment apparatus for highly varied and complex groups of participants. To further complicate matters, these types of apparatuses are usually used at rare and infrequent events, offering high payoff if the apparatus is appropriate but creating high costs if the apparatus fails. One way to mitigate these risks is by running a pilot study on a similar, lower-risk population as a validation and proof of concept of apparatus design.

Pilot Studies

Pilot studies are smaller-scale versions of a full study that are often used for pre-testing a research instrument (Teijlingen & Hundley, 2002). The benefits of pilot studies are obvious in certain respects; they include the opportunity to practice a survey’s administration, refine its logistics, and verify the appropriateness of its burden on participants (Peat, Mellis, Williams and Xuan, 2002). During a pilot study, researchers typically ask participants to provide feedback on questions to increase their clarity. Pilot studies also provide a basis for initial statistical analysis of items and subscales, and they create an opportunity to revise the survey prior to administration in a larger study; helping establish the validity and reliability, or lack thereof, of a survey. Potential aspects of miscommunication or any lack of robustness in information focus and analytic rigor can be illuminated at this stage and addressed. While this does not guarantee a successful main study, as more issues may arise after the pilot study, the rewards of a well-executed pilot study outweigh the risks.

Pilot studies do have some drawbacks. The results of pilot studies are often underrepresented in the literature, and the limited-scale results of a pilot study cannot be taken as a sure indicator of an apparatus. For instance, negative results in a pilot study may reveal more about the limited participant population versus the capability assessed by the survey itself, or the quality of the survey. In small communities, such as the coalition military interoperability population, pilot studies may contaminate data in the full-scale study, as the same pilot participants later become the full-scale study participants and are double-tapped, i.e., asked to take the same (or new, or modified) questions during the pilot study and again in the full-experiment, two or three months apart.

PILOT TESTING APPARATUSES AT BOLD QUEST

BQ is a Coalition Capability Demonstration and Assessment event focused on providing a purposeful data collection environment. Sponsored by the Joint Staff J6, it regularly includes upwards of 1000 survey questions administered to hundreds of military participants from up to 14 nations and their services. The BQ15.1 event, working in conjunction with the U.S. Army’s Maneuver Center of Excellence Army Expeditionary Warrior Experiment (AEWE) Spiral J, provided an opportunity to collect data from a company’s worth of US Soldiers, a squad of US Marines, and coalition participants, all of whom interacted with technologies in live fire, field and tactical mission cases. The study described below highlights the pilot testing of three surveys in the most recent BQ/AEWE event.

Methodology

The apparatus used during BQ15.1/AEWE covered a range of topics, but the pilot test surveys described in this paper focused on (1) assessment of TTPs used at the event, (2) TTPs used in active areas of operations, and (3) a collection of questions about participant ethos. The questionnaires assessing TTPs were seeking to assess the content validity of the questions being presented to the audience; the ethos survey was also seeking to refine the content validity, as well as the test-retest reliability. Participants completed each survey at least twice, at the beginning and end of the exercise. For analytical purposes related to the impact of the environment on perception, the ethos-focused collection was additionally administered at a midpoint. Simultaneously, a more mature survey about systems-feedback was administered to all participants under the same circumstances. Overall, 207 participants from 8 nations completed the surveys; of these, 40 were non-native English speakers. The pilot test surveys were presented under one of two circumstances: (1) uncommented testing of the apparatus (*Passive Condition*) and (2) careful explanation of the validation, verification, and purpose of the collection (*Active Instructions Condition*). Refer to Table 1, below.

All survey administrations were provided through an anonymous, non-networked, computer-based survey database, and, as previously mentioned, they were actively proctored. For the uncommented apparatus administration (Passive Condition), the participants were given no more information on the survey than was included in the instructions and during their initial intake and consent process. For the commented apparatus (Active Instructions Condition), all participants were told to stop when they reached the apparatus and then given further scripted instruction explaining that the surveys were being administered for validation and verification purposes and that the researchers were seeking feedback about the apparatus as well as the participants' honest responses to them. For the systems survey, participants were told that the apparatus had been validated and verified in a previous exercise through statistical analysis and pilot testing, but that the researchers were still seeking feedback as well as the participants' honest responses.

Table 1. Study design for the BQ15.1/AEWE survey data collection

	Pre-Test	**Midpoint Test**	**Post-Test**
Passive Condition (*n* = 102)	• Pilot: TTPs • Pilot: Ethos • Systems Feedback	• Pilot: Ethos • Systems Feedback	• Pilot: TTPs • Pilot: Ethos • Systems Feedback
Active Instructions Condition (*n* = 105)			

The passive condition is the standard for survey administration: Although a proctor is present to answer questions and guide participants through the process and instructions are included at the top of the page or on a start screen, no other elicitation is used to draw responses out of participants. The literature for pilot tests stressed the utility of using a different administration technique where the non-experimental purpose of the apparatus is explained and participants are encouraged to provide additional feedback. This led to the hypothesis that Active Instructions would cause the pilot-tested apparatuses to have a higher degree of trainee engagement, leading to more free-text responses provided on each apparatus as well as differing qualities of free-text responses.

Results

The research leads for each survey analyzed the Likert-type scale responses and free-text answers for their respective apparatuses. Analysis of specific items and scales for internal consistency were performed by each researcher, and apparatuses were adjusted based on their statistical analysis and the free-text responses provided by participants.

The results provided in this paper are an analysis of the free-text responses using the Descriptive Coding method, with associated sub-codes (Saldaña, 2012). When using the descriptive coding method, researchers assign a word or phrase that summarizes the data being reviewed. In this instance, each free-text response provided by participants in their surveys were coded. The primary topic codes were: (1) Frustration, (2) Answering Questions, (3) Confusion, and (4) Direct Feedback on Question. Sub-codes were based on types of participant responses; the primary codes and sub-codes are shown in Table 2, and the results of the qualitative coding are summarized at a high level in Table 3.

Table 2. Qualitative codes and sub-codes used to categorize participant feedback to pilot apparatus

(1) Frustration	(2) Answering Questions	(3) Confusion	(4) Direct Feedback on Question
1.1 Frustration at Question 1.2 Frustration at Event	2.1 Productive response (i.e., negative or positive answer to the question, but actively engaging with question) 2.2 Neutral response (e.g. "n/a", "I don't know", etc.) 2.3 Non-Productive Response (i.e., vitriolic response to question)	*No sub-codes for this topic	4.1 Comment on Context (e.g., requesting further clarification, suggesting the context for the answer provided) 4.2 Administrative comment (e.g., identification of phrasing or hard to understand words) 4.3 Encouragement

Table 3. Summary of Coded Participant Free-text Responses

	Ethos Survey (9390 total questions answered)		TTP Survey (2000 total questions answered)		Systems Feedback (19340 total questions answered)	
	Active (183 free text responses)	Passive (156 free text responses)	Active (339 free text responses)	Passive (308 free text responses)	Active (409 Free-text responses)	Passive (229 Free text responses)
(1) Frustration						
1.1 Frustration at Question	1%	1%	–	–	–	–
1.2 Frustration at Overall Event	2%	–	–	–	1%	1%
(2) Answering Question						
2.1 Productive Response	78%	55%	68%	32%	80%	68%
2.2 Neutral Response	10%	19%	30%	56%	7.%	12%
2.3 Non-productive Response	2%	2%	–	1%	12%	19%
(3) Confusion						
	1%	12%	**–**	3%	–	–
(4) Direct Feedback On Question						
4.1 Comment on Context of Question	1%	11%	1%	–	–	–
4.2 Administrative Feedback	4%	–	1%	7%	–	–
4.3 Encouraging (e.g. "Good question!)	2%	1%	–	–	–	–

Each survey garnered different quantities of answers based on the targeted audience of each pilot test. A comparison of the types of comments made was conducted because we hypothesized that participants in the Active Instructions Condition would be more apt to provide meaningful feedback.

To evaluate this, a one-way ANOVA (Analysis of Variance) was conducted on the interaction of type of survey administration with participant responses. There was a significant difference between groups for the Ethos apparatus's responses: $F(1,337) = 4.149$, $p = .042$, partial squared eta = .012, observed power .528; for the TTP apparatus's responses: $F(1,645) = 4.53$, $p = .034$, partial squared eta = .007, observed power .566. For both pilot tested apparatuses, the active administration produced decreases in neutral responses compared to passive administration ($F(1,395) = 10403.9$, $p = .000$), as well as increases in productive responses ($F(1,1083) = 14704.12$, $p = .000$). The systems survey did not have any significant difference in free-text responses as a factor of type of survey administration. As shown in table 3, despite the same active and passive administration types, participants answering the systems feedback surveys did not offer any direct feedback on the questions; they did, however, have a higher rate of non-productive responses.

Next a comparison between the responses of native and non-native English speakers was conducted because of the earlier mentioned challenges of administering surveys of a highly technical or emotional content in a coalition environment. There was no statistically significant difference between the types of free-text responses offered by native and non-native English speakers regardless of proctoring condition, though there was a trend to non-native English speakers offering overall more neutral responses that did not reach statistical significance. With a multi-national population, there is the inherent potential that one might get a set of results that are the same (i.e. similar rates of neutral responses for all apparatus in both administration contexts) but for different reasons. Neutral responses or picking the middle ground could have served as a means for non-native English speakers to express confusion or to communicate decreased understanding of the questions; alternatively, the overall increases in neutral response during passive survey administration could have stemmed from a number of factors including unwillingness to express confusion. Much of the discussion with participants during the active condition was explaining semantics within the questions to all participants.

Next Steps

The answers and additional feedback provided to the writers of the pilot-tested apparatus were taken into account, and led to the solidification of scales, necessary word changes to increase participant understanding, and refinement of questions prior to the next large scale employment of their surveys at the next Bold Quest event and other coalition data collection opportunities. Based on positive feedback both from the researchers involved and the participants who took these developmental instruments, we will continue to purposefully offer audiences the opportunity to engage in pilot-testing new surveys and assessment apparatuses.

RECOMMENDATIONS

Based on the data about pilot tests above, experience, and the literature, we have the following recommendations to improve the quality of surveys in coalition environments, prior to pilot testing of similar developmental or fully validated surveys, and in those instances where developing a parallel apparatus in all participants' first languages is not feasible.

1) **Reduce idioms and cross-cultural translations to improve readability**. Idioms are those innocuous words and phrases that, when used by a native speaker, have a figurative meaning as well as a literal meaning. The English language is peppered with these phrases (which is an excellent example of a phrase not to use for a cross-cultural audience.) Additionally, apply the active tense in your questions and avoid words related to flexible concepts of time (e.g., 'later', 'soon', etc.) (Mores, 1985). When possible, collaborate with bilingual researchers associated with your participant group who understand both the challenges of working in a second language and the goals of your research apparatus.

2) **Create a sense of insider trust for multiple groups by asking for help from the community (or communities) you're assessing**. The difficulties in design and analysis of a new apparatus come from the

fact that a variety of combination variables could potentially be the causative force to produce the results. By expanding your definition of 'SME validation' and approaching leads from groups with which you work or leads of those domains you're assessing, you can check not only the appropriateness of your English language content but also the appropriateness of the words in a military or cross-cultural context.

3) **Emphasize the purpose of the survey**. Make sure the respondents comprehend the importance of the survey, are comfortable in their anonymity, and believe in the potential of their responses to create meaningful outcomes (e.g., impact materiel and non-materiel capabilities in the field). If possible, broach the topic of purpose before the participants are in the room for administration, such as during a large group presentation at the beginning of an exercise.

4) **Minimize the survey load on participants**. After performing a four-hour mission warfighters are tired. They understand the importance of your twenty questions about the mission they just spent hours completing, requesting detailed recall of events, and the successes, failures, and knowledge gaps in employing the systems or processes they were just trained on, but it's still a tiring process. Add in the additional burden of having to translate questions in your head, look up words, or confer with others in your unit to make sense out of each question, and the quality of response you receive could be impacted. Minimize the load on all participants by making sure your questions are tightly written and have a clear purpose.

5) **Be sensitive to other factors that might impact responses**. Even if you, as a researcher, have worked hard to address all of the above factors, you may still see anomalous data, or as described above, similar answers provided for different reasons. Your analysis will ascribe purpose and meaning to the results, but be flexible and view it from the perspective of your larger coalition of participants when performing your analysis.

CONCLUSIONS

While in a perfect world, all new assessment and feedback tools are properly translated and pilot-tested in the native languages of all participants, a pilot test of an English-language apparatus using participants from all targeted members of the coalition being assessed can still produce a valid instrument apparatus.

By creating better assessment and feedback tools in a coalition environment, we can improve the performance of people and capabilities that are deployed with during the next contingency operation. A successful model for cross-cultural apparatus design and implementation would be a valuable tool that could be extended to other training and simulations demands. The framework established can then be adapted to focus on particular areas, issues or challenges faced by coalition warfighters. In-depth analysis of an appropriate apparatus result designed with multi-cultural sensitivities as part of its infrastructure could help to determine causative differences and their relative significance with regard to the results produced. There is more work to be done in the field of apparatus design under challenging conditions, with tight timelines and hard goals and requirements for collection successes—recommending actively administered pilot tests is just a first step.

ACKNOWLEDGEMENTS

This work was supported in part by the U.S. Joint Staff (Contract #W91247-11-D-0010-2B03). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the US Joint Staff or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes.

REFERENCES

Comish, J. (2002). Response Problems In Surveys: Improving response & minimising the load. *Proceedings of UNSD Regional Seminar on 'Good Practices in the Organisation and Management of Statistical Systems' for ASEAN countries, Yangon Myanmar*, 11-13 December 2002.

Department of Defense (2014) Instruction Number 8910.01 Information Collection and Reporting. Washington, DC.

Marshall SL & While AE (1994) Interviewing respondents who have English as a second language: challenges encountered and suggestions for other researchers. Journal of Advanced Nursing, 19(3):566-71.

Mores P. (1985) *Health Care in Multiracial Britain.* Health Education Council, Cambridge.

Murray, C.D. and Wynne, J. (2001) Using an interpreter to research community, work and family. *Community, Work and Family*, 4(2), 157-170.

Office of the Secretary of Defense (2014) Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E). Washington, DC.

Peat, J., Mellis, C., Williams, K. and Xuan W. (2002), *Health Science Research: A Handbook of Quantitative Methods*, London: Sage.

Saldaña, J. (2012). *The coding manual for qualitative researchers*. Sage.

Sperber, A.D., Devellis, R.F, and Boehlecke, B. (1994). Cross-Cultural Translation Methodology and Validation. *Journal of Cross-Cultural Psychology,* December 1994 vol. 25 no. 4 501-524.

Van Teijlingen, E. R., Hundley V. (2002). The importance of pilot studies. *Nursing Standard*, Jun 19-25;16(40):33-6.