

IMPLEMENTING SURVEYS IN AN INTERNATIONAL CONTEXT: AN OVERVIEW

Claudia Tamassia, Consultant, United States

***Abstract.** This paper discusses the implementation of surveys in an international context. It relies on experiences from OECD Programme on International Student Assessment (PISA) and IEA Third International Mathematics and Science Study (TIMSS) to illustrate the steps and discuss issues involved in this process. It also highlights additional complexities when these are applied to an international context that involves multiple languages and cultures. Implementation of a survey involves the development of a framework, the selection of the most appropriate means of collecting data, development of the instrument, analysis and further studies. It also discusses ways in which an evaluation of educational facilities could be implemented, either independent or linked to current studies so relationship with performance would be possible, and discusses further alternatives.*

Introduction

The relationship between the environment in which learning takes place, namely schools or other types of learning facilities, has been widely documented and shown to play an important role in student performance. The physical conditions of buildings such as the air quality, temperature and building structure are variables that will have either positive or negative impact on performance, depending on their conditions. These variables will also influence quality of teaching, affect attendance of students, and increase operating and maintenance costs. Therefore, continuous evaluation will be essential to ensure that the facilities are meeting their intended needs. These evaluations can be implemented as stand alone studies, with the sole purpose of describing conditions or in conjunction with other existing surveys allowing them to be linked to student level variables, including performance.

Evaluations are complex and multidimensional. The simple agreement of what are the important elements to be examined can be difficult to be achieved. When these variables are examined within the national or local perspective, it is easier to reach an agreement concerning the appropriate level of comparison and even the set of variables to be examined. When a national perspective is replaced by an international one, the complexities of designing and implementing such an evaluation increase and the definition of what is important becomes entangled with the methodological complexities.

In a national evaluation there is a possibility of linking ready-available performance data coming from large-scale assessments at the state or national levels with the new data collected on facilities. The use of previously collected performance data would only be suitable if these were collected through a standardised assessment, thus offering data that are consistent and valid across schools.

In the case of an international evaluation, comparable performance data across countries are rarely available and therefore, the data collection process of student-level information should be designed as part of the experiment.

International surveys, a method of collecting data from a sample or population, provide an opportunity to collect outcome information (*e.g.* performance) and contextual factors (*e.g.* students, teachers or schools background information), and afterwards relate them through statistical analyses. Although, traditionally, international studies such as the OECD Programme for International Student Assessment (PISA)¹ and the IEA Third International Mathematics and Science Study² (TIMSS) have asked general questions about school infrastructure and facilities, these have often been superficial with little explanatory factor or direct associations possible.

The collection of contextual information has been embedded in many educational studies, particularly the international ones. Mullis (2002) mentioned that contextual information are collected for a variety of

purposes such as:

1. To identify factors that are related to student performance.
2. To describe the population of students participating in the survey,
3. To evaluate potential bias associated with non-participation,
4. To examine opportunity-to-learn issues.
5. To examine the distribution of educational aspects and facilities among different group of students.

Additionally, within the context of PISA, Harvey-Beavis (2002) described the value of examining relationships between contextual variables and performance to: *i)* identify differences between countries in the relationship of achievement and student- and school-level factors, *ii)* examine the proportion of variation in achievement between and within schools, *iii)* examine the impact of schools in moderating or increasing the effects of individual-level variables on achievement, and *iv)* address and examine these relationships over time.

Although a variety of methods exist for this purpose, on the international setting, the most widely used method continues to be the paper-and-pencil questionnaire due to its practicality, easiness to administer and analyse, with insurable comparability. These instruments are often targeted to students, teachers or schools on a variety of issues. These questionnaires provide valuable information on the environment in which learning takes place and allow for the association between these contextual variables and performance. This association, in turn, provides an in-depth understanding of which variables affect performance and in which way this association takes place. Within the international setting, questionnaires also offer higher control over the instrument being administered, without the cultural influence of coaching or impact of the interviewer technique as in the traditional interviews. There are however, also limitations to this method. For example, respondents may answer to questions in superficial way due to lack of interest, response rates tend to be lower, and there is no alternative to written communication (*e.g.* gestures, intonation, or other visual cues).

Other methods of surveys include structured or telephone interviews, case-studies, review of records or observational techniques. Although most surveys involve a single mean of data collection, a combination of methods is also possible. For example, a paper-and-pencil questionnaire followed by an interview for more complex issues.

On the theme of school facilities, PISA 2000 collected data on the quality of schools' physical infrastructure by asking school administrators to judge the extent to which learning of 15-year-olds was hindered by: poor condition of buildings, poor heating/cooling and lighting systems, and lack of instructional space (*e.g.* in classrooms) (OECD, 2001). In 2003, TIMSS and PISA collected data on the degree to which the following factors hinder instruction: school building and grounds; heating/cooling and lighting systems; and instructional spaces (*e.g.* classrooms).

The impact of multiculturalism

In international contexts, the impact of multiculturalism and multiple languages should be considered throughout the process: from development to analysis of results. Their implications extend to all aspects of surveys and therefore, care should be taken possibly through the involvement of multiple reviews with multi-cultural committees in all stages.

International surveys such as PISA, TIMSS or PIRLS include paper-and-pencil questionnaires as part of their instruments but these instruments are usually implemented in the local language, thus adding complexity through a translation process. Translation is the most common cause of errors that invalidate questions and their results. Instruments are originally developed in one or two languages (*e.g.* PISA uses

English and French), which are then used as the source languages for translating additional languages. In this case, the emphasis is on developing source version(s) that will be linguistically and methodologically correct so they can serve their purpose of “models” to local translations. Although countries would still need to understand one or both source languages in order to translate them, the actual respondents would be answering in their local languages. A detailed translation manual needs to be available providing information on the translation methodology (*e.g.* back translation as used in TIMSS, double-translation followed by reconciliation as used in PISA), guidelines for translation/adaptation (*e.g.* differences between translation questionnaires and cognitive instruments) and common interpretation problems. As an additional measure, PISA includes notes and guidelines for translation within items.

Day and Evers (2002) examined this impact on using a “questionnaire to assess cultural factors in the acceptance of computer interfaces” (p. 2). In this case, questionnaires were implemented in English to both English and non-English speaking countries. Therefore, the issue of appropriate English level was crucial as the same instrument would be answered by both English and non-English speakers.

Validity of responses is another important consideration on international surveys. Often cultural issues affect the quality of the responses as some participants tend to respond what is “socially acceptable”. For example, when asked how many hours you read a week, rather than responding “none”, some students may decide to respond on the highest category, *e.g.* “two or more hours” as they believe this is the expected or most correct response. This type of issue invalidates data and relational analysis with other variables, particularly with performance.

Methodological issues

The development of contextual instruments or surveys is neither single-stepped nor simple. Consistent with the results from other measures from social science, education surveys or research will always have a certain amount of error associated with the results. The purpose of this section is to highlight some of the sources of errors and to mention how some of these errors can be minimised, thus making the results more accurate and valid.

Development of frameworks

Frameworks provide “the theoretical work guiding the development of the context questionnaires” (Harvey-Beavis, 2002, p. 35). They present the overall goal of the questionnaire and are used to guide the instrument development and the interpretation of results. Additionally, they offer a common language to discuss definitions and suppositions of what is important and the constructs that will be examined. One characteristic of a questionnaire framework, compared to a domain specific framework such as reading literacy, is that it is not based on a single dimension or definition. A questionnaire is usually made up of several indices and aspects of students, teachers and schools together and usually limited by time. In education, contextual instruments are usually targeted at a multitude of concepts ranging for demographical and home background information to specific concepts related to teachers, classrooms or schools, thus requiring a compromise as to how much it can ask. Questionnaires are thus made up of several constructs. These constructs have to be related in a meaningful way in order to be valid. Figure 1 illustrates a grid used in PISA to guide the development process (replicated from Harvey-Beavis, 2002: 35). The cells marked in bold (cells 2, 4, 5, 8, 10, 11 and 12) represent the core elements that PISA can cover in its current design. The design covers school and student levels, with little information collected at the classroom level.

Figure 1. Mapping the coverage of the PISA 2000 questionnaires

	Antecedents	Context	Content
System	1. Country features	2. Institutional settings and policies	3. Intended schooling outcomes
School	4. Community and school characteristics	5. School conditions and processes	6. Implemented curriculum

Class	7. Teacher background characteristics	8. Class conditions and processes	9. Implemented curriculum
Student	10. Student background characteristics	11. Student classroom behaviours	12. Attained schooling outcomes

Every survey will have a theoretical scheme guiding its work and the levels will depend on the type of research questions and nature of the data being collected. Without such guidance, the results would be a sequence of variables without a theoretical framework linking them.

Selection of the most appropriate means of collecting data

Once the domain of what is to be assessed is identified, a second step is the selection of which type of instrument better matches the objective. If the interest is on collecting quantitative data from a large sample of cases, a paper-and-pencil questionnaire is probably an appropriate option. Alternatively, if the interest is on fully understanding some characteristics from a smaller sample, in depth analyses such as case studies or face-to-face interviews could be more appropriate. A thorough assessment of the suitable method should be the second step, following the definition of a framework.

Paper-and-pencil instruments offer a quick way to collect quantitative data from a large sample of respondents. Response rate is often a problem as respondents often decide not to answer a set of questions or the full questionnaire, either due to lack of interest, extended length of the instrument, complexity of answers or even the type of information being collected (*e.g.* personal information).

Alternative designs include interviews, case studies or observational techniques, which can all be applied with a design employing statistical sampling. Interviews and case studies are briefly discussed below.

Interviews are based on a process where the interviewer asks questions, and records the respondent's answers. The interviewer plays a strong role throughout this process and thus, comprehensive training is necessary, which adds complexities to the use of interviews. Interviewers have an impact on response rate, on the way a question is asked and on the way the answers are recorded. A training session and a structured scheme for the questions should be prepared to limit the impact of interviewers by ensuring that questions are consistently asked across respondents and bias eliminated. This consistency limits the variation across interviewers. In an international setting, this training is particularly important as cultural differences tend to influence how questions are posed. Structured face-to-face interview could provide an even deeper understanding of these variables than paper-and-pencil. However, some of the drawbacks include the high cost associated with the training of appropriate interviewers, its implementation and analyses, and the complexity in transcribing and coding the responses into a comparable format. These issues become even more apparent and important when they are applied to a context where these interviews would be implemented in multiple languages (in the case of PISA in more than 20 languages).

Case study represents in-depth analytic descriptions of event, processes, facilities or programmes based on either a single case or multiple cases. Case studies are used when the purpose is to collect a large amount of data from only a few cases. Due to complexities, case studies are also appropriate as a second step in an evaluation process, where a data was first collected from a large number of cases through questionnaires with resulting data used to select a few unique cases. A more in-depth qualitative procedure, such as case studies, would then be conducted in these few cases to complement the data collected through the questionnaires. For example, as part of TIMSS, a video study of classroom teaching was implemented as a follow up from the 1995 and 1999 surveys to examine teaching and learning in mathematics and science on a selected group of high-performing countries.

Independent of which method is selected to collect data, the "population" issue will be central to the success of the study. This will have a direct impact on the type of inference, level of confidence and

comparability of results. Census, where every individual is surveyed, versus a sample, where only a small subset from the full population is surveyed, is probably the first decision. Additionally, samples can be either by convenience (as often used in preliminary data collections or field trials) or probability-based (a necessity during the final data collection). If a sample is selected, it is important to ensure that all individuals from the population have an equal chance of being selected. Sample is often by method chosen by educational studies, which often gather data from a subset of the population of students, teachers or schools for inferences about the population. In this case, the goal is to minimise sampling errors by ensuring that the selected sample can accurately represent the full sample. The quality of a sample is more often judged by the process that was used when selection cases than by the characteristics of the sample or the results.

Development of the instrument

The development of items for a questionnaire is not simple and one of the most important steps if results are to be meaningful and meet the research objectives. This process should involve people from various background including content specialists, age-group specialists, language experts and national or international experts on survey. Content specialists will ensure that the questions are addressing the intended construct and measure all necessary elements for the analysis phase by also matching the framework. Age-group specialists will play a role in ensuring that the question are written to the appropriate age group - questions written to an adult population may not be suitable to a teenage population. Language experts will ensure that the questions are linguistically accurate, clear and concise, and written in a non-offensive way and without bias. Survey specialists will ensure that the methodology is the most appropriate, the underlying concepts will be understood by a multi-cultural set of respondents, and the items are free from translation barriers or offensive characteristics.

The format of the questions will impact the quality of results and the rate of non-responses. Open-ended questions can be answered by either a single word or a long statement and are appropriate when respondents know how to express themselves and a variety of responses or unanticipated answers are the target. When open-ended questions are used in self-administered questionnaires, they tend to provide limited data because when an interviewer is not present, often answers are incomplete or deviate from the question objective making them difficult to compare with answers from other respondents (Fowler, 1993).

In many quantitative surveys, in order to avoid analytical problems and high costs, the choices of responses is limited through a closed-response format that requires a specific answer or usually a check mark on the selected category. In this case, it is important to ensure that every possible response category can be included within one of the offered alternatives. They can also be useful when the respondents' level of motivation is not high as they are quickly answered. Once the question is finalised, any alteration will also alter the pattern of responses, particularly changes implemented to question after the trial period.

In closed-response formats, the scale used for responses will also have an impact on the validity. Likert-scales are widely used when the purpose is to assess attitude. PISA utilises a four-point scale while Day and Evers (2002) utilises a six-point scale with the justification that this would "prevent responses centring by subjects whose culture might discourage them from taking a distinct position on the issue" (p. 3). Data resulting from Likert scales are ordinal, that is, they have an inherited order or sequence but not equal distance between levels.

Rather than conducting analyses at the individual item level, another common technique is the development of indices, where individual items are grouped to for a construct or index. Wolfram (2002) describes that this methodology is called complex indices and it summarises "responses to a series of related question selected from larger constructs on the basis of theoretical considerations and previous research" (p. 217). The PISA 2000 index of the quality of the schools' physical infrastructure is based on the response to "in your school, how much is the learning of 15-year-old students hindered by": *i*) poor condition of buildings; *ii*) poor heating, cooling and/or lighting systems; and *iii*) lack of instructional space (e.g. classrooms). These were answered using a four-point scale with the categories *not at all*, *a little*, *somewhat* or *a lot*.

Also important and affecting the rate of response is the reading level embedded in the questions. Day and Evans (2002) found that in order to accommodate subjects' lack of English knowledge, some questions were written in extensive form which resulted in many respondents not reading them in detail. Whenever possible, questions should (de Vaus, 1996):

- ∞ Be simple, and as short and direct as possible;
- ∞ Avoid jargon, be unambiguous and without many technical terms;
- ∞ Be independent - should ask a single question at a time;
- ∞ Be written in a way that is not leading in any way – respondents should not feel that a single answer is expected or that they are incorrect; and,
- ∞ Should be positively worded.

When designing the final instrument, it is also indispensable to:

- ∞ Include an introduction to the purpose of the survey.
- ∞ Present clear directions on how to respond to the questions.
- ∞ Arrange the questions in a thematic fashion with thematic headings whenever possible.
- ∞ Present an appropriate spacing between questions and sufficient response space for open-ended questions.
- ∞ Finish with a thanking statement to participants.

A trial phase is essential during the development phase of a survey. Survey questions are very sensitive to small changes, which can have a strong impact on the actual responses, response rates and interpretability. Even small changes, implemented for clarification, can invalidate results or comparisons as respondents are likely to understand them differently.

Analysis

Once data are collected, they need to be analysed. The type of analysis will be driven by the research questions to be answered by the survey. de Vaus (1996) describes three factors that will affect the way data are analysed: the number of variables, the level of measurement, and the purpose of the analysis as either descriptive or inferential. During the development of a survey, there are two important analyses phases: during development and during the interpretation of results.

Although the interpretation of results is towards the end of the process, it is integrated throughout the study as an analysis plan that is related to the purpose of the study (*e.g.* research questions) and the theoretical framework. The selection of questions is a second step.

The analysis during the development phase, particularly after a trial session, will ensure that that data are methodologically and theoretically appropriate, that is, do the questions being asked measures the intended construct? This process start with a literature review on the constructs being assessed followed by an in-depth review of questions that were used in previous surveys that assess the same domain, including their structure and results. The validation process will use structural equation modelling and factor analysis to confirm the theoretically expected dimensions or re-specify the dimensional structure, results which should always be published with the study. These analyses will also check possible ranges or response and inconsistencies.

Once questions are developed but before they are submitted to a large field trial, cognitive laboratory interviews (or cognitive labs) can bring benefits as they ensure that the comprehension and understanding of respondents was as expected. They are used to examine the mental processes of respondents when answering to the questions and should apply to a small sample of volunteers. PISA implemented these in only a few countries on a small number of respondents. This is a widely used method for developing and validating surveys. Respondents are brought into labs on a one-by-one basis and advised to think aloud during the procedure. The think aloud procedure is particularly useful in identifying parts of the questions that they were interpreting differently from expected and thus, prevented them from responding according to the expected guidelines.

Further studies

The paper-and-pencil questionnaires have been widely used in large scale assessment due to its advantages in collecting data from a large sample in a limited time – data that are reliable, comparable across countries and that allows for quantitative analysis to be performed. However, this is not the only way, as previously mentioned, and neither the end of the study. It is possible that once quantitative data are analysed, further research questions emerged that require the collection of more in-depth data on a few selected cases. One alternative would be to conduct a follow-up study on a few cases through either face-to-face interview or case studies. This would complement the data collected through the questionnaire and in many cases, validate the previous results.

Even though quality control procedures should be implemented throughout the survey, validity is still a concern. The overall purpose of a survey is to collect data that are as accurate as possible comparable to the real data – any deviations are considered errors. According to Fowler (1993) there are some common threats to validity. The first one is the lack of understanding across respondents to what the question is asking. Survey developers need to ensure that the question is written in a way that will ensure a common understanding and that definitions are included in the instrument whenever appropriate for clarification purposes. The second one is a lack of knowledge of the possible types of errors, for example, errors coming from questions or from the design. This includes requiring an answer from a topic that the respondent is not responsible for or is not the appropriate person to report on it. The third one is social desirability related to what is either socially accepted or too sensitive that respondents decide not to report accurately. This is related to the socially acceptable responses and highly affected by cultural differences across countries. Research is essential during the development and analysis phases to determine if questions were consistently understood by all respondents, if answers can be trusted and are comparable across respondents. Lack of confidence that the responses are accurate will invalidate the outcomes.

Conclusions

Educational research is complex and these complexities are even more evident in international studies by the addition of a level of analysis that involves multiple languages and cultures. The need to ensure that data are comparable across countries involve the establishment of strong procedures for quality control, detailed documentation and manuals, training sessions, validity and reliability analyses and the commitment of all involved that these procedures will be followed.

The evaluation of educational facilities is not different from the evaluation of other variables affecting learning – the level of details will differ. This can either be done independently through paper-and-pencil questionnaires, observational methods, interviews or case-studies. The most appropriate method depends on the purpose of the study. If a direct link between these variables and learning is important, a simultaneous data collection of both aspects is perhaps the best alternative. As some international studies are already established with methodologies accepted by the educational community, there would be advantages in adding an educational facilities study to the current design. In the case of PISA, this decision would need to be taken by the PISA Governing Board considering the level of interest and costs. Similar procedures will be taken by other studies.

Another possible alternative is to implement an in-depth data collection procedure of school facilities as an optional component. This is similar to what was done in TIMSS with the video study and in PISA

with the self-regulated learning questionnaire. Assuming that such an evaluation is of interest to only a few countries, it would still be possible to implement it study on an optional basis where countries decide whether to participate on individual bases. In this case, the component would still be implemented through the international consortium, thus allowing for international comparisons and an international database.

Notes

1. PISA was developed by the OECD in 1997 and implemented in 2000 and 2003 to “measure how well young adults, at age 15 and therefore approaching the end of compulsory schooling, are prepared to meet the challenges of today’s knowledge societies” (OECD, 2001, p. 14). It assesses reading, mathematical and scientific literacy and was implemented in 43 countries in 2000 and in 41 countries in 2003 (OECD, 2001; OECD, 2003; OECD, 2004).

2. TIMSS, the original Third International Mathematics and Science Study was first implemented in 1995, TIMSS-Repeat in 1999 and the Trends in International Mathematics and Science Study in 2003. It is implemented by the International Association for Evaluation of Educational Achievement (IEA).

References

Day, D. and Evers, V. (2002), “Questionnaire Development for Multicultural Data Collection”, <http://hcs.science.uva.nl/usr/evers/IWIPS99DON.pdf>.

de Vaus, D.A. (2002), *Surveys in Social Research* (5th ed.), Routledge, London.

Fowler, F.J. Jr. (1993), *Survey Research Methods* (2nd ed.), Sage Publications, Newbury Park, CA.

Mullis, I.V.S. (2002), “Background Questions in TIMSS and PIRLS: An Overview”, Paper commissioned by the National Assessment Governing Board, <http://www.nagb.org/release/Mullis.doc>.

OECD (2001), *Knowledge and Skills for Life. First Results from PISA 2000*, OECD, Paris.

Harvey-Beavis, A. (2002), “Student and School Questionnaire Development”, in Ray, A. and Wu, M. (eds.), *PISA 2000 Technical Report*, OECD, Paris, pp. 33-56.

OECD (2003), *Literacy Skills for the World of Tomorrow. Further Results from PISA 2000*, OECD, Paris.

OECD (2004), *Learning for Tomorrow’s World. First Results from PISA 2003*, OECD, Paris.

Wolfram, S. (2002), “Constructing and Validating the Questionnaire Indices”, in Ray, A. and Wu, M. (eds.), *PISA 2000 Technical Report*, OECD, Paris, pp. 217-252.