HANDS-ON SURVEY RESEARCH IN A VIRTUAL ENVIRONMENT ®

Peter Theuns and Walentina Cools
Vrije Universiteit
Brussel

In survey research, sub optimal sampling methods or formats of the questions asked can result in biased data, and so in poor results. Teaching this topic is hard because students can only "believe" the teacher and try to understand why and how biases can occur and contaminate the data. This paper introduces a new generic electronic learning environment that gives students hands-on experience with how their methodological choices affect the data. The learning environment consists of three modules. In the population module, the teacher defines a population. In the sampling module, the student can apply different sampling plans. In the survey module, the student can design a questionnaire and actually execute the survey. The resulting data file can be analyzed and compared to the population data. It is concluded that hands-on experience in a problem-based approach can support a deep understanding of several types of sampling errors and response biases.

INTRODUCTION

Developing generic academic competences like analyzing and solving problems systematically, the skills to decide which data are relevant, how to collect them, and the ability to assess the quality of the obtained results remain of core importance in higher education. Modern information and communication technology can support education of these competencies by making learning more attractive and thorough by presenting knowledge in an appropriate context (Norman & Schmidt, 1992) and make education become more related to actual experiences (Gijselaers, 1999; European Round Table of Industrialists, 1997). New technical developments allow learning to take place in a virtual reality that meets the above requirements.

The view on education that is adopted here builds partly on Dewey (1916), who stressed the importance of real life situations for learning "problem solving skills" which he found more important than "knowledge", and Bruner (1961) who stressed discovery learning. To both authors the term "action" was fundamental. Media in general, but surely computers can play a crucial role in the specific kind of action that is referred to as "mediated action" (Harel & Papert, 1991, and Wertsch, 1991). The resulting "situated cognition theory" further develops these notions and stresses the social and cultural dimensions of knowledge and learning (Bruner, 1990; Harré & Gillett, 1994; Kirschner & Whitson, 1997). Anchored instruction accentuates the idea that knowledge and competencies must be presented in contexts that correspond to contexts that the students will face in their professional lives.

According to Mandl and Reinmann-Rothmeier (1995) learning environments developed within a constructivist framework should meet four principles. Learning must start from authentic problems and situations; it must refer to multiple contexts; it must contain several perspectives of the learning object and it must be part of a social context. The approach of learning in a virtual reality that is introduced here fits well in this context. We are not aware of a theory on learning in a virtual reality, but some similarities are found in constructivism, anchored instruction, and in situated cognition (Kirschner & Whitson, 1997; Harel & Papert, 1991).

In this contribution a new learning environment is introduced that supports teaching survey methods and related research methods where sampling participants and operationalization of researched constructs are important. A typical objective of a course on survey methods is that students realize that the quality of the data they will eventually collect in a survey largely depends on the methods chosen. Indeed, when setting up a survey, many choices must be made. First, one must decide how the questions will be presented: will it be a personal interview, a telephone survey, a mailed questionnaire or -why not- web-based? Second, a sampling method must be opted for. Third, the whole build of the questionnaire is to be determined: its length, the question order, the question wording, and the response format(s) for the questions are to be decided on.

One of the difficulties in teaching survey methods is that is practically impossible to let students experience how badly methodological choices may affect their data. One of the most important problems in this context is that mostly, in practice, the population is unknown (and cannot be known). Typically teaching on these problems comes to either lecturing about do's and don'ts, and/or discussing cases or examples in which sampling errors and response biases occur or are dealt with. Assessing the size of the error effects and experimenting with alternative approaches is often difficult if not impossible.

In this paper a learning environment is presented that allows teachers and students to actively experiment with simulated data. In this learning environment methodological choices can interactively be altered and the consequent effects on the obtained data can be assessed.

SOME TERMINOLOGY

Here a population is conceived as a (finite or infinite) set of elements (individuals) that embody *constructs* (often called latent variables) considered relevant for the content of the survey. Constructs are considered to follow theoretical distributions for which type and parameters are set by the teacher and may differ in the respective population strata. For example, if the elements are people, then their age can be described with some (series or combination of) theoretical distributions, one for each population stratum. The constructs in the population may be either directly accessible =*empirical constructs* (e.g. "body weight", "age", "gender", etc.) or only indirectly accessible =*hypothetical constructs* (e.g. "in love", "education", "social background", "attitude towards religion", etc.). Typically, in order to assess constructs a researcher will select *variables* (the questions or items in a survey context) that operationalize the constructs. So, a population (and a sample thereof) holds constructs, researchers conducting a survey on that population obtain values for variables.

In this learning environment a population is not conceived as homogeneous in the sense that all constructs be independently distributed. Rather, a population is considered to consist of strata, sub populations in which specific distributions (or parameters) of constructs apply and different relations among constructs may exist. For example, it could very well be that in a population the average income differs among the genders or that a correlation would exist between intelligence and salary for the one gender while no such relation would exist for the other gender. These stratum specific distribution specifications and interrelations among constructs are described in *scripts*. Scripts operationally define strata.

One way to imagine how these terms combine in the virtual world that we are have constructed is to consider the population of some imaginary city. Different "kinds" of people, for example the "class of workmen living in district x" or "rich elderly people" or "working women", etc. constitute strata. Note that the set of strata is a partition of the population. Obviously the people within one stratum are not all the same, however, they are modeled to show constructs that are drawn from theoretical distributions that are typical for that stratum. If, for example, women live longer than men, and the stratification would be based on age, then in a stratum of elderly people the probability of the construct "gender" = "male" may be lower than in strata that describe younger people. Even the distribution from which a construct is drawn may differ among strata. It is important to accept that we do restrict the description of people to a finite and known set of constructs that we assume relevant to determine participants' responses to a survey.

Probably constructs that will determine responses are both constructs of interest (the actual constructs the researcher wants to assess) plus a number of participant constructs of disinterest, but which may affect the way the participant reacts to the survey. In case the survey is about the relation between income, schooling and consumer behavior then constructs of interest may be salary, bonuses, bank interests, income from real estate etc., highest level of education, professional training etc., and shopping habits, eating habits etc. Constructs of disinterest could in this case be, for example, personality, boredom, vision, fatigue, reading ability, distraction, bias towards social desirability etc. It is clear that both types of constructs will affect the way in which people react to the survey. Not only may both types of constructs interact to shape (or bias) the responses, also they will affect cooperation and eventual non-response.

While the learning environment consists of three modules (population description, sampling, and survey), from a didactical standpoint it makes more sense to look at how the environment is used by the teacher and the students. On the teacher side all descriptions about the population, the sampling scenario's and the survey items take place. On the student side we have the experimental part where samples are drawn, the survey questionnaire is developed and the virtual administration of the survey is simulated.

LEARNING ENVIRONMENT – TEACHER SIDE

A main task of the teacher is to describe the population. However, the teacher may find it easier to first make up a comprehensive list of (potential) questionnaire items to figure in the student environment later on. Doing so may be helpful for completing the next.

Once a list of items is drafted, the teacher must make up the set of constructs (of both interest and disinterest) that are considered relevant to generate responses to the questionnaire items. For example, if one of the questions is “How old are you?” then the construct of interest will be the actual age of the person; a construct of disinterest may be the person’s propensity to lie about it. Next, the number of strata and their respective sizes must be decided on. The strata can be used to distinguish groups of elements that show specific (distributions of) values for one or more constructs or that show specific interrelations among constructs.

Now the actual parameterization of the population can begin. Each script describes the distribution (type and parameters) of all the constructs within its corresponding stratum. Also, the interrelations among constructs as they appear within the stratum are fixed in the script. Values for all parameters, distribution types and correlations as they occur in one stratum make up a script. For each stratum such script must be defined. Once the constructs, the stratum sizes, and the scripts are defined, a population is generated (sampled) by the learning environment.

Second tasks for the teacher consist of writing sampling “scenarios” and determine their theoretical equivalent. Rather than to model only the different theoretical sampling schemes (random, systematic, stratified, etc.) it is opted for that students would be able to experiment with different sampling schemes that are presented as a number of “stories” that reflect actual survey administration practices. Here, for example, it could be described that a researcher has hired pollsters that walk around in certain districts where they then interview people in the street. Such approach would likely be modeled as “non-probability sample” because probably some strata would be over/under represented in the sample. This part completes the teacher environment. It turns out that the teacher environment produces a population file in the first place, but on top of that it also produces a comprehensive list of survey items, and scripts. All three (population data, items, and scripts) are fed into the student environment where they will serve to generate virtual responses to the items.

LEARNING ENVIRONMENT – STUDENT SIDE

For students the learning environment is a ‘place’ where they can experiment with both sampling methods and questionnaire design, but most of all they can experience the impact of these on the obtained data. Students can draw samples from the population according to a number of implemented plans among which simple random, systematic with random start, stratified, cluster, staged, purposive, and quota sampling. Depending on the scenarios implemented by the teacher these plans can be presented as “stories” or can be selected from a list. The (construct) data in a sample are a strict subset of the population data. An interesting feature of the learning environment is that the population constructs matrix $\mathbf{C}_p$ is known and has the same type of format as the sample constructs matrix $\mathbf{C}_s$. So, bias resulting from a sub optimal choice of sampling approach can be assed data-analytically from a comparison of $\mathbf{C}_p$ and $\mathbf{C}_s$.

Next, students can experiment with questionnaire design. Here students can select items and response formats from the list that was previously designed by the teacher. They can choose to design a long or short questionnaire, select a single response type (e.g. only true/false items) or mixed (some rating scales, some true/false, some itemized scales). While the teacher has made up a list of possible items and response formats, and the scripts that incorporate the interrelations among constructs, responses to items can be generated in the survey module. Although models exist that describe relations between latent variables (constructs) and observed or measured variables (see for example Jöreskog, 1973, and Hayduk, 1987), these were not adopted here because they were developed mainly to assess latent variables from observed variables. Here however, the latent variables are considered “known” while the processes that are assumed to translate combinations of latent variables into overt responses are considered manifold and not (necessarily) linear. Therefore the following new model is implemented.

$$\mathbf{R} = b_{\mathbf{C}}(\mathbf{C}\beta) + \mathbf{E}$$

where **R** = the ($N \times k$) matrix (r_{ij}) where r_{ij} denotes the response obtained from element i to item j in the questionnaire. N denotes the sample size; k equals the number of variables or items in the questionnaire administered.

C = the ($N \times m$) constructs matrix (c_{ij}) (which is equal to either $\mathbf{C}_p$ or $\mathbf{C}_s$) where c_{ij} denotes the amount (measure) of construct j that is present in element i. Moreover, m denotes the number of considered constructs.

β = the ($m \times k$) matrix of weights that describes how constructs combine to yield the true score of a variable

$b_{\mathbf{C}}$ = denotes a (composed and often non-linear) function that determines the bias in individual responses due to respondent behavior (inclination towards social desirability, non-response, exaggeration, etc.). This bias is modeled as depending on the constructs present in an element.

E = matrix describing random error in the responses (for example due to careless errors, lucky guess, any distracters present when responding to the questions asked, …)

The underlying assumption to this model is that when elements, say people, respond to a question, they solicit their relevant resources (all constructs that may determine the response, both constructs of interest like “knowledge” on the item, but also constructs of disinterest like reading ability or fatigue) to generate a “true” response to that question. This “true” response is conceived as the actual rating or judgment made by the element, which, by the way, is assumed to be made on interval level. However, this judgment is assumed to be translated into the actual response in a process where systematic (modeled by the bias function $b_{\mathbf{C}}$) and random error (modeled by **E**) distort the response and where the resulting response is matched to one of the presented response categories.

An interesting result of this approach is that when the sample size N = population size, then responses are obtained for the complete population. So, the obtained results from a sample can always be compared to those for the entire population.

CONCLUSION

The presented learning environment can be used in class demonstrations as well as for problem based learning approaches and student assignments. Teachers can define populations and can then easily (and repeatedly) generate new population data. Due to the generic nature of the environment very diverse kinds of situations can be simulated, adapted to the interests of the students. Examples include psychological testing, demographic research, marketing studies, and many more. Also, the complexity of the simulated data can be very different depending on the number of constructs, the number and the diversity of the strata considered, and the kind of items. Moreover, bias in the simulated data can be turned on or off, allowing to further increase/decrease realism or simplicity.

Although the teacher’s job is not an easy one, requiring both effort and substantive knowledge in order to construct cases that are experienced as “realistic”, there is not a true alternative because in practice populations can only rarely be studied as intensively as in this learning environment. Only a very intensive study where the same participants would be studied repeatedly with ever new though similar questionnaires could yield similar, though probably much smaller and static data. The hands-on experience offered by this learning environment with always slight adaptations to the developed method is invaluable to obtain a deep understanding of many aspects of survey research.

In order to support the teachers in setting up learning experiences where specific points on survey research can be demonstrated further development of the environment is required. In the near future cases will be developed specifically for demonstrations on specific issues in respondent behavior, sampling, questionnaire design, etc.

ACKNOWLEDGEMENTS

The participating institutions Vrije Universiteit Brussel, Universiteit Gent, and Katholieke Universiteit Leuven and Hogeschool Gent sponsored the development and implementation of the learning environment. Co-financing was obtained from the Federal Ministry of Education.

REFERENCES

Hayduk, L.H. (1987). *Structural equation modeling with LISREL: Essentials and advance*. Baltimore: The Johns Hopkins University Press.

Bruner, J.S. (1961). The act of discovery. *Harvard Educational Review, 31*, 21-32.

Bruner, J.S. (1990). *Acts of meaning*. London: Harvard University Press.

Dewey, J. (1916). *Democracy and education*. New York: The Free Press.

European Round Table of Industrialists (1997). *Investing in knowledge: The integration of technology in European education*. Brussels.

Gijselaers, W.H. (1999). Is PGO millenniumproof? *Renovatie en innovatie van een succesvol onderwijsmodel Tijdschrift voor hoger onderwijs, 17,* 69-82.

Harel, I., & Papert, S. (Eds.) (1991). *Constructionism*. New Jersey: Alex Publishing Corporation.

Harré, R., & Gillett, G. (1994). *The discursive mind*. Thousand Oaks: SAGE.

Jöreskog, K.G. (1973) A general method for estimating a linear structural equation system. In A. S. Goldberger and O.D. Duncan (Eds.), *Structural equation models in the social sciences.* New York: Seminar Press/Harcourt Brace.

Kirschner, D., & Whitson, J.A. (Eds.) (1997). *Situated cognition*. London: Lawrence Erlbaum.

Mandl, H., & Reinmann-Rothmeier, G. (Eds.) (1995). *Unterrichten und Lernumgebungen gestalten*. Weinheim: Beltz.

Norman, G.R., & Schmidt,H.G. (1992). The psychological basis of problem-based learning: A review of the evidence. *Academic Medicine*, *67*(9), 557-565.

Wertsch, J.V. (1991). *Voices of the mind: A sociocultural approach to mediated action*. Cambridge, Massachusetts: Harvard University Press.