Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

AGENT BASED LEARNING ENVIRONMENT FOR SURVEY RESEARCH

Jayendran Venkateswaran
Sayli Shiradkar
Deepak Choudhary

Industrial Engineering and Operations Research
Indian Institute of Technology Bombay
Mumbai 400076, INDIA

ABSTRACT

Survey-based research methodology is commonly used in various disciplines ranging from social sciences to healthcare. However, considering practical constraints, it is difficult to provide real world experience of survey sampling methodologies to students and novice researchers. In this paper, we propose development of a virtual learning environment based on agent modeling to help learn different aspects and challenges in survey-based research. A study scenario of adoption of an improved cookstove is developed as an agent model, where each household (sample point) is an agent. The agent's behavior is defined using statecharts and system dynamics models. The agent-based environment has been used to illustrate various learning points for students and novice researchers.

1 INTRODUCTION

Survey-based research is a research method that involves the use of standardized questionnaires to collect information (data) from a sample of individuals in a systematic manner (Bhattacherjee 2012; Ponto 2015). The surveys can be quantitative (e.g., numerical questions or questions on Likert scale), qualitative (e.g. open-ended questions), or mixed. Among these, quantitative survey research is quite popular, especially in the social sciences, management and healthcare. The popularity has primarily been due to the fact that the numeric data obtained from the surveys can be quantitatively analyzed using statistical tools, including descriptive analysis and inferential analysis (Bhattacherjee 2012; Blackstone 2012).

Typically, a quantitative survey research involves the following stages (Bhattacherjee 2012; Blackstone 2012): research design (problem scoping, defining the hypothesis), questionnaire design (conceptualization, operationalization, framing effective questions), sampling, data collection (conducting the survey), and analysis (data coding, cleaning and statistical analysis). Typically, surveys are conducted in predetermined locations with participants, and the collected data is analyzed by employing various statistical techniques such as linear regression to understand the relationship between output and input variables (Urpelainen and Yoon 2015; Trani et al. 2017). The reader is kindly referred to social science research methods books such as by Bhattacherjee (2012) or Blackstone (2012) for details.

Now, quantitative surveys, when properly designed and executed can help provide data that is accurate, meaningful, and generalizable (Chen and Eisenberg 2020). Such data, when properly analyzed can help derive meaningful insights. However, it is not without methodological challenges that includes, designing effective questions & questionnaire design, appropriate and representative sampling, cost & time to conduct survey, managing lack of responses in samples, data cleaning, use of appropriate statistical methods, etc

(Chen and Eisenberh 2020; Wagner et al. 2019; Ponto 2015). Various strategies are also proposed to mitigate or overcome the challenges posed. These strategies include use of pretest questions & graphics in questionnaires, appropriate sampling techniques such as stratified & multimode sampling to ensure adequate coverage and representation, and so on (Chen and Eisenberh 2020; Wagner et al. 2019; Ponto 2015). Understanding and navigating through various methodologies to conduct an effective survey-based research can become quite daunting especially for students and novice researchers. These skills (in conducting effective survey research) get honed only by hands-on practical experience in the real world, over the years. This has motivated the authors to ask themselves: Can we develop a simulation-based virtual environment, for researchers to use and hone their survey research skills? Can such a simulation environment be helpful for learning, to better understand the strengths and weaknesses of quantitative survey research?

A review of literature revealed very few works in this area. Gilbert (1978) had presented a simulation approach for survey sampling where an interactive package was developed that allowed to compare different survey methods like random sampling, stratified, quota and cluster sampling. Chang et al. (1992) designed a computer program simulating samples drawn from the synthetic population of a country offering flexibility to students to choose sampling methods. Sample output is then analyzed by students in a different statistical software. This program was designed for a course on sampling methodology.

In this paper we present our preliminary work on developing an agent-based model of a community adopting a technology intervention (improved cookstove), and illustrate its potential use for reinforcing learning points for students / novice researchers of survey-based research. It is noted that there are several works related with analysis of simulation output data, simulation metamodel, etc (Santos and Santos 2007; Gore et al. 2017; Mertens et al. 2017). However, these approaches are not directly applicable since the purpose of this paper is not to analyze the simulation output, but rather use simulation as a virtual learning environment. In this paper a basic regression model only is considered to help illustrate the use of the proposed virtual simulation environment by students to help learn survey-based research.

The rest of the paper is organized as follows. In Section 2 we describe the proposed virtual learning environment and a list of possible learning points. Section 3 presents the details of the virtual world that is modeled (i.e., technology intervention and adoption). Section 4 presents the details of the agent-based model, the sampling technique involved, statistical method (logistic regression) used for sample data analysis, and the implementation details. Sections 5 presents the different scenarios, which are used in Section 6 to illustrate the various learning points. Section 7 presents the conclusion and future work.

2 VIRTUAL LEARNING ENVIRONMENT

2.1 The Agent-Based Simulation

At the core of the virtual learning environment is an agent-based model of the community. Agents, simply defined, are autonomous decision-making units. Each agent can be defined by their properties and their actions or behavior (Wilensky and Rand 2015). The behavior may be based on their interaction with other agents and/or cognition of their environment. An agent-based model, allows us to understand and analyze the emergent behavior of the system. In survey-based research, the information from a sample of individuals are collected to understand their preferences, thoughts, and behaviors in a systematic manner (Bhattacherjee 2012). Hence an agent-based modeling approach provides the best suited platform, where each individual (person, household or institution) to be surveyed, can be represented as an agent.

The agent-based model thus built can be used by researchers (or students or learners) as their "survey location". That is, the researcher can conduct the survey by sampling agents from the model during the simulation run. The responses or sample data collected can then be subject to appropriate statistical analysis to make inferences about the population. Now, the primary advantage of using simulation is that the entire simulation (all agents) data can be used to estimate the 'population' characteristics. Thus, a comparison of the sampling-based inferences (from survey) against the 'population' characteristics (from simulation) can

be used to readily reinforce key concepts in survey-based research. Figure 1 illustrates this interaction between the researcher and the virtual world. It is noted that the agent-based model referred here is not intended to replace field surveys, but to capture the behavior of individuals within a community in sufficient detail to reinforce learning goals.

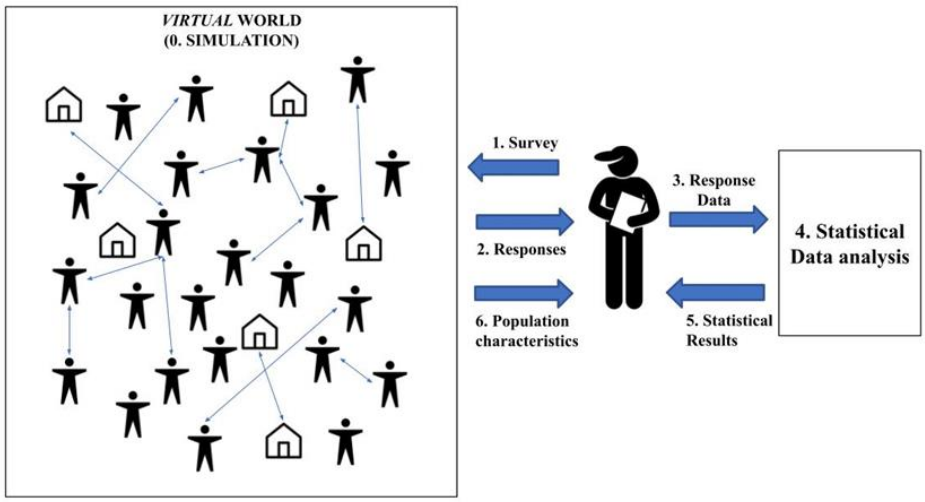

Figure 1: Proposed virtual worlds framework.

2.2 Learning Points

Learning points (LPs) or goals refers to specific concepts/ topics/ ideas that a student is supposed to learn about. Virtual learning environments can be effectively used when tied to specific learning points. In this paper, the following LPs are demonstrated using the proposed virtual learning environment.

- LP1: *How soon should we measure (survey) the impact of an intervention*? Suppose one plans to conduct a survey to measure the long-term impact, or make a statement about the success/ failure of an intervention. In this case care must be taken to ensure sufficient time has elapsed from the time of intervention so that the steady state dynamics are captured. This can be demonstrated by surveying the same sample (individuals) at different points in time, analyzing them using statistical tools and comparing the observations.
- LP2: *What is Type I and Type II error*? This is a classical question faced by all students working in inferential statistics. Type I error, or false positives, refers to the case when we conclude (based on samples) a factor to be significantly affecting the outcome when it is not. Type II error, or false negatives, refers to the case when we conclude (based on samples) a factor to *not* significantly affect the outcome when it actually does. This can be demonstrated by comparing the statistical results of the samples with the results from the population itself.
- LP3: *Can we identify factors that weakly impact the outcome in reality*? When conducting survey-based research we reasonably assume and actively look for factors having strong correlations. Factors with weaker correlations may be difficult to discern. This can be demonstrated by creating 'strong correlation' and 'weak correlation' scenarios in the virtual world, sample the scenarios and compare the statistical results.
- LP4: *Are these factors distinct or combined*? This refers to the situation when decisions in reality are made by aggregating certain attributes of the individual (sample), but when sampling and

subsequent statistical analysis, the attributes were estimated separately. This is demonstrated by creating scenarios in the virtual world where the impact is based on, say, the sum of attributes, but the analysis is done by taking the attributes separately.

It is noted that the above LPs (not exhaustive list) are listed as per the experience of authors in designing and conducting surveys, data analysis, interacting with students, and commonly mentioned challenges/ examples in various books (Bhattacherjee 2012; Blackstone 2012).

3 STUDY CONTEXT

In this paper the scenario of adoption of improved cookstoves (ICS) by rural households is considered for modeling and analysis. There are numerous studies on uptake and adoption of improved cookstoves. Some studies focus on adoption of ICS offered through a specific program/scheme. Relationships between dependent variables and independent variables are established using regression models. A study conducted on the ICS adoption in Northwest Ethiopia, sampled about 10% of the population for conducting a quantitative survey at a single time point (Adane et al. 2020). ICS adoption status was an independent variable measured as a binary variable (Yes/No). Dependent variables were: household and settings related (gender and education of head of HH, family size, number of rooms, ownership, fuel use, multiple cookstoves), Cookstove technology related (fuel processing, health benefit, safety benefit, time-saving benefit), user's knowledge and perception related (social interaction, demonstration of stove, traditional suitability) and financial related (stove price and availability). Logistic regression analysis of the survey data showed that gender, education level, house ownership, location of kitchen, source of fuel, fuel processing, durability, optimistic previous social interaction, health and fuel saving benefits, live demo, price and availability were significant factors in adoption of ICS (Adane et al. 2020). Krishnapriya et al. (2021) analyzed the survey data obtained from six countries to estimate the impact of ICS on time spent on cooking. Linear regression was utilized where outcome variable was time spent on cooking and dependent variables covered household characteristics such as the wealth index, household size, gender, education level and age of the household head, number of adults and children, the primary cook's age and education, the empowerment index for the female respondent and primary cookstove (Krishnapriya et al. 2021). Another study utilized a longitudinal survey method (survey conducted with same participants over different time points (Jueland et al. 2020). Logistic regression was employed with independent variables as adoption of ICS by HH and dependent variables were socio-economic factors like education level of head of household and primary cook, caste, number of children, number of rooms, land owned, monthly expenditure, loan, community/ relationship with NGO. Number of children, education, number of rooms, monthly expenditure, relationship with NGO were found to be significant variables in adoption of ICS.

Our virtual environment is constructed based on the findings from literature. The agents (sample points) are the households (HHs). The key outcome variable of interest is the 'adoption of ICS', which is measured by the, "*Is ICS the primary cookstove for the household?*" as a binary value. The answer "Yes" represents adoption (value 1) and "No" (value 0) otherwise. The independent variables are demographic factors - Number of rooms in the household, number of children, number of adults, age of primary cook, education level of primary cook and house type.

4 MODEL & ANALYSIS METHODOLOGY

This section describes the scenario of adoption of an improved cookstove as modeled in the proposed virtual environment.

4.1 ABMS Based Model

Agent-based model for the adoption of improved cookstoves (ICS) is developed. Figure 2 presents the framework of the agent model, the attributes and their behavior. The (virtual) geographical region consists of 3000 households scattered across multiple villages. Each household is defined by their attributes such

as, number of rooms in the household, number of children, number of adults, age of primary cook, education level of primary cook and house type. Each HH can be in either of 3 states, 'Use TCS only' to indicate that the HH uses only traditional cookstove; 'Use TCS & ICS' to indicate that the HH uses traditional and ICS cookstove; and 'Use ICS only' to indicate that the HH uses only ICS. Initially all the HHs have traditional cookstoves (state 'Use TCS only'). Upon intervention, HHs are provided with new improved cookstoves. At this stage, HHs have access to both traditional and improved cookstoves. The fraction cooking needs of HH that are satisfied with ICS is captured using a system dynamics (SD) stock-flow diagram model. Each HH has an independent 'goal for ICS usage' attribute that refers to the maximum fraction of cooking the HH is willing to shift to the improved cookstove. HHs also have a 'Average Adjustment time' attribute that defines how long on an average it takes to reach the goal. The actual Fraction cooking in ICS is modeled as a stock for each HH, that is adjusted using a simple negative feedback loop to reach the goal. It is noted that the goal is NOT accessible to the surveyor. Thus, the adoption behavior of agents dynamically changes over time. The agent attributes are static. In the model, the HHs' attributes are initialized at random, as per the settings given in Table 1. It is noted that the attribute 'Goal of the ICS usage' of HHs is determined based on certain conditions on the other attributes of the HH (see Section 5).

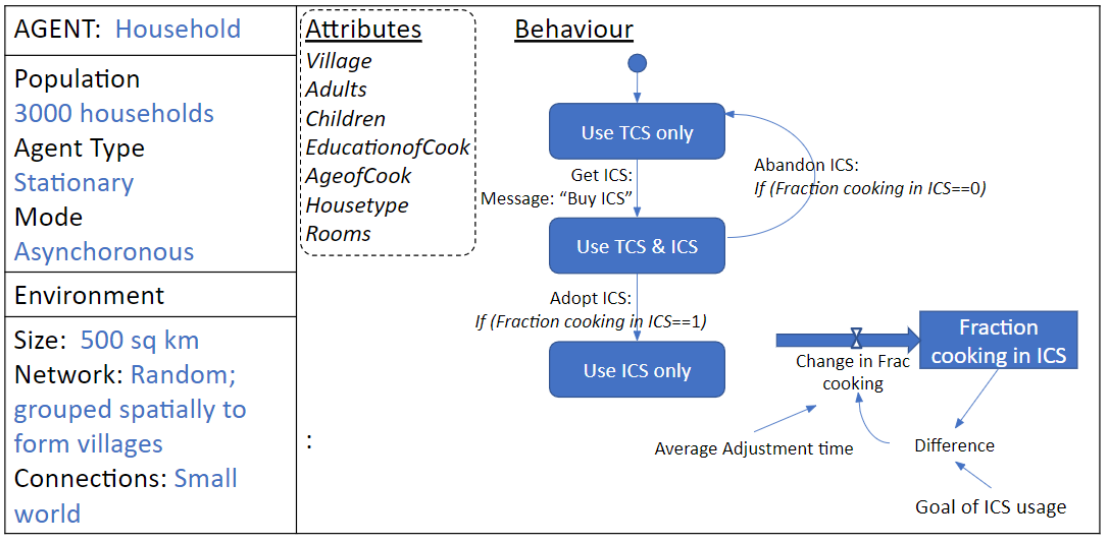

Figure 2: Canvas capturing the agents, its attributes and behavior.

4.1 Sampling

Literature on field surveys suggests utilizing maximum up to 10% of the population as a sample (Adane et al. 2020; Gilbert 1978). Thus, in a virtual learning environment, 10% of the population is sampled as a survey. In this preliminary work, we use a random sampling method. Five sets of 300 households are selected through random sampling from four villages for each scenario. The following data is sampled for each HH: the HH (or agent) ID, 'village', 'number of rooms', 'number of adults', 'number of children', 'age of cook', 'education level of cook', 'house type' and 'Is primary cookstove ICS?'. The last data point of 'Is primary cookstove ICS?' is provided by the HH as follows. If at the time of sampling, the value of stock 'fraction of cooking in ICS' is >= 50% then the answer value 1 is returned by the HH, else it is taken as 0. Kindly note that this is done irrespective of the value of the goal of the HH.

Table 1: Variables definitions in virtual world of ICS adoption.

Attributes	**Values**	**Notes**
Village	Discrete Uniform[1, 4]	4 villages are modeled
Number of Adults	Discrete Uniform[2, 6]	Adult is a person with age >=18
Number of Children	Discrete Uniform[0, 4]	Children is a person with age <18
Education Level	Discrete Uniform[1, 4]	Education is coded as None (1); below or up to class 8 (2); class 9 to class 12 (3); graduate and above (4)
House type	Discrete Uniform[1, 2]	Kachha (made of dung, straws) (1) Pakka (made of cement, bricks, concrete) (2)
Number of Rooms	Discrete Uniform[1, 4]	
Age of Cook	15 + Poisson(15)	Primary cook's age is minimum 15 years
Average Adjustment Time	Discrete Uniform[2, 6]	

4.2 Analysis of Samples using Regression

In order to understand the relationship between adoption of ICS and the various predictor variables a logistic regression on binary outcome is typically derived. The logistic regression model (Newson 2012) investigates transformation of outcome variable Y_i, repeatedly measured at various time points for each household. The logistic regression equation is as follows

$$logit(Yi) = \alpha + \beta * Xi + \varepsilon i$$

where Yi is the outcome variable estimated for each household, α is the intercept, β are the regression coefficients, Xi are the predictor variables, and εi is the error term. The complete logistic regression equation used in the analysis of ICS adoption in our paper is as follows:

$logit$(Adoption of ICS) = α + β0* Village + β1 * Number of rooms + β2 * Age of cook + β3 * Education level + β4* Number of children + β5 * Number of adults + β6* House type + ε

Here, the dependent variable is 'Adoption of ICS' and the independent variables are Village, Number of rooms, Age of cook, Education level, Number of children, Number of adults and House Type. Village, education level and house type are categorical variables.

4.3 Implementation

The agent-based model has been implemented in Anylogic, and the output (samples) were exported to comma separated files. These samples were then analyzed using *R* software, where the *glm*() function was used to fit a logistic regression model against various predictor variables of interest.

5 SCENARIOS

As mentioned in Section 4.1, the 'Goal of the ICS usage' is a key attribute of each household. This attribute is assumed to be determined by each household as per their attributes. In order to represent various situations

that may occur in reality, different scenarios are created. Here, a scenario refers to the choice of rules used to determine the 'Goal of ICS usage' attribute of the HH. The general rules-construct that defines a scenario is as follows:

If (condition based on HH attributes is TRUE) then
    'Goal of ICS usage' is drawn from probability distribution $f(x)$
Else
    'Goal of ICS usage' is drawn from probability distribution $g(x)$

The above rule/condition is triggered when the household initially takes up the ICS. A full factorial design experiment is considered based on three factors: the IF-condition (4 levels), function $f(x)$ (2 levels) and function $g(x)$ (2 levels). Thus, a total of 16 different scenarios are generated by using different conditions used in the IF-statement, and/or different functions for $f(x)$ and $g(x)$, as summarized in Table 2. Care has been taken to ensure that there is inherent randomness in setting the goal of usage. Also, it is noted for each scenario, the sample HHs were 'sampled' at month 3, month 6 and month 12 from the time the ICS was introduced in the HH.

6 EXPERIMENTS & RESULTS

In this section, we discuss the various learning points (presented in Section 2.1), based on the results obtained through the virtual environment, under the scenarios presented in Section 5. The key results discussed below are about which factors are found to be significant based on the logistic regression of the samples. These sample-based regression results will be compared against the results obtained by fitting a regression for the entire population data (taken at month 24). For each scenario, 5 sample sets (replications) were taken and the regression carried out. All experiment files included the data files can be accessed from the site https://bit.ly/3RjrEqD.

6.1 LP1: Timing of the Survey

Table 3 presents the regression results for data sampled (one sample set only) for 8 scenarios in months 3, 6 and 12. In the Table, NS stands for non-significant variables while '*' symbol represents a significant variable. In case of scenario 1, 'House Type' is a significant variable when data is obtained at 3 months. However, survey data at 6 months and 12 months show 'Age of Cook' and 'House type' as strongly significant variables. For month 3 survey data, 'Age of Cook' and 'House type' were not found to be significant for scenarios 2 to 8. Except scenarios 4 and 8, both or either of the variables 'Age of Cook' and 'House type' are found to be significant with respect to time. This observation shows that relationships between variables can vary with respect to time.

In case of adoption of ICS, the goal of the ICS usage is set initially for all households, but it takes some months for the HHs to reach that goal. The time to reach the goal is also different for each household as given in Table 1. Thus, if a survey is conducted without giving sufficient time for the community to adjust/ adopt the new technology, the conclusions from the surveys will be incorrect.

Table 2: Scenarios in the virtual world.

Scenario	**Condition**	$f(x)$	$g(x)$
1	(House Type ==2 AND Age of Cook <= 30)	U(0.4, 1)	U(0, 0.7)
2	(House Type ==2 AND Age of Cook <= 30)	U(0.4, 1)	U(0,0.9)

3	(House Type ==2 AND Age of Cook <= 30)	U(0.2, 1)	U(0,0.7)
4	(House Type ==2 AND Age of Cook <= 30)	U(0.2, 1)	U(0,0.9)
5	(House Type ==2 OR Age of Cook <= 30)	U(0.4, 1)	U(0,0.7)
6	(House Type ==2 OR Age of Cook <= 30)	U(0.4, 1)	U(0,0.9)
7	(House Type ==2 OR Age of Cook <= 30)	U(0.2, 1)	U(0,0.7)
8	(House Type ==2 OR Age of Cook <= 30)	U(0.2, 1)	U(0,0.9)
9	(House Type ==2 AND Age of Cook <= 30 AND children+ adults => 6)	U(0.4, 1)	U(0, 0.7)
10	(House Type ==2 AND Age of Cook <= 30 AND children+ adults => 6)	U(0.4, 1)	U(0,0.9)
11	(House Type ==2 AND Age of Cook <= 30 AND children+ adults => 6)	U(0.2, 1)	U(0,0.7)
12	(House Type ==2 AND Age of Cook <= 30 AND children+ adults => 6)	U(0.2, 1)	U(0,0.9)
13	(House Type ==2 OR Age of Cook <= 30 OR children+ adults => 6)	U(0.4, 1)	U(0,0.7)
14	(House Type ==2 OR Age of Cook <= 30 OR children+ adults => 6)	U(0.4, 1)	U(0,0.9)
15	(House Type ==2 OR Age of Cook <= 30 OR children+ adults => 6)	U(0.2, 1)	U(0,0.7)
16	(House Type ==2 OR Age of Cook <= 30 OR children+ adults => 6)	U(0.2, 1)	U(0,0.9)

Table 3: Comparison of results from samples surveyed in months 3, 6 and 12.

	Scenario 1	Scenario2	Scenario 3	Scenario4	Scenario5	Scenario6	Scenario7	Scenario 8
Factors	**Results when Survey done at 3 months from ICS introduction**							
AgeofCook	NS	NS	NS	NS	NS	NS	NS	NS
HouseType	*	NS	NS	NS	NS	NS	NS	NS
	Results when Survey done at 6 months from ICS introduction							
AgeofCook	***	*	***	NS	***	*	*	NS
HouseType	***	*	***	NS	***	NS	*	NS
	Results when Survey done at 12 months from ICS introduction							
AgeofCook	***	*	*	NS	***	*	*	NS
HouseType	***	***	***	NS	***	*	NS	NS

*** indicates p-value < 0.001, ** indicates p < 0.01, * indicates p < 0.05, NS indicates Not Significant

6.2 LP2: Understanding Type I and Type II Errors

Table 4 presents the results obtained for Scenarios 1 through 8, when regression was carried out for all 5 sets of sample data obtained from month 12. The values in the table indicate the difference between the percentage of sample sets (out of 5) where a particular factor (predictor variable) was found to be significant, and the expected percentage based on population regression. For example, the negative value.

say, -20% indicates the corresponding factor was not found to be significant in 1 out of the 5 sample sets of that scenario, when it ought to have been. This indicates False Negatives. Likewise, the positive value, say, +40% indicates that the corresponding factor was found to be significant in 2 out of the 5 sample sets of that scenario, when it ought not to have been. This indicates False Positives. Blank cells indicate that a nonsignificant factor was indeed not found to be significant in any of the sample sets. This indicates True Negatives. The '***' instead of value indicates that the significant variable was indeed found to be significant in all the sample sets. This indicates True Positives. Table 4 also has color coding to provide visual clarity: cell color Blue indicates some samples were False Positive; Red indicates False Negatives; Green indicates samples that reflect population results.

Table 4: Fraction of Type I and Type II errors based on samples surveyed in month 12 (all 5 sample sets) for all scenarios.

Factors	**Results based on population data**	**Results based on samples** (Blue: False Positives, Red: False Negatives)							
		Scn 1	**Scn 2**	**Scn 3**	**Scn 4**	**Scn 5**	**Scn 6**	**Scn 7**	**Scn 8**
Village	Not Significant in all scenarios	20%		20%				20%	
Rooms	Not Significant in all scenarios	40%	20%	40%	20%				
Adult	Not Significant in scenarios 1, 2, 3, 6; Significant in scenarios 4, 5, 7, 8	20%	20%	20%	-80%	-100%		-80%	-80%
Children	Not Significant in all scenarios								
Age of Cook	Significant in all scenarios	***	-20%	-20%	-60%	***	-20%	-20%	-100%
Education Level	Not Significant in all scenarios	40%	40%	40%	60%	20%	60%	40%	40%
House Type	Significant in all scenarios	***	***	***	-20%	***	***	-20%	-60%

As can be observed from Table 4, there are many instances when we get false positives (type I error). For example, education level was found to be a significant factor (falsely) in many instances based on the sample. Sometimes we get false negatives (type II error), which could be more harmful. For example, the age of cook was not found to be significant (type II error) in many instances. It is especially noted that it missed it 100% of the time in Scenario 8. These results are just by chance. The purpose of this table is to explicitly illustrate the same, highlight the need for a better sampling strategy, and caution novice / students new to the field to not completely trust the sample data collected & its analysis.

6.3 LP3: Weaker Correlations Are Harder to Discern

In designing the scenarios, two distributions $f(x)$ and $g(x)$ were used to generate the goal depending on the condition (see Section 5): $f(x)$ is either U(0.4, 1) or U(0.2, 1), and $g(x)$ is either U(0, 0.7) or U(0, 0.9). Suppose $f(x)$ is U(0.4, 1) then the HH will not take any values less than 0.4 when the If-condition is true. However, if $g(x)$ is also U(0, 0.9) then the HH can have a high value of the goal (>0.5) even if the condition is false. Table 5 presents the correlation between each of the factors and the 'Adoption of ICS' (0 or 1) for the population of 3000 data. The highlighted columns in blue represent the relatively highly correlated variables with the output 'Adoption of ICS'. Each pair of scenarios (1 and 4), (5 and 8), (9 and 12), (13 and 16) has the same if-condition. Now, as seen from the table, for the Age of Cook the correlation reduces from scenario 1 to scenario 4, since in scenario 4 the overlap between $f(x)$ and $g(x)$ is larger. Similar observations can be made between scenarios 5 and 8; scenarios 9 and 12 and scenarios 13 and 16. Also, it is intuitive that as the correlation between the factor and the output (adoption) reduces, it becomes difficult to discern the same based on samples.

Table 5: Correlation among the attributes of the population of 3000 HHs under different scenarios.

	$f(x)$	$g(x)$		**Village**	**Rooms**	**Adults**	**Children**	**AgeofCook**	**EduLevel**	**House type**
Scenario 1	U(0.4, 1.0)	U(0, 0.7)		0.02	-0.02	-0.01	0.01	**-0.23**	0.01	**0.28**
Scenario 4	U(0.2, 1.0)	U(0, 0.9)		0.01	0.00	-0.04	-0.02	**-0.09**	0.02	**0.08**
Scenario 5	U(0.4, 1.0)	U(0, 0.7)		-0.01	0.01	-0.02	-0.02	**-0.24**	-0.01	**0.24**
Scenario 8	U(0.2, 1.0)	U(0, 0.9)		-0.01	0.00	-0.03	-0.01	**-0.09**	0.01	**0.05**
Scenario 9	U(0.4, 1.0)	U(0, 0.7)		0.01	-0.01	0.07	0.09	**-0.16**	0.00	**0.17**
Scenario 12	U(0.2, 1.0)	U(0, 0.9)		0.01	0.00	-0.01	0.00	**-0.06**	0.02	**0.04**
Scenario 13	U(0.4, 1.0)	U(0, 0.7)		-0.04	-0.01	0.04	0.08	**-0.10**	0.00	**0.08**
Scenario 16	U(0.2, 1.0)	U(0, 0.9)		-0.02	0.00	-0.01	0.02	**-0.03**	0.02	**-0.01**

When performing regression using the sample sets, one can expect the sample data to perform better in identifying the significant variables in Scenario 1 or 5 as opposed to Scenario 4 or 8. This is shown in Table 6. Only the significant factors in the population, Age of cook and House type are shown. The negative value, say, -20% indicates the corresponding factor was not found to be significant in 1 out of the 5 sample sets of that scenario, when it ought to have been. This indicates False Negatives or Type II error. It is noted that Scenarios 1 and 4 use the same condition "(House Type ==2 AND Age of Cook <= 30)". As seen from table, in Scenario 1, when $f(x)$ and $g(x)$ have smaller overlap, the sample-based analysis was able to correctly identify the significant factors in all sample sets. On the other hand, in Scenario 4, when $f(x)$ and $g(x)$ have larger overlap, the sample-based analysis was unable to identify the significant factors (type 2 error) in some or all sample sets. A similar observation is made between Scenarios 5 and 8, both of which use the condition "(House Type ==2 OR Age of Cook <= 30)". The key observation here is that it becomes harder to distinguish 'weak' influencers based on the samples.

Table 6: Regression results for Scenarios 1 & 4; and Scenarios 5 & 8.

Factors	**Scn 1**	**Scn 4**	**Scn 5**	**Scn 8**
Age of Cook	***	-60%	***	-100%
House Type	***	-20%	***	-60%

6.4 LP4 Combined Effects may be Missed unless Looking for it

In order to illustrate this learning point, the following experiment is conducted. The if-condition to set the goal of ICS usage in households is modified to include the sum of adults and children in the household (see scenarios 9 to 16 in Table 2) to indicate that the goal of ICS usage is affected by the total people in the HH. However, this information is unclear when analyzing based on the sample, where the regression model uses 'Adults' and 'children' as distinct independent variables. Table 7 presents the regression results for Scenarios 9, when Adults and children are taken as separate variables, and when 'Adults + Children' are taken as a single variable. Only the significant factors in the population: Adults, Children, Age of cook and House type are shown. The values indicate the percentage of sample sets (out of 5) where a particular factor (predictor variable) was found to be significant. '***' indicates the factor was significant in all the sample sets. The results show that when the factors (Adults and children) were taken separately, they were both found to be significant in only 40% of the sample sets. However, when combined as an 'Adults+ children' single variable, it was found to be significant in 80% of the sample sets. This becomes a tricky issue to handle.

Table 7: Regression results for Scenario 9, without and with combined factors.

Scenario 9			
Factors		**Factors**	
Adults	40%	Adults + children	80%
Children	60%		
Age of Cook	80%	Age of Cook	80%
House Type	***	House Type	***

7 CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a virtual learning environment, based on agent-based simulation, that can be used as an exploratory and explanatory tool for survey-based research. Preliminary work was done to show how the agent-based environment can be used to model and illustrate various learning points for students and novice researchers. The agent model was based on the adoption of an improved cookstove (technology solution) among households in a region. The behavior of the households (agents) was determined by system dynamics models, with other attributes initialized randomly.

Typically, it is difficult to provide real world experience of survey sampling to all students considering practical constraints. This virtual world has the potential to act as an effective tool in providing hands-on experience through simulated scenarios. The learning points also can be further expanded to include different sampling methods, descriptive statistics, longitudinal studies, etc. Future research in many other dimensions are also required, including, measuring the learning uptake among students, making the community dynamics more realistic by including social interactions, improved user interface etc.

REFERENCES

Adane, M. M., G. D. Alene, S. T. Mereta, and K. L. Wanyonyi. 2020. "Facilitators and Barriers to Improved Cookstove Adoption: A Community-Based Cross-Sectional Study in Northwest Ethiopia". *Environmental Health and Preventive Medicine* 25(1):1–12.

Bhattacherjee, A. 2012. *Social Science Research: Principles, Methods, and Practices*. Global Text Project. http://scholarcommons.usf.edu/oa_textbooks/3, accessed 19th April 2022.

Blackstone, A. 2012. *Principles of Sociological Inquiry – Qualitative and Quantitative Methods*. Saylor Foundation. https://open.umn.edu/opentextbooks/textbooks/139 accessed 1st April 2022.

Chang, T. C., S. L. Lohr, and C. G. McLaren. 1992. "Teaching Survey Sampling Using Simulation". *The American Statistician* 46(3):232–237.

Chen, T., and M. L. Eisenberg. 2020. "Challenges in Survey Based Research". *The Journal of Sexual Medicine* 17(11):2115–2117.

Gilbert, G. N. 1978. "A Simulation Approach to Teaching Survey Sampling". *Teaching Sociology* 5(3):287-294.

Gore, R., S. Diallo, C. Lynch, and J. Padilla. 2017. "Augmenting Bottom-up Metamodels with Predicates". *Journal of Artificial Societies and Social Simulation* 20(1):4.

Jeuland, M. A., S. K. Pattanayak, S. Samaddar, R. Shah, and M. Vora. 2020. "Adoption and Impacts of Improved Biomass Cookstoves In Rural Rajasthan". *Energy for Sustainable Development* 57:149–159.

Krishnapriya, P. P., M. Chandrasekaran, M. Jeuland, and S. K. Pattanayak. 2021. "Do Improved Cookstoves Save Time and Improve Gender Outcomes? Evidence from Six Developing Countries". *Energy Economics* 102, 105456.

Mertens, K. G., I. Lorscheid, and M. Meyer. 2017. "Using Structural Equation-Based Metamodeling for Agent-Based Models". In *Proceedings of the 2017 Winter Simulation Conference*, edited by V. W. K. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. H. Page, 1372–1382. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Newsom, J. T. 2012. "Basic Longitudinal Analysis Approaches for Continuous and Categorical Variables". In *Longitudinal Data Analysis*, edited by J. Newsom, R. N. Jones, and S. M. Hofer, 143–79. Routledge.

Ponto, J. 2015. "Understanding and Evaluating Survey Research". *Journal of the Advanced Practitioner in Oncology* 6(2):168–171.

Santos, I. R. and Santos, P. R. 2007. "Simulation Metamodels for Modeling Output Distribution Parameters". In *Proceedings of the 2007 Winter Simulation Conference*, edited by S. G. Henderson, B. Biller, M. H. Hsieh, J. Shortle, J. D. Tew, and R. R. Barton, 910–918. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Trani, J. F., P. Kumar, E. Ballard, and T. Chandola. 2017. "Assessment of Progress Towards Universal Health Coverage for People with Disabilities in Afghanistan: A Multilevel Analysis of Repeated Cross-Sectional Surveys". *The Lancet Global Health* 5(8):E828–E837.

Urpelainen, J and S. Yoon. 2015. "Solar Home Systems for Rural India: Survey Evidence on Awareness and Willingness to Pay from Uttar Pradesh". *Energy for Sustainable Development* 24:70–78.

Wagner, S., D. Mendez, M. Felderer, D. Graziotin, and M. Kalinowski. 2020. "Challenges in Survey Research". In *Contemporary Empirical Methods in Software Engineering*, edited by M. Felderer, and G.H. Travassos, 93–125. Springer Cham.

Wilensky, U. and W. Rand. 2015. *An Introduction to Agent-Based Modeling: Modeling Natural, Social, and Engineered Complex Systems with NetLogo*. Cambridge: Massachusetts: The MIT Press.

AUTHOR BIOGRAPHIES

JAYENDRAN VENKATESWARAN is an Institute Chair Professor in Industrial Engineering and Operation Research at the Indian Institute of Technology Bombay, India. His research and teaching interests are in modeling and analysis of complex socio-economic systems, systems modeling and simulation, last mile supply chain ecosystem, and energy access. He has several publications in leading journals and conferences to his credit. He has guided 7 doctoral students and 40+ graduate (Masters') students. His e-mail address is jayendran@iitb.ac.in. URL: https://www.ieor.iitb.ac.in/~jayendran

SAYLI SHIRADKAR is a research scientist in Industrial Engineering and Operation Research at the Indian Institute of Technology Bombay, India. She holds a Ph.D. degree in Industrial Engineering from the University of Central Florida. Her research interests include simulation-based training, modeling and simulation and data analytics. Her email address is sayli.bhide@gmail.com

DEEPAK CHOUDHARY is a research associate in Industrial Engineering and Operation Research at the Indian Institute of Technology Bombay, India. He holds a M.Tech degree in Signal Processing & Control from National Institute of Technology. His research interests include survey and operational data analytics and machine learning. His email address is deepak.c@soulsiitb.in.