Human-Computer Interaction

Week 11 Lecture 11A
Quantitative Evaluation

COMP 3900 & COMP 6390

Semester 2, 2016

Duncan Stevenson

Australian National University

Purpose of this lecture

Remind you about the basic concepts in quantitative evaluation

Use a published example (Virtual Reality for teaching surgery) to illustrate:

- The reasons you might use a quantitative method for evaluating an interactive system
- The ways you can gather quantitative data
- The types of preparation you might need in order to conduct a quantitative evaluation

Basic concepts

Comparing two systems

- Sometimes an older system and a newer system
  [Is the newer system better than the older system?]
- Sometimes two different ways of doing the same task
  [Is one way better than the other?]
  [Are both ways equally good?]
- Purpose is to make a decision [remember the toy football game experiment]
- You need to understand the systems well enough to properly choose the variables that you plan to measure

Basic concepts

The independent variable

- Typically you have <u>one</u> independent variable, which you set at the start of each run of the evaluation
- If we are comparing two systems (System A and System B) then the <u>choice of system is our independent variable</u>
- For example, if you have two systems to compare and a number of participants, then you might:
  - Get Participant 1 to use System A: everything else about this run of the evaluation is either controlled or is an outcome measurement relating to System A
  - Get Participant 2 to use System B: everything else about this run of the evaluation is either controlled or is an outcome measurement relating to System B
  - And so on.

Basic concepts

Controlled variables

- Controlled variables are anything that could potentially vary between different runs of the evaluation.
- Examples could include:
  Age, skills and limitations on the abilities of the participants.
  The approach, personality and language of the researcher who is leading the evaluation.
  The noise and lighting levels in the evaluation environment
  The data that you give the participant
- You need to understand the evaluation task well enough to identify all the important controllable variables

Basic concepts

Dependent variables

- Dependent variables are the things you measure during the evaluation
- Examples could include:
  Time taken to complete a task
  A measure of the success of a task
  The number of user actions (such as mouse clicks) to complete a task
  The number of errors the user makes

Basic concepts

Random selection and allocation of participants

- The statistical analysis of your results relies on you randomly selecting your participants from the larger population of potential users of the system you are evaluating.
- The statistical analysis also relies on you randomly allocating your participants to the two systems that you are comparing.

Basic concepts

“Within subjects” and “Between subjects”

In our vocabulary, a “subject” is a participant in our evaluation

- “Within subjects”: each participant uses both System A and System B
- “Between subjects”: each participant uses only one of the two systems
- You need to decide how you will design your experiment

Brief discussion of the reasons for deciding which method to use.

Basic concepts

Minimising variation in the measurements

- Standard ways of measuring the dependent (output) variables
- If different people are taking the measurements then comparing each person's set of measurements with the others to look for bias
- "Blind" measurement: the person doing the measurement does not know which independent variable is being used. This can apply to evaluating the skills from two different methods of training.

Basic concepts

Statistical analysis of the measurements

For example, if you are comparing the time taken to do a task with each of System A and System B, then you

- Compute the mean and standard deviation of the times taken for each participant on System A
  [Each participant will take a slightly different time to do the task, so the times will be spread around the mean]
- Compute the mean and standard deviation of the times taken for each participant on System B
- Use a statistical test to decide if the mean times for System A and System B are "significantly" different.

Basic concepts

Statistical test of significance

Continuing our example, the statistical test will compute the probability that the difference in the mean times taken for System A and System B is as large as it is.

- This probability is computed using the assumption that there is no difference in effect between System A and System B [which is the Null Hypothesis]
- If that probability is very small then either a very unlikely event has happened or we reject the assumption that there is no difference in effect between System A and System B (and therefore conclude that there is a significant difference in effect between System A and System B)

Virtual Reality training example

- The purpose of this evaluation was to demonstrate that Virtual Reality technology could be used to teach surgical techniques.
- Its purpose was to convince the American College of
- Surgeons to allow the use of Virtual Reality technology for teaching surgery.
- It used a specific example of surgery and a specific type of virtual reality training as a case study.

Randomised Controlled Trials (RCT)

Randomised Controlled Trials (RCT)are considered to be the “gold standard” for evaluating stable interactive systems where there are measurable outcomes from using those systems.

RCTs can be expensive to run, so you would normally only run an RCT on a system if there was real value in having the result.

Case study of a randomised controlled trial

This lecture looks in detail at a journal paper that reports on a randomised controlled trial of a Virtual Reality training system.

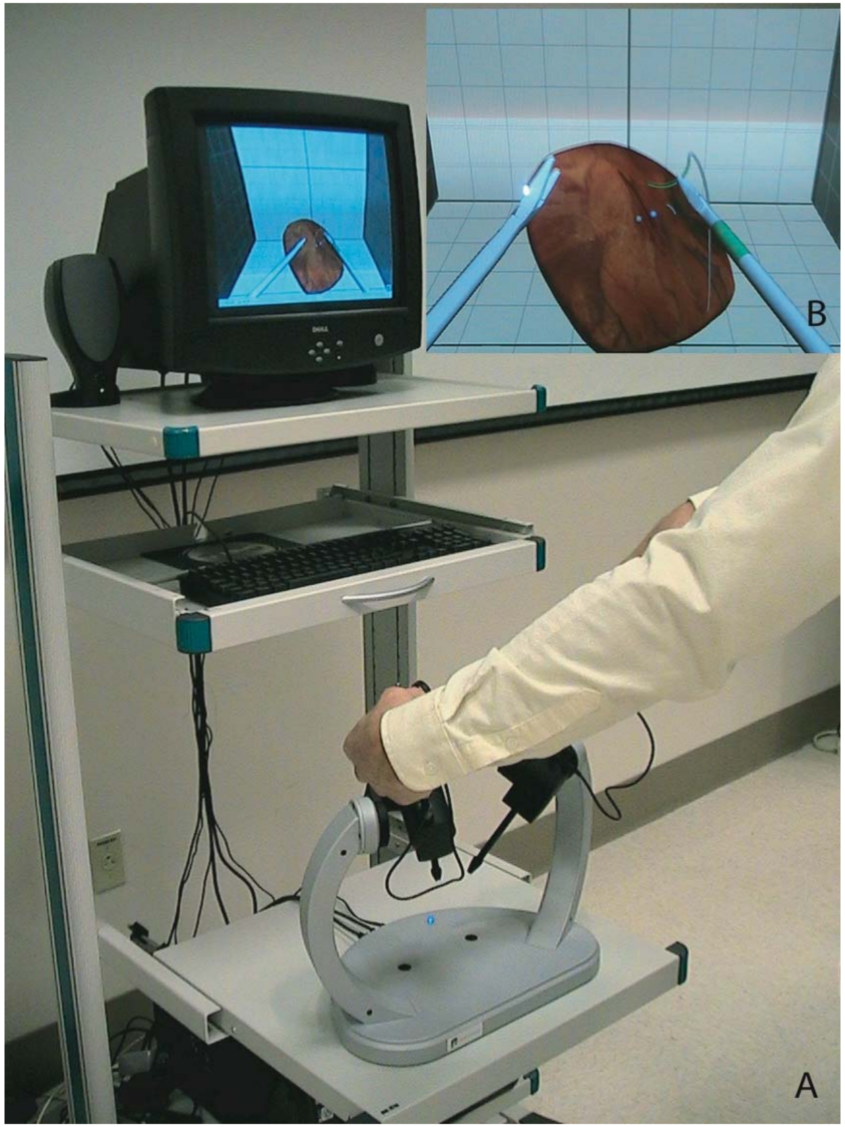
The trial compared the performance of two groups of junior surgeons.

- One group received conventional training (lectures, textbooks, observation in the operating room) – "control group"
- The other group received conventional training supplemented by Virtual Reality training – "experimental group"

There were several dependent (i.e. outcome) variables that were measured, including time taken and numbers of errors made.

Virtual reality training technology

MIST-VR trainer:
laparoscopytoday.com



Randomised controlled trial

Seymour, N., Gallagher, A. et al. (2002)

"Virtual Reality Training Improves Operating Room Performance: Results of a Randomised, Double-Blinded Study“

Annals of Surgery 236(4): 458-464

This is an example of a carefully managed experiment to demonstrate to a skeptical audience (the American College of Surgeons) that virtual reality training for surgeons could have a valuable place in the training curriculum.

The authors themselves were already convinced that their particular VR training situation would deliver a noticeable improvement to the trainee surgeons.

The journal paper is on pages 458-462. The last two pages are comments from other surgeons about the contents of the paper.

Randomised controlled trial

Objective

To demonstrate that virtual reality (VR) training transfers technical skills to the operating room (OR) environment.

The use of VR surgical simulation to train skills and reduce error risk in the OR has never been demonstrated in a prospective, randomized, blinded study.

Prospective: Creating the data as part of the study

Randomized: Randomly allocate the participants (surgical residents) to either the experimental group or the control group

Blinded: The surgeon assisting the trainee and the surgeons assessing the videotapes of the trainee's performance did not know the training status of the resident.

Randomised controlled trial

Comment

Prospective trial: you get to control the variables involved in whatever the trial is doing

Retrospective trial: you get data on activities that have already happened (but without your active control over the variables) and you assess that data. (Telehealth example – two years of virtual intensive care unit operating data was evaluated at the end of the project)

Controlled : There are two parts to the trial, a controlled part (where you run the task in the standard way) and an experimental part (where you run the task in the new way).

- We also refer to the “control group” (the participants doing the task in the standard way) and the “experimental group” (the participants doing the task in the new way).

Randomised controlled trial

Comment (continued)

The basic idea of a randomised controlled trial (RCT) is that you split your participants into two groups (the experimental and the control group).

The experimental group does the task the new way

The control group does the task the old way

You then compare the performance of the experimental and control groups (using statistical comparisons):

- If the experimental group performed significantly better than the control group then you have a result to report
- If the two groups performed similarly (not significantly different outcomes) then you could conclude that the new and the old way of doing the task were of similar quality
- If the experimental group's performance was significantly worse than the control group's then you might conclude that the new way is worse than the old way of doing the task

Randomised controlled trial

Method

Sixteen surgical residents (PGY 1–4) had baseline psychomotor abilities assessed,

then were randomized to either

- [Experimental] VR training (MIST VR simulator diathermy task) until expert criterion levels established by experienced laparoscopists were achieved (n=8),

or

- [Control] non-VR-trained (n=8).

All subjects performed laparoscopic cholecystectomy (gallbladder dissection) with an attending surgeon blinded to training status.

Videotapes of gallbladder dissection were reviewed independently by two investigators blinded to subject identity and training, and scored for eight predefined errors for each procedure minute (interrater reliability of error assessment r 0.80).

Randomised controlled trial

Results

No differences in baseline assessments were found between groups.

- Gallbladder dissection was 29% faster for VR-trained residents.
- Non-VR-trained residents were nine times more likely to transiently fail to make progress (P<.007, Mann-Whitney test) and five times more likely to injure the gallbladder or burn non-target tissue (chi-square 4.27, P<.04).
- Mean errors were six times less likely to occur in the VR-trained group (1.19 vs. 7.38 errors per case; P<.008, Mann-Whitney test).

Randomised controlled trial

Conclusions

The use of VR surgical simulation to reach specific target criteria significantly improved the OR performance of residents during laparoscopic cholecystectomy.

This validation of transfer of training skills from VR to OR sets the stage for more sophisticated uses of VR in assessment, training, error reduction, and certification of surgeons.

The potential exists to train a resident (trainee surgeon) to a high level of objectively measured skill before he or she is permitted to operate on a patient.

Randomised controlled trial

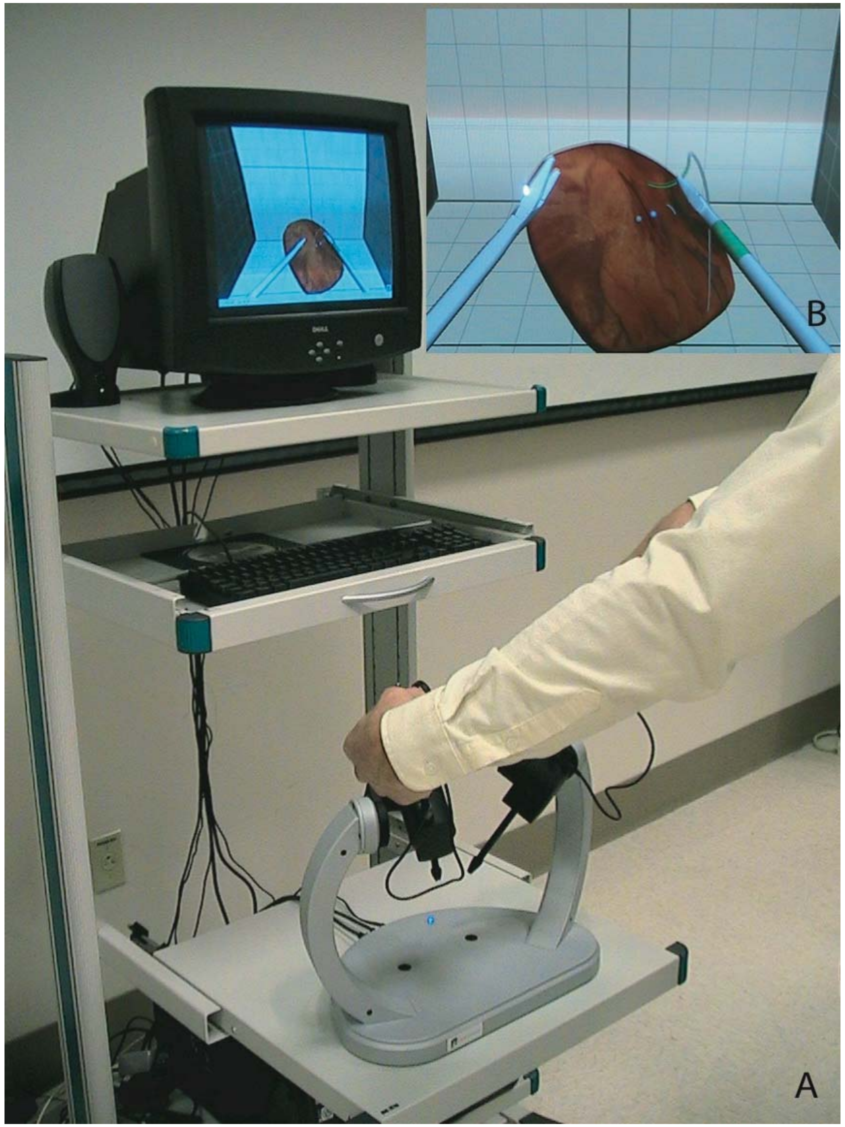
Methods – baseline assessment

Sixteen surgical residents (11 male, 5 female participated in this study. All study participants were randomly assigned to either a study group that would receive VR training in addition to the standard programmatic training, or a control group that would receive standard training only.

All residents in both groups completed a series of previously validated tests to assess fundamental abilities. Visuospatial assessment included the pencil and paper Card Rotation, Cube Comparison, and Map Plan tests.12 Perceptual ability (reconstruction of 3-D from 2-D images) was assessed on a laptop computer with the Pictorial Surface Orientation test (PicSOr).13

Psychomotor ability was assessed with the Minimally Invasive Surgical Trainer-Virtual Reality (MIST VR) system (Mentice AB, Gothenburg, Sweden) with all tasks set at medium level of difficulty.

Randomised controlled trial



MIST-VR trainer:
laparoscopytoday.com

Randomised controlled trial

Apparatus

With this system, a 3-D “box” on the computer screen represents an accurately scaled operating space.

Targets appear within the operating space according to the specific skill task selected and can be grasped and manipulated with virtual instruments.

Each of the different tasks is recorded exactly as performed and can be accurately and reliably assessed.

Comments on the training method:

Train until the resident makes no errors in 2 successive performances of the task.

Randomised controlled trial

Calibration

Four attending surgeons, all with extensive prior experience with laparoscopic procedures, completed 10 trials on the MIST VR “Manipulate and Diathermy” task. at the “Difficult” level to establish the performance criterion levels (mean error score 50, mean economy of diathermy score 2).

The training goal for residents in the VR group was to perform the same task equally well with both hands on two consecutive trials at the criterion levels set by the experienced surgeons.

Training sessions lasted approximately 1 hour. Training was always supervised by one of the authors (A.G.G. or N.E.S.), and explicit attention was paid to error reduction and economy of diathermy.

Randomised controlled trial

Pre-trial checking that the residents knew what to do

Before procedures, all were asked to view a short training video demonstrating optimal performance of excision of the gallbladder from the liver using a hook-type monopolar electrosurgical instrument. This video defined specific deviations from optimal performance that would be considered errors. After the viewing, all residents were given an eight-question multiple-choice examination that tested recognition of these errors.

Randomised controlled trial

Performing the surgical task for the trial

During surgery, after division of the carefully identified cystic structures, residents were asked to perform the gallbladder excision using a standardized two-handed method.

This phase of the procedure was video-recorded with voice audio by the attending surgeon describing any interventions (attending takeover of one or both instruments). Procedures with attending takeover were flagged for examination of audio.

Randomised controlled trial

Rating method for evaluating the residents' performances

… eight events associated with the excisional phase of the procedure were defined as errors and chosen as the study measurements.

These measurements excluded any inferences that were not directly observable. All of the events were explicitly defined to facilitate interrater agreement. Clear guidance was given as to when an event was judged to have or have not occurred.

The length of time of the gallbladder excision phase was also determined. Timing of length of procedure started with first contact of the electrosurgical instrument with tissue and ended when the last attachment of the gallbladder to liver was divided.

Randomised controlled trial

Results (details)

The duration of the dissection for the VR-trained group was 29% less than in the ST group, although this difference did not achieve statistical significance .

[Timing the activities to get a numerical measurement]

Gallbladder injury and burn of nontarget tissue errors were five times more likely to occur in the ST group than in the VR group (one of each error in VR residents as compared to five of each error in the ST residents).

[Counting the errors to get a numerical measurement]

Separate comparisons between the groups for these errors demonstrated statistical significance in both cases (chi square 4.27, df 1, P<.039)

Randomised controlled trial

Results (more details)

ST residents [i.e. the control group] were nine times more likely to be scored as lack of progress [when compared with the experimental group], with mean number of lack of progress errors per case of 0.25 versus 2.19 (VR vs. ST groups, respectively;
Mann-Whitney, Z=2.677, P<.008).

[The authors have simplified their written language here and assume that you know they are reporting a statistically significant result]

Qualitative comment:

There was one liver injury, three dissection incorrect plane, and six attending surgeon takeover errors scored, all in the ST group.

Randomised controlled trial

Results (more details)

The ST group made six times as many errors as the VR group with four times the variability in the performance of the VR residents as indicated by standard errors. The mean number of scored errors per procedure was significantly greater in the ST than in the VR group (1.19 vs. 7.38, Z = 2.76, P<.006, Mann-Whitney test).

The paper then uses graphical methods to show these results and it discusses their clinical importance.

Randomised controlled trial

This paper uses written shortcuts to describe the use of statistics to analyse the data (most papers do this). The longer version would look like:

The Null Hypothesis:

> There is no difference in the quality of the training between the standard-trained group and the VR-trained group.

If we assume that the Null Hypothesis is true and if we have properly randomly allocated our participants to the experimental and control groups then we can calculate the probability of each of the observed results.

1. For the comparison of the types of errors between the two groups, the probability of the observations would be < 0.039

Randomised controlled trial

2. The probability of the mean number of lack of progress errors per case being 0.25 versus 2.19 (VR vs. ST groups) would be less than 0.008
3. The probability of the mean number of scored errors per procedure being 1.19 vs. 7.38 (VR vs. ST groups) would be less than 0.006

Each of these three is a fairly unlikely event (small probabilities). Therefore we decide that these events did not occur by chance, so we reject our Null Hypothesis.

When we write about our results we use the adjective "significant" to show that our Null Hypothesis probabilities are less than the tolerance levels we have chosen.

Summary

Randomised: participants selected at random/randomly allocated to the experimental and control groups

Controlled trial: two groups of participants, one does the task in the normal way (the control group) and the other does the task in the new way (the experimental group)

Blinded: the people assisting with the task and/or assessing the task do not know whether the participants are in the control group or the experimental group.

Null hypothesis: a formal concept that is part of the accepted way of doing a Randomised Controlled Trial

Significant: the probability of the result is less than you pre-set limit (0.05, 0.01, 0.001 etc.)

Summary

Prospective trial: The trial creates the data under controlled conditions

Retrospective trial: The evaluation is performed on data that already exists (data gathered from trials where you did not necessarily have any control over the trial conditions)

Baseline assessments: You can conduct small test experiments with your participants to check that the groups you have set up are not significantly different with respect to attributes that matter in the experiment.

Independent variables: The variables that separate the experimental group from the control group (e.g. VR training versus standard training)

Dependent variables: The variables that are measured during the trial (e.g. number of errors, time taken)

The role of qualitative data

Qualitative data can be used to create quantitative data

- Example: Errors made by the surgical trainees during surgery
- A series of observations of the surgical trainees making errors as they proceed with the dissection task is a qualitative data element that could be analysed in a grounded theory approach to get an understanding of what the trainees did
- If the same observations are selected from a standard list of possible errors, and time-coded, then you get a numerical data element (number of each type of error per unit time).
- If you also standardize the observations across the set of observers you get a whole-of-experiment numerical data set that you can use to make quantitative comparisons.

The role of qualitative data

Qualitative data can be used to explain quantitative data and, in particular, to explain unusual data elements

Example: An unusually slow participant in a timed experiment

You might have a participant who takes an unusually long time to complete a task (when compared to the other participants).

If you have observation data you might be able to see why this happened, for example a technical failure.

If you have an exit interview you might ask the participant why they were so slow. This might uncover a variable that you did not properly control.

The role of qualitative data

Qualitative data can be used to show other team members the nature of a problem

- Example: You might be comparing a new and old system and the quantitative data shows the new system is performing more slowly than the old system. Observational data might help you find what it is that is slowing the system down.

The role of qualitative data

Qualitative data can help you explore the users' experiences of the system

- Example: You might have a new system that is “better” (in some quantitative way) than the older system but the users do not like the experience of using it.
- Qualitative data, and in particular interview data that treats the participants as intelligent partners in the research, might help explain the contradiction.