CS Students and their bias in HCI evaluation tasks

Joost Kroes
University of Twente
Calslaan 54-204
7522 MG Enschede
J.J.Kroes@student.utwente.nl

ABSTRACT
A large percentage of Human Computer Interface (HCI) research is done exclusively on groups of CS students. Students, as studies in social sciences indicated, are sometimes a poor representation of the general population.

One could argue whether students actually are a bad representation of the general population. There is no concrete data on what to expect. To gain insight in these matters, an experiment with a number of evaluation tasks performed by both CS students and a more general population sample is held.

From this experiment, it can be concluded that CS students, when compared to the general population, have stronger opinions towards interface complexity and tend to heavily favor interfaces that they themselves know well, but are not necessarily equally well-used or commonplace among the general population.

Keywords
Sampling Bias, HCI, CS students, Evaluation tasks, technological aptitude, interface evaluation.

1. INTRODUCTION
HCI is the study of Human Computer Interaction. A field of Computer Science that specializes in interaction between computers and humans and encompasses the whole gamut of work related to interface design, implementation and evaluation. [3]

1.1 Evaluation in HCI
Evaluation could be considered to be the lifeblood of HC. It is continually being used to validate interface designs and theories in myriad experiments.

The evaluation of computer interfaces has evolved over time, leading to a wide array of evaluation methods [6], starting with simple, measurable metrics such as task completion time and branching into newer (and harder to measure) concepts such as user experience. [7]

The percentage of papers using some form of evaluation sent to the ACM SIGCHI conferences for instance has risen from just over half to 97% in 2006,[1] Evaluation has become a norm for HCI research.

1.2 Research participants
With evaluation as a compulsory element of HCI research, large amounts of evaluation experiments are set-up and executed, requiring a large amount of participants. To quickly and easily access and find these participants, a large amount of HCI researchers make use of HCI or Computer Science (CS) students as participants in their evaluation tasks.

A 2007 meta-study showed that of all papers published in 2006, 48% used primarily students in their evaluation tasks. [1] Though this percentage has been decreasing (from 57% in 2000), the high percentage of CS and HCI students could indicates a sampling bias that might invalidate the findings of HCI evaluation tasks.

1.3 A resurfacing issue
Though not heavily discussed within the field of HCI yet, this problem is not new. It has been heavily debated and referenced in the social sciences. [4] [8] [10] To quote Peterson: “There seem to be few insights in the college-student-as-research-subject issue that have not been previously presented, replied to, subjected to rejoinder and subsequently forgotten.” [8]

The experience gained in these previous studies shows that student groups can sometimes yield differing results when compared to a more generic population sample. However, there is no reliable method of predicting these differences. [4]

1.4 Meet the CS student
In the case of this research, the focus lies upon the CS student, whom is the primary research subject for a large amount of HCI studies.

The group of CS students has a few traits, when compared to non-students1.

1.4.1 Age
CS students are generally aged 18-26, with the vast majority of CS bachelor studies being finished before the age of 26 and the vast majority of CS master studies being finished before the age of 27. [11]

1.4.2 Sex
CS students, both in the bachelor and the masters’ program are predominantly male. [1] [9]

1.4.3 Education
By definition, the CS bachelor or master student is highly educated or becoming highly educated, with only a doctorate offering a higher level of education.

---

1 The term non-students is used throughout this paper to indicate a sample of the general population. It does not indicate those whom do not study, but indicates the entire population minus the set of CS students.
1.4.4 Technological aptitude

CS students have extensive experience and knowledge regarding interaction with computers. Non-CS Students have varying levels of experience and knowledge regarding computers and computer programs.

1.5 The case against students

The CS student having such a distinct profile combined with the rampant use of CS students as participants in HCI could endanger the validity of HCI research.

However, as indicated in previous studies, students might or might not differ; it depends on the type of experiment and the measured variables. In the case of HCI, there is no clear indication whether CS students differ significantly in comparison with a general population sample, though it seems likely that they do.

2. RESEARCH QUESTION

The research question of this paper can be summarized as follows:

“Do CS students yield significantly different results in HCI evaluation tasks when compared to a general population sample?”

This also implies the following question(s):

“Can the difference between CS students and non-students on evaluation tasks be explained by certain features of the CS students?”

3. METHODOLOGY

To determine if CS students respond significantly differently to HCI evaluation tasks when compared to a more general population sample, a simple interface evaluation experiment was set-up by means of a questionnaire that required participants to evaluate features of a set of images of interfaces.

3.1 Experiment set-up

This set-up is similar to an old consumer studies experiment [2] by Hawkins that was used to prove that students used in consumer research differ significantly from another group as housewives.

Because this experiment attempts to do the same as Hawkins did, his experiment is used as a basis for this experiment.

The experiment is divided into two parts. First, all respondents were asked to rate 4 examples of a specific type of interface on four different features parallel to Hawkins’ study. These features are:

- **Familiarity** with this interface, ranging from unknown to very well known.
- **Beauty** of this interface, ranging from ugly to beautiful.
- **Ease-of-use** of this interface, ranging from simple to hard.
- **Complexity** of this interface, ranging from low to high.

Secondly, all participants are asked to rate one of the interfaces as most liked and one of the interfaces as most disliked. This yields a second set of data for extra comparison between both groups.

3.1.1 Interface choice

In designing the experiment, a large issue was the selection of the type of interface to be evaluated. One could argue that since this study focuses on the difference in reaction of two groups on an interface, and not the actual interface itself, that the selection of an interface is arbitrary.

However, participants with both a high and low amount of computer aptitude will have to be able to evaluate the interfaces presented, which suggest using a type of interface that is already commonplace and well known.

To this end, the type of interface was chosen to be a media player. There are many different designs and implementations available of media players, yet their basic functionality is well known and generally the same over most media players.

See appendix B for the images of media player interfaces that were used.

3.2 Questionnaire design

3.2.1 Participant information

Before starting the experiment part of the questionnaire, participants were asked to supply some personal information. Considering participant privacy, none of the information could be linked directly to any of the participants.

The following information was supplied by every participant:

- Age
- Sex
- Education level
- Current education
- Self-reported computer aptitude

As the self-reported aptitude is somewhat ‘fuzzy’ when compared to the other information that participants are asked to provide, a validation question is added that asks participants the opposite question that the aptitude question poses. The aptitude level that is used in this research is the average between the score on the aptitude question and the inverse score on the lack-of-aptitude question.

3.2.2 Interface ratings

As part of the experiment, the participants were asked to rate 4 different images of media player interfaces on the 4 different features from section 3.1.

The participants rated these interfaces on a seven-point semantic differential scale, wherein a participant is presented with a pair of bipolar adjectives between which the participant must choose along a scale of seven points.

See Appendix A for an example of a 7-point semantic differential scale as used in the experiment.

3.3 Kolmogorov-Smirnov two-sample test

To determine statistically significant differences between the results of CS students and non-student, the Kolmogorov-Smirnov two-sample test is used.

The Kolmogorov-Smirnov two-sample test is a statistical test that quantifies a distance between the empirical distribution function of one sample and the empirical distribution function of another sample. The two-sample Kolmogorov-Smirnov test is used to determine the difference between two datasets because it is sensitive to both differences in location and shape of the empirical cumulative distribution functions of both samples.

In this experiment, the Kolmogorov-Smirnov two-sample test is modified to serve as a test of the goodness-of-fit of the data gathered from both groups of participants, meaning that the test
either indicates that 2 datasets fit (and are not significantly different) or that these sets do not fit (and thus are significantly different).

Applying the Kolmogorov-Smirnov test to two data sets yields a distance D that can be used to determine if the data sets fit. The distance D must exceed a certain critical value to indicate that two sets of data are indeed significantly different. This critical value is determined as follows:

\[
\text{Critical value} = 1.36 \times \sqrt{\frac{n_1 + n_2}{n_1 + n_2}}
\]

Because we chose to use a significance level of 0.05, the expression is multiplied by 1.36. This value can be found in statistical tables for the two-sample Kolmogorov test. \(n_1\) is the number of elements in sample set 1 and \(n_2\) is the number of elements in sample set 2.

With \(n_1 = 29\) and \(n_2 = 28\), our critical value is 0.360. Any pair sample sets with a greater distance \(D\) than 0.360 can be considered to be statistically significantly different under a significance level of 0.05 (or 95% probability).

4. RESULTS

4.1 Participants

A sample of 29 CS students and 28 non-students was polled in the experiment. Though there were more non-student respondents in total, some of them failed to answer several of the evaluation questions. Their results were omitted in this experiment.

4.1.1 Students

The CS student participants were recruited in the break room of the CS faculty, making it relatively simple to reach a broad audience of CS students.

The CS student group consisted of 27 male and 2 female students, with an average age of 21.3 (\(\sigma = 2.8\)) and self-reported their aptitude with computers and programs as 1.9 (\(\sigma = 0.8\)) on a scale from -3 (lowest) to 3 (highest).

4.1.2 Non-students

To reach a large amount of non-students, several locations were visited in an attempt to prevent a sampling bias. Participants were recruited at the local public library, at a municipal swimming pool, at local cafeterias and at a church.

The non-student group proved to be harder to reach, as some potential candidates refused to participate when learning that the content of the experiment was related to computers and programs. These candidates stated that they had little ability or knowledge of computers and chose not to participate for that reason. This might skew the results of the self-reported aptitude.

The non-student group consisted of 21 male and 7 female participants, with an average age of 37.3 (\(\sigma = 10.6\)). The non-student group self-reported their aptitude with computers and programs as 0.9 (\(\sigma = 1.3\)) on a scale of -3 (lowest) to 3 (highest).

4.2 Evaluation tasks

4.2.1 Interface preference

Evaluating the results of the evaluation task defined in 3.1.1 and grouping this by non-students and CS students yielded table 1 and table 2, containing the percentages of participants that selected the given interface as most liked and most disliked per interface.

### Table 1. Non-student preference task results.

<table>
<thead>
<tr>
<th>NONSTUDENTS</th>
<th>most liked</th>
<th>most disliked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface 1</td>
<td>25%</td>
<td>39%</td>
</tr>
<tr>
<td>Interface 2</td>
<td>21%</td>
<td>17%</td>
</tr>
<tr>
<td>Interface 3</td>
<td>21%</td>
<td>27%</td>
</tr>
<tr>
<td>Interface 4</td>
<td>33%</td>
<td>17%</td>
</tr>
</tbody>
</table>

Application of the Kolmogorov-Smirnov two-sample test on both the set of preferred and the set of non-preferred interfaces between both groups indicated the set of liked interfaces is significantly different between students and non-students (\(D = 0.396\)) but that the set of disliked interfaces is not (\(D = 0.209\)).

Also, it seems that students have a very high personal preference towards media player interfaces, whereas the nonstudents seemed to have a lower preference for a specific media player. They picked all 4 players almost equally.

Some students expounded upon their reason to pick a specific media player. In the case of 16 students, the main reason to pick a specific interface as most liked was that it was “already known” and “known best”. It seems that the increased experience and knowledge students skewed their preferences towards those interfaces they already knew.

4.2.2 Interface evaluation task

Using the 7-point semantic differentials, answers on the 7 point scale were transformed into numerical values, ranging from -3 (leftmost answer) through 0 (the central answer) up to 3 (the rightmost answer).

For all features over all interfaces two sets of these numerical values were made. One set for CS students and one set for non-students.

The mean value and standard deviation of the evaluation answers are displayed in table 3. Application of the Kolmogorov-Smirnov two-sample test on the data sets of all four interfaces and over the four different features yielded surprisingly little result. The points where students and non-students disagreed significantly were:

1. The complexity of interface 1 (\(D = 0.417\)). CS students generally stated that this interface was significantly more complex than non-students stated.
2. The complexity of interface 2 (\(D = 0.423\)). CS students generally stated that this interface was significantly more complex than non-students stated.
3. The familiarity with interface 4 (\(D = 0.455\)). Nearly all CS students claimed at least moderate to high familiarity with this interface.

Almost significantly different was the feature ease-of-use feature of interface 1. It came close, but fell short of the critical value for the Kolmogorov-Smirnov two-sample test (\(D = 0.358\)).
5. CONCLUSION

Though the results in Table 3 were slightly disheartening, with this experiment being modeled after Hawkins’ experiment, where 42% comparisons turned out statistically significant (with 95% probability), as opposed to merely 19% in this experiment, the research questions can be answered somewhat satisfactorily:

CS students do respond differently on some interface evaluation tasks compared to non-students, though the only notable differences are the following:

5.1.1 Complexity
CS students as a group seem to have stronger opinions on the complexity of interfaces. CS students seem to consider interfaces that non-students deem complex even more complex and seem to consider interfaces that non-students deem simple as even more simple.

This seems to fit Peterson’s finding that student groups in experiments tend to have more pronounced statistical effects on the range of variables tested. [8]

5.1.2 Familiarity
Almost all CS students whom participated in the experiment were somewhat to very familiar to interface 4. In comparison, non-students considered this interface to be slightly unknown. That participants favor those interfaces that they already know is not a big surprise. That the group of CS students as a whole is very familiar with a specific type of interface, however, is significant.

CS students seem have a significantly larger familiarity and knowledge of interfaces and subsequently much stronger opinions towards these opinions when compared to non-students, whom as a group are not as familiar with certain interfaces and therefore have weaker opinions towards interfaces.

6. RECOMMENDATIONS

Though it comes as no big surprise that CS students do tend to sometimes differ from non-students, the student-as-research-subject problem persists and not just within HCI.

The periodic resurfacing of this issue and its subsequent fading into obscurity within the social sciences seems to indicate a basic problem when dealing with students and wildly differing studies: There is no clear way on how to predict, account for or prevent student participant bias.

There is a need for methods that allow us to translate results of experiments with CS students to the realm of the non-student and vice-versa.

To that end, an interesting avenue for future work is to evaluate the correlation between self-reported aptitude influences the way the familiarity as described in section 5.1.2 influences CS students and non-students in evaluation tasks.

Also, all comparisons are done on basis of the evaluation task results. Experiments with larger sample sizes could use participant information such as education level, self-reported aptitude, sex and age.

7. ACKNOWLEDGEMENTS

I would like to thank the employees at the public library of Enschede, the employees at the reformed church in Hattem and the employees at the municipal pool Watermulderin Wezep for graciously allowing me to recruit participants in on their premises.
8. REFERENCES


APPENDIX

A. Semantic differential scale example
An example a question and the 7-point semantic differential used in the experiment:

1.6 How high do you consider your aptitude in using computers and computer programs?

| Low | O | O | Average | O | O | High | O |

B. Media player Interfaces
The four images of interfaces used in the experiment.

Fig 1. Media player interface 1
Fig 1. Media player interface 2
Fig 1. Media player interface 3
Fig 1. Media player interface 4