
User Modeling and User-Adapted Interaction **11:** 181–194, 2001.

Empirical Evaluation of User Models and User-Adapted Systems

DAVID N. CHIN
Department of Information and Computer Sciences, University of Hawaii, 1680 East West Rd., Honolulu, HI 96822, U.S.A. E-mail: chin@hawaii.edu

(Received: 9 March 2000; in final form 14 July 2000)

Abstract. Empirical evaluations are needed to determine which users are helped or hindered by user-adapted interaction in user modeling systems. A review of past *UMUAI* articles reveals insufficient empirical evaluations, but an encouraging upward trend. Rules of thumb for experimental design, useful tests for covariates, and common threats to experimental validity are presented. Reporting standards including effect size and power are proposed.

Key words: empirical evaluation, experimental design, covariant variables, effect size, treatment magnitude, power, sensitivity.

1. What Is Empirical Evaluation?

Empirical evaluation refers to the appraisal of a theory by observation in experiments. The key to good empirical evaluation is the proper design and execution of the experiments so that the particular factors to be tested can be easily separated from other confounding factors. For example, one may want to test whether a software system with a user model works better than the same system without a user model, test the effect of different levels of user modeling or different user model parameter settings, or test different user interfaces. These factors, which are under the control of the experimenter, are termed *independent variables* because their values can be varied independently of other variables by the experimenter. *Dependent variables* are variables whose values depend on the values of other variables. They include response variables or recorded measures such as the frequency/extent of certain behaviours (e.g., system usage), qualities of a behavior in a particular situation, number of errors, error rate, time to complete a task, proportion/quality of tasks achieved, interaction patterns, learning time/rate, and/or subjective evaluations (e.g., user satisfaction). Some dependent variables can only be measured indirectly such as cognitive load measured through blood pressure or pupil dilation.

In an ideal experiment, only the independent variables are varied and everything else is fixed so that any changes in the dependent variables can be directly attributed to the variations in the independent variables. Unfortunately, such ideal control is

almost impossible. If different participants[1] are used for the different independent variable cases, then individual differences (intelligence, reading ability, spatial reasoning, perceptual abilities such as color blindness, poor eyesight, poor hearing, knowledge, etc.) will typically influence the dependent variables much more than the independent variables. If the same participant is involved in all independent variable conditions, then there is the problem that the earlier conditions will affect the later conditions due to practice effects (e.g., a participant may be able to perform a task faster the second time around).

There are also potential problems with different times, locations, or other environmental conditions influencing the dependent variables. People are sometimes more and sometimes less tired and people have different moods at different times of the day and days of the week/month, which affects their performance and/or their subjective evaluations. There may be more distracting noise at certain times (e.g., jackhammers at the construction site across the street run only in the afternoon). Computers may be slower at certain times. For example, suppose the experimenter schedules participants to test the user-adapted program on the hour and the program with no user model on the half-hour. Network access may be much slower on the hour because students in teaching labs load their programs from network servers at the start of class. Also others (e.g., the experimenter) may bias the participants by words, tone, body language, or even appearance. For example suppose a friendly, beautiful assistant runs the no user model cases and a rude, dirty assistant with bad body-odor runs the user model cases. Any of these problems can obscure the effects of user models on users, because such UM effects are typically not very large.

To overcome such problems, participants are randomly assigned to groups in order to average out the effect of 'nuisance variables' on the dependent variables. Of course, in order for averaging to work properly, large numbers of participants are needed. Statistical techniques for analysis of variance (ANOVA) are used to determine whether differences in dependent variable values among groups are due to the different independent variable treatments or due to random fluctuations. To improve the sensitivity of experiments (and thus reduce the number of participants needed), *crossed designs* use the same participants for multiple dependent variable conditions. For example, the same participant uses both the user-adapted system and the no-UM system. Crossed designs control for practice effects by varying the order of dependent variable conditions for different participants and participants are randomly assigned to the different orders.

Experiments are also performed blind or double-blind. In the blind experiment, participants do not know if the software has a user model and so is 'supposed to be better.' Just as placebos lead to large improvements in drug trials, so too the mere belief that a user model or other advanced technology is present in a program will likely bias participants. Even the appearance of computers (more or less

[1] Participants were formerly called subjects, but this terminology is no longer deemed politically correct.

modern) or neater desks may bias the beliefs of participants. In the double-blind experiment, the experimenter is also not aware, and so cannot inadvertently influence participants (this is standard practice for drug trials). In tests of audio equipment,[2] experimenter body language biased participants, *even when experimenters were trying* ***not*** *to* (Shanefield, 1980).

This article cannot possibly cover all of the facets of proper experiment design and data analysis. Luckily, there are many books and articles on this subject. I can recommend (Keppel, 1991; Stevens, 1992; Neter et al., 1985; Huck et al., 1974; Campbell and Stanley, 1963). There are also many on-line sources for statistics such as www.statistics.com, www.stat.ufl.edu/vlib/statistics.html, www.psychstat.smsu.edu/introbook/sbk00.htm, and stat.tamu.edu/stat30x/notes/trydouble2.html. There is even a Java-based web-browser analysis program at members.aol.com/johnp71/javastat.html and web-based psychological experimentation labs where researchers can post web-based experiments (see www.psych.unizh.ch/genpsy/Ulf/Lab/WebExpPsyLab.html for one such lab and for links to other lab sites).

2. Why User Models Need Empirical Evaluation

User models cannot and should not be separated from the software systems that use them.[3] After all, what good is a user model if it will not be used for anything? A system with an unused user model might as well not have a user model at all. If the user model is indeed being used to make a difference in a software system such as adapting the software system to the user, then one should ask whether the user model adaptations actually improve the software system. Also, what types of users benefit from the adaptations? It may very well be that some user model adaptations are less beneficial or even detrimental to some classes of users.

In general, adding a user model to any software system will most likely make it more complex, less predictable, and more buggy. Consequently, it is a very reasonable question to ask whether or not the user model will actually improve the system. Even when a user model adapts a system to follow the user s explicit wishes exactly, there is still a question as to whether this is a good idea. For example, the user s preferred configuration may actually be slower or more error-prone than an ideal configuration. Or, the plethora of different configurations may make it difficult for users in a group to cooperate, thus decreasing overall efficiency, even though efficiency for individual users working alone may be improved. A common adaptation for user models is information filtering, which may seem to be always helpful, especially in today s information-overloaded society. However, eliminating seemingly irrelevant information can confuse users, thus decreasing performance. For

[2] Audio equipment tests are extremely sensitive: 0.2-dB louder equipment, which is imperceptibly louder, is often rated better (Lipschitz and Van der Kooy, 1981; Illényi and Korpássy, 1981).

[3] This is not to say that user models cannot be tested for accuracy separately from their use in systems. For example, Corbett and Anderson (1993) and Brusilovsky and Eklund (1998) test the accuracy of their user models.

example, eliminating irrelevant streets from maps may make them more readable unless the system has eliminated some of the user's landmarks (e.g., the user may look for a particular unique crosshatch pattern of streets to quickly locate a particular neighborhood). Likewise, eliminating irrelevant links from a web page may confuse the user's navigation when one of the irrelevant links is part of a path that the user relies on. There may be a more direct path elsewhere in the page, but that does not help because the user does not know it. So, we must test the usefulness of user model adaptations through experiments before we can claim that they are helpful.

3. UM Evaluations in the Past

A quick scan of the first nine years of *UMUAI* reveals that only about one third of the articles (excluding surveys, reviews, and special issue introductions) includes any type of evaluation. This is much too low of a percentage. Even worse, some of the evaluations report only preliminary results. Several do not have control groups, some are just pilot studies (and acknowledged as such by their authors) that do not include enough participants to produce statistically significant results, and some are just informal studies without strict controls. Eliminating such preliminary studies leaves only about a quarter of *UMUAI* articles reporting significant empirical evaluations.

There is a strong correlation between topic areas and empirical evaluations. All of the empirical evaluation papers in the first nine years of *UMUAI* can be classified into four broad topic areas: ten papers in adaptive-hypermedia/information-filtering (Jennings and Higuchi, 1993; Kaplan et al., 1993; Boyle and Encarnacion, 1994; Vassileva, 1996; Mathé and Chen, 1996; Raskutti et al., 1997; Newell, 1997; Hirashima et al., 1997; Alspector et al., 1997; Balabanović, 1998), nine in student modeling (Nwana, 1991; London, 1992; Carbonaro et al., 1995; Corbett and Anderson, 1995; Kashihara et al., 1995; Shute, 1995; Webb and Kuzmycz, 1996; Mitrovic et al., 1996; Milne et al., 1996; Sison et al., 1998), nine in plan recognition/mixed-initiative interaction (Calistri-Yeh, 1991; Albrecht et al., 1998; Gmytrasiewicz et al., 1998; Chiu and Webb, 1998; Chu-Carroll and Brown, 1998; Guinn, 1998; Ishizaki et al., 1999; Green and Carberry, 1999; Virvou and du Bulay, 1999), and three in user interfaces/help systems (Tattersall, 1992; Krause et al., 1993; Debevc et al., 1996).

Student modeling has had a long history of empirical evaluation stemming from the educational psychology roots of intelligent tutoring and computer-aided instruction systems. Student modeling systems are typically evaluated by comparing systems with and without student models. As a preliminary step, the accuracy of student models can also be tested. For example, one can compare predicted student actions/results with actual actions/results (e.g., Corbett and Anderson, 1993; Shute, 1995) or compute the percentage of recognized bugs (e.g., Webb and Kuzmycz, 1996; Sison et al., 1998). Systems that use machine learning methods to acquire user models in any area can

evaluate the acquired user models using standard machine learning measures that compare the user model against a reserved set of test data that was not used for training (typically an 80/20% split for training/testing). For a review of machine learning and predictive statistical models see Webb et al. (2001), and Zukerman and Albrecht (2001) respectively. Adaptive hypermedia and information filtering build on the practice of empirical evaluation of information retrieval systems through measures developed in library sciences such as recall and precision (e.g., Mathé and Chen, 1996; Raskutti et al., 1997) and similarity/ relevance metrics (e.g., Newell, 1997; Hirashima et al., 1997; Alspector et al., 1997; Balabanović, 1998). For a review of adaptive hypermedia, see Brusilovsky (1996, 1998, 2001). Hopefully, we in the field of user modeling will also build up our own conventions of empirical evaluation.

Plan recognition/mixed-initiative interaction and user interfaces/help systems have indirect ties to computer-human interaction and the tradition of empirical evaluation that has slowly built up in that field. The user models of plan recognition systems are typically evaluated by the percentage of actual plans recognized in a test corpus of plans (e.g., Calistri-Yeh, 1991), by the frequency and accuracy of predicted next actions (e.g., Albrecht et al., 1998; Chiu and Webb, 1998) or by comparison with an expert human plan recognizer (e.g., Virvou and du Boulay, 1999). Mixed-initiative interaction systems are typically evaluated by comparing system responses choices with human choices (e.g., Green and Carberry, 1999) or by comparing the efficiency of the dialog (usually measured in number of dialog turns) needed to achieve an information transfer task with either human-human dialogs or with theoretically minimum dialogs, in which case simulations can be used to generate the tasks and dialogs (e.g., Guinn, 1998; Ishizaki et al., 1999). The user model in mixed-initiative interaction can be tested separately by comparing the initiative predictions of the model in actual dialogs (e.g., Chu-Carroll and Brown, 1998). User interfaces/help systems are typically evaluated by subjective user satisfaction, task completion speed (e.g., Krause et al., 1993; Debevc et al., 1996), and/or error rate/quality of task achievement (e.g., Krause et al., 1993). For a review of plan recognition see Carberry (2001), for a review of mixed-initiative systems see Zukerman and Litman (2001) and for a review of user modeling in human-computer interaction see Fischer (2001).

Although the absolute percentage (1/4 to 1/3) of research articles that contain empirical evaluations is much too low (ideally all research articles on user modeling should include empirical evaluations), there is some good news. The most recent four years of *UMUAI* articles contain almost twice as many articles with empirical evaluations as the first four years. We, both as authors and reviewers, should strive to improve upon this admirable trend.

4. Designing Your Own Empirical Evaluation

In order to avoid the uneven influence of nuisance variables on the experiment, here are some rules of thumb:

- Randomly assign enough participants to groups.
- Randomly assign time slots to participants.
- Test room should not have windows or other distractions (e.g. posters) and should be quiet. Participant should be isolated as much as possible.
- The computer area should be prepared ergonomically in anticipation of different sized participants.
- If a network is used, avoid high load times.
- Prepare uniform instructions to participants, preferably in a written or taped (audio or video) form. Check the instructions for clarity with sample participants in a pilot study. Computer playback of instructions is also helpful.
- Experimenters should not know whether or not the experimental condition has a user model. Each experimenter should run equal numbers of each treatment condition (independent variable values) to avoid inadvertent bias from different experimenters. Experimenters should plan to minimize interactions with participants. However, the experimenters should be familiar with the user modeling system and be able to answer questions.
- Be prepared to discard participant data if the participant requires interaction with the experimenter during the experiment.
- Follow all local rules and laws about human experimentation. For example, in the USA all institutions receiving federal funds must have a local committee on human subjects (CHS) that approves experiments. Typically, participants should at least sign a consent form.
- Allow enough time. Experiments typically take months to run.
- Do run a pilot study before the main study.
- Brainstorm about possible nuisance variables.

Nuisance variables can also be handled explicitly, in which case they are called *covariant variables, covariates,* or *concomitant variables.* Covariates are first measured (before the experiment) and their influence on dependent variable values is later factored out by the statistical analysis of covariance (ANCOVA). Which covariates should be used depends on the particular experiment. Common covariates include age, gender, socioeconomic status, ethnic background, education, learning styles, previous experience, prior knowledge, and aptitudes. Here are some common measurement tests for covariates that may be useful for certain types of user model evaluations:

Aptitude tests:

- Ball Aptitude Battery®, www.ballfoundation.org/ci/bab.html.

Cognitive tests:

- Ekstrom and French, Kit of Factor-Referenced Cognitive Tests (incl. visualization, visual memory, memory span, perceptual speed) from Educational Testing Service, www.ets.org/research/ekstrom.html.

- Taggart and Torrance, Human Information Processing Survey and other tests from Scholastic Testing Service, walden.mvp.net/~stlsts/gift.shtml.
- Oltman, Raskin and Witkin, Group Embedded Figures Test from Consulting Psychologists Press, www.acer.edu.au/scripts/product.php3?family_code=SH.
- Nelson-Denny Reading Test, www.riverpub.com/products/group/reading.htm.

Personality Tests:

- Meyers-Briggs Type Indicator (MBTI) and others, from CAPT, capt.org or www.cpp-db.com/products/mbti/index.asp (one must be trained to give and interpret the MBTI test).
- Rotter, Locus of Control (attribution theory), www.psychtests.com/lc.html or www.queendom.com/lc.html, (additional personality tests at www.queendom.com/tests.html#personality).
- Kolb, Learning Style Inventory, pss.uvm.edu/pss162/learning_styles.html.

5. Problems in the Empirical Evaluation of User Modeling Systems

There are some common problems in the empirical evaluation of user modeling systems that bear repeating so that researchers can be wary of such problems in their own evaluations. Experiment design problems include the failure to use a control group when one is needed (e.g., a control group for the system without user modeling). Sometimes the experimental procedure itself generates a variable (e.g., thinking aloud modifies the user s problem solving strategy). Data can be contaminated (e.g., incorrect recording or transcription of data). There may be unwarranted assumptions about scales for variables (e.g., eye blink rates are not linearly related). Nuisance variables can be confounded with relevant variables (e.g., the user modeling system turns on interface recording, which slows down the system considerably). Previous training of participants should be taken into account (e.g., participants who have used a similar system previously do better in the experiment). Frequently, there is a failure to include a sufficient number of observations to provide the needed precision. There is also the human tendency to favor one outcome rather than another, which will inadvertently bias the results. This bias can come from either experimenters or participants. Experimenters often fail to recognize the rarity of an event (e.g., winning at gambling leads to expectations of winning greater than the actual odds). The experimental procedure itself can affect conditions to be observed (e.g., knowing that video cameras are present changes behavior).

The internal validity of an experiment refers to whether the independent variables made a difference in the study and, if so, can the researcher infer a cause and effect relationship? Essential to internal validity are control of extraneous variables and selection and measurement procedures. Without internal validity, results are hard to interpret. Internal validity can also be threatened by factors such as *history,*

maturation, testing, instrumentation, statistical regression, mortality, and *selection.* History refers to some other event that affects the dependent variable. The longer the time between pretest and posttest, the greater the chance of history. Maturation refers to biological or psychological processes that occur with the passage of time and are independent of external events. For example, users become more expert over time. Testing refers to the tendency to score higher in subsequent tests when the series of tests are similar to one another. Instrumentation problems occur when there is any change in the observational techniques (machines or judges). Statistical regression refers to the tendency for the mean of extreme scores to drift back to the middle. For example, the mean expertise of very expert users will tend to drift back slightly toward intermediate simply because some of the experts will fail to keep up with their expertise. Mortality refers to the loss of subjects between a pretest and a posttest. If the participants who drop out differ from those who remain, mean scores between the tests could differ because some participants were not measured in the posttest. Selection is when participants who seek exposure to the treatment are compared with participants who do not seek this exposure, since the two groups are likely to differ in motivational levels.

External validity refers to how well the results of the study can be generalized. How representative are the results to other populations, variables, and situations? Threats to external validity include *population* and *ecology*. Population affects external validity when the experimentally accessible population differs from the target population (e.g., computer science students are used for testing, whereas the actual users have much different abilities). There may also be interactions between treatment effects and participant characteristics. Ecological threats include incorrectly describing independent variable(s), incorrectly describing or measuring dependent variable(s), multiple-treatment interference, interaction of history and treatment effects, interaction of time of measurement and treatment, pretest and posttest sensitization, the Hawthorne effect where expectation leads to improvement, the novelty and disruption effect (e.g., any change in business practices leads to improvement, at least in the short term), and experimenter influence on participants (Rosenthal effect, Pygmalion and Golem effects).

The final problem is in the interpretation of the results. Even if significant results are obtained, the meaning is often unclear. For example, if users significantly prefer a user-adapted system, does that make the user-adaptations worthwhile? If the purpose is to sell more systems, then it is almost certainly worthwhile (within a cost-benefit analysis). However, if the purpose is to improve the productivity or achievement quality of users, then other measures are needed. Also, significant results may only apply to some subset of the participants. For example, Specht and Kobsa (1999) found that only the subgroup of learners with low previous knowledge benefited from certain adaptive hypertext strategies. It is often difficult to tell which subgroupings are important and unless participants are pretested (or posttested before changes in their profile have occurred), it is impossible to separate out the subgroups. This makes it especially important to either measure covariant

variables (see the previous section) as part of the experiment, perform pilot studies to rule out the usefulness of suspected covariants (which unfortunately often takes as many participants as full studies) or rely on previously published results that have shown certain covariants to be irrelevant to your particular type of experiment.

6. Proposed Reporting Standards

For future empirical evaluations of user models in this journal and in other forums, I propose that authors should report certain common measures. These include:

- the number, source, and relevant background of the participants,
- the independent, dependent, and covariant variables,
- the analysis method,
- the post-hoc probabilities,
- the raw data (in a table or appendix) if not too voluminous[4],
- the effect size (treatment magnitude), and the power (inverse sensitivity), which should be at least 0.8.

The last two measures are often left out of reports even in top psychological journals. However, they are especially important for new areas of empirical evaluation such as user modeling where there are few previous reports of effect size. The effect size is also called treatment magnitude. It gives the magnitude of the change in dependent variable values due to changes in the independent variables as a percentage of the total variability. There are several different measures of effect size, but the most common is omega squared (ω^2).

$$\omega^2 = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_{S/A}^2},$$

where σ_A^2 is the variance attributable to the effects of varying the independent variable and $\sigma_{S/A}^2$ is the random variance among participants. So ω^2 is the proportion of the total variance ($\sigma_A^2 + \sigma_{S/A}^2$) that is attributable to the effects of user modeling. If the effect size is smaller, then larger numbers of participants will be needed to accumulate the signal from the independent variable manipulations before it becomes visible over the uncorrelated noise from nuisance variables. For researchers about to test user modeling systems similar to something previously evaluated, the previously reported effect size is invaluable for experiment planning. For comparison, in the social sciences ω^2 of 0.01 is considered small, 0.06 medium, and 0.15 large.

Power (also called sensitivity) tells how easily an experiment can detect differences (the technical definition is the probability of rejecting a false null hypothesis). It is measured as the fraction of experiments that for the same design, same number of participants (called sample size) and same effect size would produce the given

[4] If voluminous, the raw data can instead be put on a website referenced by your paper.

significance. So, a power of 0.5 means half of repeated experiments would find non-significant results. A power of 0.8 means 80% of repeated experiments would find the same (significant) results. The general consensus in experimental psychology is that researchers should strive for a power of 0.8 even though the average power of the *Journal of Abnormal and Social Psychology* papers is only 0.5. I recommend that we UM researchers should strive for power ratings of 0.8 in our experiments, especially when reporting results in this journal and other forums.

Designing your experiments to have a high power rating not only ensures greater repeatability of results, but more importantly, it makes it *more likely that you will find your effect.* Consider the case of a power rating of 0.5. This means your experiment will have only a 50% chance of getting a significant result and a corresponding 50% chance of failure. Rather than repeating your experiment if you don't get a significant result, you should design your experiment so that you have a higher chance of getting significant results in the first place. Since it is unlikely that others will duplicate your user modeling experiment, this somewhat self-serving purpose is a much more compelling reason to design your experiments with a power rating of 0.8. Sometimes in designing your experiment, increasing the number of participants in order to increase power is not a viable approach. An alternative is to lower the significance threshold. Although a significance threshold of 0.05 is traditional, it is just an arbitrary number and a higher significance level of say 0.1 may be better in some cases.

Power is best calculated by using programs. Gatti and Harwell (1988) describe the many advantages of using programs rather than power charts. A pilot study (small preliminary experiment) can be used to estimate factors that affect power such as effect size.

The power measure is especially important to report when describing an experiment with non-significant results. If the power is low, then it may just mean that there were not enough participants in the study rather than there was no difference. I would like to recommend that reviewers place equal value on reports of non-significant results as significant results if the reports contain effect size and power measurements. My reasoning is that such non-significant results give important upper bound estimates on the effect size (how much the user model affects user performance or satisfaction). Such information is quite valuable for practitioners who are deciding whether to add user modeling to their software systems. If they know that adding this type of user modeling to this type of system will most likely improve performance or satisfaction less than the reported amount, then they can make a much more informed decision.

7. Alternative Techniques

Besides numeric methods, qualitative methods can also be quite useful for empirical evaluation of user models. One advantage of qualitative methods is that they typically require fewer participants. Some qualitative methods that might be appli-

cable to evaluation of user models include ethnographic field studies, content analysis, case studies, self reports, and interviews. For more information about qualitative evaluation, see Yin (1994), Miles and Huberman (1994), Silverman (1993), Marshall and Rossman (1999), Weber (1990), the *Qualitative Research in Information Systems* journal and web links at www.auckland.ac.nz/msis/isworld/.

8. Conclusions

Although the user modeling community is starting to embrace empirical evaluation, there is still a long way to go. As researchers and as reviewers, our community can do more to educate both one another and ourselves about empirical evaluation techniques and to encourage more empirical evaluation. We can also adhere to better reporting standards for empirical evaluations and avoid bias against reports of properly designed and executed experiments with statistically insignificant results. Hopefully the next ten years of UMUAI will see empirical evaluations become common practice in user modeling.

Acknowledgements

This work was supported in part by the Office of Naval Research under contract N00014-97-1-0578. I would like to thank Martha Crosby, who helped develop much of this material as part of our joint tutorial on 'Evaluating the Effectiveness of User Models by Experiments' presented at the Seventh International Conference on User Modeling, Banff, Canada, June, 1999 (available at www.ics.hawaii.edu/~chin/UM-99-tutorial.html). I would also like to thank David Albrecht, Peter Brusilovsky, Judy Kay, Alfred Kobsa, and Diane Litman for their very helpful comments.

References

Albrecht, D. W., Zukerman, I. and Nicholson, A. E.: 1998, Bayesian models for keyhole plan recognition in an adventure game. *User Modeling and User-Adapted Interaction* **8**(1–2), 5–47.

Alspector, J., Kołcz, A. and Karunanithi, N.: 1997, Feature-based and clique-based user models for movie selection: a comparative study. *User Modeling and User-Adapted Interaction* **7**(4), 279–304.

Balabanović, M.: 1998, Exploring versus exploiting when learning user models for text recommendation. *User Modeling and User-Adapted Interaction* **8**(1–2), 71–102.

Boyle, C. and Encarnacion, A. O.: 1994, MetaDoc: An adaptive hypertext reading system. *User Modeling and User-Adapted Interaction* **4**(1), 1–19.

Brusilovsky, P.: 1996, Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction* **6**(2–3), 87–129.

Brusilovsky, P.: 1998, Methods and techniques of adaptive hypermedia. In: P. Brusilovsky, A. Kobsa and J. Vassileva (eds.), *Adaptive Hypertext and Hypermedia.* Kluwer Academic Publishers, Dordrecht, pp. 1–43.

Brusilovsky, P.: 2001, Adaptive hypermedia. *User Modeling and User-Adapted Interaction* **11**(1–2), 87-110 (this issue).

Brusilovsky, P. and Eklund, J.: 1998, A study of user-model based link annotation in educational hypermedia. *Journal of Universal Computer Science* **4**(4), 429–448.

Calistri-Yeh, R. J.: 1991, Utilizing user models to handle ambiguity and misconceptions in robust plan recognition. *User Modeling and User-Adapted Interaction* **1**(4), 289–322.

Campbell, D. T. and Stanley, J. C.: 1963, Experimental and quasi-experimental designs for research. In: N. L. Gage (ed.), *Handbook of Research on Teaching*. Rand McNally and Co., Chicago.

Carberry, S.: 2001, Techniques for Plan Recognition. *User Modeling and User-Adapted Interaction* **11**(1–2), 31–48 (this issue).

Carbonaro, A., Maniezzo, V., Roccetti, M. and Salomoni, P.: 1995, Modelling the student in Pitagora 2.0. *User Modeling and User-Adapted Interaction* **4**(4), 233–251.

Chiu, B. C. and Webb, G. I.: 1998, Using decision trees for agent modeling: Improving prediction performance. *User Modeling and User-Adapted Interaction* **8**(1–2), 131–152.

Chu-Carroll, J. and Brown, M. K.: 1998, An evidential model for tracking initiative in collaborative dialogue interactions. *User Modeling and User-Adapted Interaction* **8**(3–4), 215–253.

Corbett, A.T., Anderson, J. R. and O'Brien, A. T.: 1993, The predictive validity of student modeling in the ACT programming tutor. In: P. Brna, S. Ohlsson and H. Pain (eds.), Artificial Intelligence and Education, 1993: The Proceedings of AI-ED 93. Charlottesville, VA: AACE.

Corbett, A. T. and Anderson, J. R.: 1995, Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction* **4**(4), 253–278.

Debevc, M., Meyer, B., Donlagic, D. and Svecko, R.: 1996, Design and evaluation of an adaptive icon toolbar. *User Modeling and User-Adapted Interaction* **6**(1), 1–21.

Fischer, G.: 2001, User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction* **11**(1–2), 65–86 (this issue).

Gatti, G .G. and Harwell, M.: 1998, Advantages of computer programs over power charts for the estimation of power. *Journal of Statistics Education* **6**(3). Also available at www.amstat.org/publications/jse/v6n3/gatti.html.

Gmytrasiewicz, P. J. and Noh, S. and Kellogg, T.: 1998, Bayesian update of recursive agent models. *User Modeling and User-Adapted Interaction* **8**(1–2), 49–69.

Green, N. and Carberry, S.: 1999, A computational mechanism for initiative in answer generation. *User Modeling and User-Adapted Interaction* **9**(1–2), 93–132.

Guinn, C. I.: 1998, An analysis of initiative selection in collaborative task-oriented discourse. *User Modeling and User-Adapted Interaction* **8**(3–4), 255–314.

Hirashima, T., Kashihara, A. and Toyoda, J.: 1997, Information filtering using user's context on browsing in hypertext. *User Modeling and User-Adapted Interaction* **7**(4), 239–256.

Huck, S., Cormier, W. and Bounds, W.: 1974, *Reading Statistics and Research*. Harper and Row, New York.

Ishizaki, M., Crocker, M. and Mellish, C.: 1999, Mixed-initiative dialogue using computer dialogue simulation. *User Modeling and User-Adapted Interaction* **9**(1–2), 79–91.

Illényi, A. and Korpãssy, P.: 1981, Correlation between loudness and quality of stereophonic loudspeakers. *Acustica* **49**(4), 334–336.

Jennings, A. and Higuchi, H.: 1993, A user model neural network for a personal news service. *User Modeling and User-Adapted Interaction* **3**(1), 1–25.

Kaplan, C., Fenwick, J. and Chen, J.: 1993, Adaptive hypertext navigation based on user goals and context. *User Modeling and User-Adapted Interaction* **3**(3), 193–220.

Kashihara, A., Hirashima, T. and Toyoda, J.: 1995, A cognitive load application in tutoring. *User Modeling and User-Adapted Interaction* **4**(4), 279–303.

Keppel, G.: 1991, *Design and Analysis: A Researcher's Handbook* 3rd Edn. Prentice-Hall, Englewood Cliffs, NJ.

Krause, J., Hirschmann, A. and Mittermaier, E.: 1993, The intelligent help system COMFOHELP: Towards a solution of the practicability problem for user modeling and adaptive systems. *User Modeling and User-Adapted Interaction* **3**(3), 249–282.

Lipschitz, S. P. and Van der Kooy, J.: 1981, The great debate: Subjective evaluation. *Journal of the Audio Engineering Society* **29**(7/8), 482–491.

London, R.: 1992, Student modeling to support multiple instructional approaches. *User Modeling and User-Adapted Interaction,* **2**(1–2), 117–154.

Marshall, C. and Rossman, G. B.: 1999, *Designing Qualitative Research* 3rd Edn. AltaMira Press, Walnut Creek, CA.

Mathé, N. and Chen, J. R.: 1996, User-centered indexing for adaptive information access. *User Modeling and User-Adapted Interaction* **6**(2–3), 225–261.

Miles, M. B. and Huberman, A. H.: 1994, *Qualitative Data Analysis: An Expanded Sourcebook* 2nd Edn. Sage Publications, London.

Milne, S., Shiu, E. and Cook, J.: 1996, Development of a model of user attributes and its implementation within an adaptive tutoring system. *User Modeling and User-Adapted Interaction* **6**(4), 303–335.

Mitrovic, A., Djordjevic-Kajan, S. and Stoimenov, L.: 1996, INSTRUCT: Modeling students by asking questions. *User Modeling and User-Adapted Interaction* **6**(4), 273–302.

Neter, J., Wasserman, W., and Kutner, M. H.: 1985, *Applied Linear Statistical Models* 2nd Edn. Richard D. Irvin, Homewood, IL.

Newell, S.: 1997, User models and filtering agents for improved internet information retrieval. *User Modeling and User-Adapted Interaction* **7**(4), 223–237.

Nwana, H. S.: 1991, User modelling and user adapted interaction in an intelligent tutoring system. *User Modeling and User-Adapted Interaction* **1**(1), 1–32.

Raskutti, B., Beitz, A. and Ward, B.: 1997, A feature-based approach to recommending selections based on past preferences. *User Modeling and User-Adapted Interaction* **7**(3), 179–218.

Shanefield, D.: 1980, The great ego crunchers: Equalized, double-blind tests. *High Fidelity,* March, 57–61.

Shute, V. J.: 1995, SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction* **5**(1), 1–44.

Silverman, D.: 1993, Interpreting qualitative data: Methods for analysing talk, text and interaction. Sage Publications, Thousand Oaks, CA.

Sison, R. C., Numao, M. and Shimura, M.: 1998, Discovering error classes from discrepancies in novice behaviors via multistrategy conceptual clustering. *User Modeling and User-Adapted Interaction* **8**(1–2), 103–129.

Specht, M. and Kobsa, A.: 1999, Interaction of domain expertise and interface design in adaptive educational hypermedia. In *Workshops on Adaptive Systems and User Modeling on the World Wide Web* at WWW-8, Toronto, Canada and UM99, Banff, Canada.

Stevens, J.: 1992, *Applied Multivariate Statistics for the Social Sciences* 2nd Edn. Lawrence Erlbaum, Hillsdale, NJ.

Tattersall, C.: 1992, Generating help for users of application software. *User Modeling and User-Adapted Interaction,* **2**(3), 211–248.

Vassileva, J.: 1996, A task-centered approach for user modeling in a hypermedia office documentation system. *User Modeling and User-Adapted Interaction* **6**(2–3), 185–224.

Virvou, M. and du Bulay, B.: 1999, Human plausible reasoning for intelligent help. *User Modeling and User-Adapted Interaction* **9**(4), 323–377.
Webb, G. I. and Kuzmycz, M.: 1996, Feature based modelling: a methodology for producing coherent, consistent, dynamically changing models of agent's competencies. *User Modeling and User-Adapted Interaction* **5**(2), 117–150.
Webb, G. I., Pazzani, M. J. and Billsus, D.: 2000, Machine learning for user modeling. *User Modeling and User-Adapted Interaction* **11**(1–2), 19–29 (this issue).
Weber, R. P.: 1990, *Basic Content Analysis* 2nd Edn. Sage Publications, Beverly Hills.
Yin, R. K.: 1994, *Case Study Research: Design and Methods* 2nd Edn. Sage Publications, London.
Zukerman, I. and Albrecht, D. W.: 2000, Predictive statistical user models for user modeling. *User Modeling and User-Adapted Interaction* **11**(1–2), 5–18 (this issue).
Zukerman, I. and Litman, D.: 2000, Natural language processing and user modeling: Synergies and limitations. *User Modeling and User-Adapted Interaction* **11**(1–2), 129–158 (this issue).

Authors' Vita

David N. Chin is a Professor of Information and Computer Sciences at the University of Hawaii at Manoa. He received his B.S. degrees in Physics and in Computer Science/Engineering from M.I.T. and his Ph.D. degree in Computer Science from the University of California, Berkeley. Dr. Chin has worked in user modeling, natural language processing, intelligent interfaces, intelligent agents, geographic information systems, software maintenance, and empirical evaluation.