

Damaged Merchandise?

A review of experiments that compare usability evaluation methods

Wayne D. Gray, Ph.D.
Marilyn Salzman
Human Factors & Applied Cognition
George Mason University

Assertions

- Demand in HCI for quick answers to difficult questions
- To meet this demand, the field of HCI has uncritically touted flawed results and misleading conclusions as guidance for practitioners
- Shortcutting the scientific method is no virtue and that saying that we “don’t know” or that we need more time and resources to do valid studies is no vice

Scope of Review

- To examine validity issues in experiments performed to compare the effectiveness of usability evaluation methods

> KEYWORDS

- EXPERIMENTS
- UEMs (“you-ems”)
- VALIDITY ISSUES
 - cause and effect
 - generalizability

Studies Reviewed (1)

■ The Big Five

- Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). User interface evaluation in the real world: A comparison of four techniques. *CHI'91* (pp. 119-124). New York: ACM Press.
- Desurvire, H. W., Kondziela, J. M., & Atwood, M. E. (1992). What is Gained and Lost when Using Evaluation Methods Other than Empirical Testing, Proceedings of the *HCI'92 Conference on People and Computers VII*, (pp. 89-102).
- Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of Empirical Testing and Walkthrough Methods in User Interface Evaluation. *CHI'92* (pp. 397-404). New York: ACM Press.
- Nielsen, J. (1992). Finding Usability Problems Through Heuristic Evaluation. *CHI'92* (pp. 373-380). New York: ACM Press.
- Nielsen, J., & Phillips, V. L. (1993). Estimating the Relative Usability of Two Interfaces: Heuristic, Formal, and Empirical Methods Compared. *INTERCHI'93* (pp. 214-221). New York: ACM Press.

Summary of UEMs and participants

	Heuristic Evaluation	Cognitive Walkthru	Guidelines	User Testing	Other	
Jeffries, et al., 1991	HF Experts	SW Eng	SW Eng	PC users		
Karat, et al., 1992				mixed	team/wt-mixed	ind/wt-mixed
Desurvire, et al., 1992	HFES/SWEs	HFES/SWEs		18-25yrs		
Nielsen & Phillips, 1993					GOMS-students	HEst-HFESs
R. W. Bailey, et al., 1992	M&N90			students		
Cuomo & Bowen, 1994	dbl HFES	dbl HFES	dbl HFE	domain Exps IBM employees	Forum-IBM	Beta-IBM
Smilowitz, et al, 1993					TAL/UT-students	
Virzi, et al., 1993	dbl HFES			students		
G. Bailey, 1992; 1993				retired missionaries	HFES	SWEs
Desurvire & Thomas, 1993					PAVE-HFES/SWEs/ non-Experts	
Nielsen, 1992	HFES/dbl HFES/nov					

gray@gmu.edu

Damaged Merchandise?

Overview of talk

- The *unique* role and burden of *experiments* in empirical studies of HCI
- Introduction to five threats to validity
- Body of talk
 - Details of threat
 - Examples showing how 1-2 of the big four fail on this threat
 - Summary of all 5 studies on the threat
- Conclusions

gray@gmu.edu

Damaged Merchandise?

The *unique* role and burden of *experiments* in empirical studies of HCI (1)

- “The unique purpose of experiments is to provide stronger tests of causal hypotheses than is permitted by other forms of research” (Cook & Campbell, 1979, p. 83).

gray@gmu.edu

Damaged Merchandise?

The *unique* role and burden of *experiments* in empirical studies of HCI (2)

- Cause and effect
 - Experiments are conducted to determine the effect of some independent variable on some dependent variable
 - Beyond correlation --> causality
- Generality
 - Is the effect found limited to the exact circumstances of the study or can it be generalized to other circumstances?

gray@gmu.edu

Damaged Merchandise?

The *unique* role and burden of *experiments* in empirical studies of HCI (3)

- Few studies are generalizable across all times and places; many have one or more limits in their claim to causality
 - I.E., science is cumulative. Except in trivia cases, all validity issues can **not** be resolved in one experiment
- ➔ HOWEVER, there seems to be something about conducting a UEM study that “causes” researchers to ignore all limits!

Five threats to validity

- Cause-effect issues (was the study done well?)
 - (1) statistical conclusion validity
 - (2) internal validity
- Generality issues (how far can we generalize the findings?)
 - (3) construct validity
 - (4) external validity
- Fifth type of validity (beyond C&C?)
 - ★ (5) conclusion validity

Plan for body of the talk

- Introduce a threat
- Discuss how it is manifested in one or two of the “big four”
- Show a summary table of how it is manifested throughout the 5 studies we reviewed

Threat #1: Statistical Conclusion Validity

- Are there real differences between groups?
- Threats
 - low statistical power
 - random heterogeneity of variance (the *Star Effect*)
 - too many comparisons (fishing)

SCV: Data Summary

	comparisons	respondents per group	Type of statistics		
			parametric	non-parametric	eyeball
Jeffries, et al., 1991	17	4-3-3-6(UT)	v. few	none	most
Karat, et al., 1992	24	6-6-6-6-6-6	v. few	few	most
Desurvire, et al., 1992	54	3-3-3-3-3 ^a each; 18(UT)	none	none	all
Nielsen & Phillips, 1993	na	12-10-15-19-20(UT)	few but variance reported	none	most
Nielsen, 1992	14 ^b	31-19-14	some	none	many

SCV: Threat Summary

	statistical power	random heterogeneity of variance	fishing
Jeffries, et al., 1991	low	v. likely	likely
Karat, et al., 1992	moderate	likely	unlikely
Desurvire, et al., 1992	low	v. likely	yes
Nielsen & Phillips, 1993	adequate	unlikely	n/a
R. W. Bailey, et al., 1992	moderate	moderate	no
Cuomo & Bowen, 1994	low	v. likely	yes
Smilowitz, et al, 1993	adequate	moderate	unlikely
Virzi, et al., 1993	moderate	moderate	likely
G. Bailey, 1992; 1993	adequate	moderate	no
Desurvire & Thomas, 1993	low	v. likely	yes
Nielsen, 1992	adequate	unlikely	likely

Threat #2: Internal Validity

■ Are observed (statistically valid) differences causal as opposed to correlational?

■ Threats

- Instrumentation: A concern whenever human evaluators are used to rate usability problems
- Selection
 - general: characteristic pertaining to general background or experience of participants (e.g., age and experience when those are NOT being studied)
 - specific: past experience directly relevant to the UEM, software, or some other aspect of the experimental design (e.g., participants in different groups have different prior familiarity with software being evaluated)

Internal Validity: Threat Summary

	instrumentation	selection	
		general	specific
Jeffries, et al., 1991	bad	bad	bad
Karat, et al., 1992			likely
Desurvire, et al., 1992	bad		
Nielsen & Phillips, 1993		bad	bad
R. W. Bailey, et al., 1992			
Cuomo & Bowen, 1994			
Smilowitz, et al, 1993			
Virzi, et al., 1993			
G. Bailey, 1992; 1993			
Desurvire & Thomas, 1993	bad		
Nielsen, 1992			

Threat #3: Construct Validity

- Are the experimenters manipulating what they claim to be manipulating (the *causal* construct)
- Are they measuring what they claim to be measuring (the *effect* construct).
- Threats
 - Causal construct validity
 - Effect construct validity
 - Interaction of different treatments
 - Mono-operation and mono-method bias

Causal Construct Validity: Example

- Jeffries, et al. (1991) point out that one difference between their use of HE and Nielsen and Molich's (1990) use is that they used UIS whereas Nielsen & Molich, "proposed that software developers apply the technique" p.120.
- Virzi, Sorce, and Herbert (1993) point out that it is Jeffries, et al.'s "guideline" condition and not their HE condition that comes closest to what Nielsen and Molich (1990) would call HE .
- In an attempt to clarify this issue, in an email exchange we specifically asked Jeffries: "Your description of the HE sounds much like an expert review, during which the experts freely explored the interface to identify problems." She replied, "Yes, it was an expert review" (R. Jeffries, personal communication, May 18, 1995).
- Note that even the father of HE shares this confusion over what Jeffries, et al. (1991) did and whether it was an HE . For example, Nielsen (1992) claimed that Jeffries, et al.'s provided, "independent research [that] has found HE to be extremely cost-efficient" p.373.

Effect Construct Validity

- Two parts:
 - how usability is defined
 - how it is measured
- Complications:
 - Surplus construct irrelevancies
 - Construct underrepresentation

Effect Construct Validity

An important distinction for analytic UEMs is the difference between intrinsic versus pay-off (Scriven, 1977) evaluation of usability.

If you want to evaluate a tool . . . say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axeman. (p. 346)

Effect Construct Validity?

Intrinsic versus payoff evaluation?

- “HE picks up minor usability problems that are often not even seen in actual user testing” (p. 378, Nielsen, 1992).
- “Seventeen of the 40 core usability problems that had been found by HE were confirmed by user test” (p. 45, Nielsen, 1994).
- Nielsen argues that those not found were not false alarms for HE but were due to the characteristics of the users who were involved in user testing, “it would therefore be impossible to find these usability problems by user testing with these users, *but they are still usability problems*” p. 46.

Effect Construct Validity!!

To paraphrase Scriven (1977), a valid measure of usability “proceeds via an examination of the effects of the interface on user performance, and these alone.”

Do the effects measured by HE have construct validity?

- R. W. Bailey, et al., 1992
- Did usability test on MANTEL interface (from Molich & Nielsen, 1990)
- Five conditions
 - > original MANTEL
 - > MANTEL + 2 changes
 - > MANTEL + 4 changes
 - > MANTEL + 5 changes
 - > ideal-MANTEL (all 29 of M&N’s changes)
- Only two changes (first two) had any effect on user performance

Interaction of different treatments

- Participants per group in Desurvire et al., 1992.

	HE	CW	Total
HF Experts	3	3	6
non-experts	3	3	6
Software Engineers	3	3	3

Construct Validity: Threat Summary

	causal construct validity	effect construct validity	interaction of different treatments	mono-op & mono-method bias
Jeffries, et al., 1991	not HE	likely		
Karat, et al., 1992		moderate	bad	good
Desurvire, et al., 1992	not examined	not examined	bad	
Nielsen & Phillips, 1993		likely		
R. W. Bailey, et al., 1992				good
Cuomo & Bowen, 1994				
Smilowitz, et al., 1993				
Virzi, et al., 1993		likely		good
G. Bailey, 1992; 1993		good		
Desurvire & Thomas, 1993	not examined	not examined	bad	
Nielsen, 1992	likely	likely		

Threat #4: External Validity

- External validity concerns generalizing to particular target persons, settings, and times, and generalizing across types of persons, settings, and times (Cook & Campbell, 1979). The distinction is between generalizing to a population versus across subpopulations.
- Claims that exceed the scope of the settings and persons that the experiment can generalize to or across are said to lack external validity.
- Replication!!!

Threat #5: Conclusion Validity

- When problems with the first four types of validity are disregarded, are the claims consistent with the results and/or do they follow from what was done?
 - does not apply to claims that are invalid due to one of the four Cook and Campbell validity problems
 - does apply when a claim that is stated as a logical conclusion of the study was either (1) not investigated in the study or (2) contradicted by the results of the study

External Validity & Conclusion Validity: Threat Summary

	External Validity	Conclusion Validity
Jeffries, et al., 1991	problem	concern
Karat, et al., 1992	problem	problem
Desurvire, et al., 1992	problem	problem
Nielsen & Phillips, 1993		problem
R. W. Bailey, et al., 1992	good	
Cuomo & Bowen, 1994	problem	
Smilowitz, et al., 1993	well handled	
Virzi, et al., 1993	well handled	
G. Bailey, 1992; 1993	well handled	
Desurvire & Thomas, 1993	problem	problem
Nielsen, 1992		concern

Jeffries et al., 1991: Goals

- Compare four UEMs they call Heuristic Evaluation, Cognitive Walkthrough, guidelines, and user testing to determine
 - (1) how the techniques compare
 - (2) the interface problems the techniques best detect
 - (3) who (developers or UIS) can use the techniques most effectively
 - (4) the relative costs and benefits of each technique.
- Due to a combination of threats to validity, the study falls far short of being able to reach conclusions about any of these factors.

Jeffries et al., 1991: A sample of claims made

- “Overall, the [H]euristic [E]valuation technique as applied here produced the best results. It found the most problems, including more of the most serious ones, than did any other technique, and at the lowest cost” p. 123.
- Heuristic Evaluation is dependent upon “having access to several people with the knowledge and experience necessary to apply the technique” p. 123.
- “Another limitation of [H]euristic [E]valuation is the large number of specific, one-time, and low-priority problems found and reported.”
- “Usability testing did a good job of finding serious problems . . . and was very good at finding recurring and general problems, and at avoiding low-priority problems.”
- User testing “was the most expensive of the four techniques. . . . despite this cost, there were many serious problems that it failed to find.”
- “The guidelines evaluation was the best of the four techniques at finding recurring and general problems.”

Jeffries et al., 1991: A sample of claims made

- “Overall, the [H]euristic [E]valuation technique as applied here produced the best results. It found the most problems, including more of the most serious ones, than did any other technique, and at the lowest cost” p. 123.
- Heuristic Evaluation is dependent upon “having access to several people with the knowledge and experience necessary to apply the technique” p. 123.
- “Another limitation of [H]euristic [E]valuation is the large number of specific, one-time, and low-priority problems found and reported.”
- “Usability testing did a good job of finding serious problems . . . and was very good at finding recurring and general problems, and at avoiding low-priority problems.”
- User testing “was the most expensive of the four techniques. . . . despite this cost, there were many serious problems that it failed to find.”
- “The guidelines evaluation was the best of the four techniques at finding recurring and general problems.”

Jeffries et al., 1991: What can be concluded?

We maintain that there is nothing that can be safely concluded from the study as conducted

Karat et al., 1992: Goals

- Attempt to compare user testing with a walkthrough technique to:
 - (1) compare for each UEM the number of problems found, their severity, and resources required
 - (2) determine if differences between UEMs can be generalized across systems
 - (3) examine the characteristics of walkthroughs (individual vs. team evaluations, evaluator characteristics, task scenarios vs. self exploration, and heuristics vs. no heuristics) as they influence effectiveness. (N.B., the issue of evaluator characteristics was not part of the experimental design)

Karat et al., 1992: A sample of claims made

1. User testing "identified the largest number of problems, and identified a significant number of relatively severe problems that were missed by the walkthrough conditions" p. 402.
2. Jeffries, et al. (1991), Desurvire, et al. (1991), as well as the current study "provide strong support for the value of [user interface] expertise" p. 403.
3. "These methods [user testing and walkthroughs] are complementary and yield different results; they act as different types of sieves in identifying usability problems" p. 403
4. "Team walkthroughs achieved better results than individual walkthroughs in some areas" p. 403.
5. "All walkthrough groups favored the use of scenarios over self-guided exploration in identifying usability problems. This evidence supports the use of a set of rich scenarios developed in consultation with end users."
6. "The results also demonstrate that evaluators who have relevant computer experience and represent a sample of end users and development team members can complete usability walkthroughs with relative success" p. 403

Karat et al., 1992: A sample of claims made

1. User testing "identified the largest number of problems, and identified a significant number of relatively severe problems that were missed by the walkthrough conditions" p. 402.
2. Jeffries, et al. (1991), Desurvire, et al. (1991), as well as the current study "provide strong support for the value of [user interface] expertise" p. 403.
3. "These methods [user testing and walkthroughs] are complementary and yield different results; they act as different types of sieves in identifying usability problems" p. 403
4. "Team walkthroughs achieved better results than individual walkthroughs in some areas" p. 403.
5. "All walkthrough groups favored the use of scenarios over self-guided exploration in identifying usability problems. This evidence supports the use of a set of rich scenarios developed in consultation with end users."
6. "The results also demonstrate that evaluators who have relevant computer experience and represent a sample of end users and development team members can complete usability walkthroughs with relative success" p. 403

Karat et al., 1992: What can be concluded?

- All in all, we feel forced to conclude that although this study promised much, upon close reading it offers little that we can take away or recommend to practitioners.

Desurvire, et al., 1992: Goals

1. to compare the effectiveness of three types of evaluators; Human Factors Experts (HF Experts), Software Engineers, and non-experts on two analytic UEMs, Heuristic Evaluation and Cognitive Walkthrough
2. to compare all of the above with user testing
3. to determine differences in the types of problems identified by the different methods and evaluators
4. to determine what user testing finds that the analytic UEMs do not and, conversely, what analytic UEMs find that user testing misses.

Desurvire, et al., 1992: A sample of claims made

1. "HE is a better method than the Cognitive Walkthrough for predicting specific problems that actually occur in the laboratory, especially for Experts" (pp. 98-99).
2. "Experts in the HE condition named almost twice as many problems that caused task failure or were of minor annoyance in the laboratory, than Experts in the cognitive condition" (p. 99).
3. "HE seems to facilitate the identification of potential problems and improvements that go beyond the scope of the tasks, more so than the CW" p. 99.
4. "Experts focused on problems that violate the heuristic, 'provide feedback', where they are more focused on the user's interaction with the system than the SWE who

Desurvire, et al., 1992: A sample of claims made

1. "HE is a better method than the Cognitive Walkthrough for predicting specific problems that actually occur in the laboratory, especially for Experts" (pp. 98-99).
2. "Experts in the HE condition named almost twice as many problems that caused task failure or were of minor annoyance in the laboratory, than Experts in the cognitive condition" (p. 99).
3. "HE seems to facilitate the identification of potential problems and improvements that go beyond the scope of the tasks, more so than the CW" p. 99.
4. "Experts focused on problems that violate the heuristic, 'provide feedback', where they are more focused on the user's interaction with the system than the SWE who

Desurvire, et al., 1992: What can be concluded?

- The authors fail to recognize the limitations of their study and base many strongly worded conclusions upon scant data. The prerequisites for an experimental study, statistical conclusion validity and internal validity, are severely lacking. We conclude that *there is nothing that can be safely concluded from the study as conducted.*

Nielsen & Philips, 1993: Goals

■ To compare

1. performance time estimates
2. the relative costs of four analytic UEMs and user testing. The four analytic UEMs are: Cold, Warm, and Hot heuristic estimates ; and GOMS analysis.

Nielsen & Philips, 1993: A sample of claims made

1. & 2. "User testing was also much more expensive than 'cold' heuristic estimates and somewhat more expensive than GOMS analyses" pp. 220-221.
"GOMS and heuristic estimates were about equal for relative estimates, so heuristic estimates might be recommended based on its lower costs" p. 221.
3. "Heuristic estimates were better in the hot condition where estimators had access to running versions of the two interfaces, than in the cold condition based on specifications only" p. 221.
4. "Performance estimates from both heuristic estimation and GOMS analyses are highly variable" p. 221

Nielsen & Philips, 1993: A sample of claims made

1. & 2. "User testing was also much more expensive than 'cold' heuristic estimates and somewhat more expensive than GOMS analyses" pp. 220-221.
"GOMS and heuristic estimates were about equal for relative estimates, so heuristic estimates might be recommended based on its lower costs" p. 221.
3. "Heuristic estimates were better in the hot condition where estimators had access to running versions of the two interfaces, than in the cold condition based on specifications only" p. 221.
4. "Performance estimates from both heuristic estimation and GOMS analyses are highly variable" p. 221

SCV & Internal validity

conclusion validity

conclusion validity

Nielsen & Philips, 1993: What can be concluded?

Our conclusions from this article are very different than the authors'.

1. the Warm group must be ignored.
2. the Cold heuristic estimation group does about as well as the Hot heuristic estimation group
3. with training and expertise stacked against them, the GOMS group did amazingly well. (Indeed, with four replications, the accuracy of the GOMS predictions for this particular interface may be the one finding in the entire HCI literature with the greatest external validity.)
4. Contrary to our personal intuitions, GOMS can be performed quite accurately with very little training or experience.

Conclusions (1)

- Cause-effect issues (SCV & Internal validity)
 - Experimental research in HCI has wrongly adopted the methodological standards of user testing

Conclusions (2)

- Construct validity of effect
 - Usability is wrongly treated as a monolithic, atheoretical construct
 - Purpose of analytic UEMs is to predict usability pay-offs (the “speed of the cuts”) from an examination of intrinsic attributes (“the grade of hickory in the handle”).
 - No one-to-one correspondence
 - correspondence cannot be assumed but is a theoretical question to be resolved by empirical methods.

Call to action!

- Researchers
 - must be willing to put an increasing amount of time, effort, and resources into experiments of lesser scope and more restricted conclusions
- Reviewers
 - must demand more than a timely, controversial, or fresh finding; they must insist that the foundation upon which the findings rest, the design and methodology of the experiment, be reliable and valid
- Practitioners
 - must learn to suspect easy answers to difficult questions and to come to believe that research at a bargain price may be simply damaged merchandise.