
Make Evaluation Poverty History

Gilbert Cockton

School of Computing and
Technology, University of
Sunderland, St. Peter's Way,
Sunderland SR6 0DD, UK
Gilbert.Cockton@sunderland.ac.uk

Abstract

A Tyranny of Evaluation could be palatable if it wasn't so impoverished. Henry Lieberman complained at the CHI 2003 Fringe about having to do bad evaluation. As an evaluation evaluator, I sympathise, but rather than banish skinny evaluation warlords, I'm trying to fatten them up and civilise them. Like a hungry wolf, evaluation can savage and devour design, but design must feed evaluation to tame it. I've drawn up a new HCI menu for this, but I've not managed to cook and taste much yet. It's a banquet too large for my local research kitchen. Adding a few super chefs and their teams won't be enough. It's now time for CHI 2.0 – high bandwidth collaborative social networking for the biggest method bake in HCI history.

Keywords

Worth Systems, Worth/Aversion Maps, Total Iteration Potential, Worth-Based Evaluation, Menu Design.

ACM Classification Keywords

H.1.2 User/Machine Systems; D.2.9 Management

Henry Lieberman vs. Shumin Zhai

Henry, an MIT IUI guy, presented his *Tyranny of Evaluation* [13] at the CHI 2003 Fringe. Shumin, an IBM input device guy retorted [22]. It all goes back to their Grad Schools. Henry did computer science, Shumin did Human Factors. They learned different

Copyright is held by the author/owner(s).

CHI 2007, April 28 – May 3, 2007, San Jose, USA

ACM 1-xxxxxxxxxxxxxxxxxxxx.

stuff. When CHI reviewers started wanting Henry to do Shumin stuff, he was unhappy: “our methodologies for quantitatively evaluating user interfaces suck”. Henry would happily test a bit, but not live out evaluation warlords’ dreams: “Don't get me wrong. ... I do user testing. I think you can learn a lot of things from user testing. But, in my experience, it's chancy. Everybody's got a wonderful anecdote about how user testing uncovered some surprising and important ... When that happens, it's great. But ... you can't depend on it, and you can't insist on it.”

Shumin worried that a lack of evaluation would “turn HCI into a ‘faith-based’ enterprise’ ... Granted, quantitative and controlled experiments or other forms of evaluation ... can be burdensome and limited, and ... seem not to be worth the time and effort.” And yet this “by no means should imply we can discard evaluation and simply accept inventors’ and designers’ claims.” So time-wasting and worthless activity beats faith?

How did we get here?

Unusually, for much of HCI history, most designers didn't formally evaluate and evaluators didn't design much. Software development could be even worse, with no-one designing or evaluating, but instead *making*, a craft process of ‘play with the clay’. The idea of evaluation as something separate from design is very disconcerting. Designers *do* evaluate. You simply cannot design without doing it. If you choose one design option from several alternatives, then you must evaluate them to choose. Now, designers may make a bad job of this, but they still do it. Also, it's not as if that the current ‘good’ job is much better [22].

Empirical human sciences have dominated HCI research and practice: cognitive psychology in first wave HCI; ethnography in the second [3]. Both have acted as ‘design police’, albeit with different truncheons, and similar views are found in software industry as well as HCI research [11]. From the dawn of HCI, designers and developers have complained about the largely negative impact of much evaluation: a list of things that have gone wrong, but little credible and/or actionable advice on how to fix them [17].

There is evidence from several sources that HCI's longstanding *split* between design and evaluation is becoming an *overlap*. Chapters by Kasper Hornbæk and Stephanie Rosenbaum in a forthcoming volume on usability [12] both cover changes in usability practice, with practitioners moving from finding problems, to suggesting fixes, and even providing user research and design input throughout development. In new media and design agencies, one increasingly sees practitioners span the whole interaction design skill set from user research through design to user testing. Even so, the HCI methods that they use are still not well integrated, and it is not clear if they can be in their original forms.

Largely imported from outside disciplines with little tradition of design research, some HCI methods are not fit for the purpose for which they get used. This is reflected in many papers on making ethnography more relevant to design or improving the downstream utility of usability evaluation. Good interaction designers and usability practitioners do successfully adapt research methods to local needs, but this is hardly the basis for a mature research discipline. Firstly, we should not leave so much ‘finishing off’ to practitioners. Secondly, as Lieberman complained (and Zhai effectively agreed),

researchers are not allowed to depart too far from 'proper' uses of methods from psychology and sociology. If they do, they risk critical rejection from reviewers with psychology or sociology backgrounds.

How bad is it?

It's not that bad, but it does matter. It's not bad, because even inappropriate hit and miss empirical methods can be better than no empirical methods at all. It will remain 'business as usual' until we have something better, but we must have something better. Progress matters for all the reasons covered by Henry Lieberman. Design is about creativity and innovation, not about negativity and repression. We need evaluation approaches that are more positive in outcome [17] and more reliably effective in impact.

Evaluation gets stuck

In 1988, the first Handbook of HCI published a chapter on usability engineering [20]. Whiteside, Bennett and Holtzblatt summarised their experience at IBM and DEC of psychologically motivated evaluations, and how they evolved towards a more contextual ethnographically informed approach. This marked the beginning of second wave HCI. In many ways the usability engineering that the IBM and DEC teams *abandoned* is as sophisticated as much current practice (and much better than much current research). However, targets based on experimental psychology variables were a double-edged sword. They enabled developer focus and measurement of progress, but also failed "to reflect usability as it will be judged in practice": "it had better be true that the specified goals are the ones users really want". This was often *not* so, so this group of usability engineers at DEC and IBM moved away from experimental psychology measures in the hope that

usability specifications and objectives could "be defined through an interpretation of data from the field". Despite the long evolution of Contextual Design since, such a move from contextual data to usability objectives has never been demonstrated. Without objectives that express design purpose within a usability perspective, evaluation cannot properly focus.

Putting evaluation on a new diet

What is the purpose of evaluation? Distinctions between formative and summative evaluation and their different purposes are commonplace, but not necessarily well grounded. Formative evaluation's current purpose is to drive iteration by finding problems and contributing to solutions (it is a category mistake to expect evaluation to be the sole source of re-design proposals). Summative evaluation's purpose is seen as measuring how well a design performs relative to targets. However, we need to know *what matters*, and express that as evaluation targets. Summative evaluation can also be used to drive iterations, although with little qualitative data on *why* things aren't good enough. Formative evaluation tends to better isolate such causal insights.

An alternative view is that *purpose of evaluation is to assess achievement of design purpose*. This highly relative definition admits no proper evaluation without clear prior indication of design purpose. It cannot succeed unless *design feeds evaluation*. Evaluation poverty, and the resulting anti-social behaviour, is the fault of design. Evaluation needs to be nourished by a full menu of design purpose.

Design rises above 'making' through reflection on alternatives. It can do so without fixing purpose,

despite a stream of goals, aims, objectives or visions in design briefs, proposals and 'crit' sessions. It may not be until planning a product launch that a "value proposition" finally crystallizes. The driver here is more often what consumers will believe and desire rather than what a product really can deliver.

Evaluation must feed off a diet of design purpose cooked up early, that lets purpose be tested, revised and validated, and be associated with design options and evaluation measures. We need a new method diet.

Why is our method diet inadequate?

Figure 1 maps design using 'Whose Advantage' as latitude, and 'Design Canvas' as longitude. The idea is that different design approaches occupy different areas relative to these axes. An approach can be well suited to a blank canvas, or to filling specific gaps. Purpose may be primarily for its creator, or be genuinely for the advantage of others. No design approaches are at extreme points, but instead occupy an area.

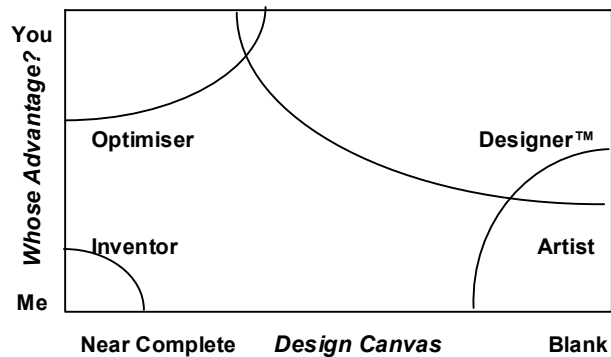


Figure 1: IADO Map of Design Areas

The areas are: Inventor, Artist, Designer and Optimiser (hence IADO). All are genuine design approaches. *Inventors* rarely get it right first time, and thus design even though they may only be guided by their own view of what is valuable. Furthermore, inventors tend to attack well understood and scoped problem spaces (e.g., the better mouse trap). They rarely work on a blank canvas, but focus on problem aspects of existing designs (e.g., power in a clockwork radio). Invention is well represented in HCI research, although as

Lieberman complained, it is a target for evaluation tyrants, with only some SIGCHI conferences being inventor friendly (e.g., UIST, IUI).

Artists by contrast operate with potentially no constraint, but like inventors, may primarily create something for themselves rather than others. They may have an idea of what people could want or like, but creation can be as much about discovering what these could be, of finding out about ourselves, rather than selecting a design alternative that best serves a predefined purpose. HCI embraced art-based approaches with the invention of cultural probes. These were followed by technology probes [10], designed more to inform *artist-designers* than to serve probed users. This "take it or leave it approach" has been criticised [3], but it has been very effective at exploring wants, needs, experiences and responses that may not emerge from extensive systematic field research prior to presenting designed artefacts.

Design in HCI however remains overwhelming that of the *optimiser*. This approach of much engineering design relies on an well-populated design canvas with only a few gaps to fill in with the best value for some critical parameters. Psychology is largely the basis for HCI engineering design, filling the CHI literature with experiments that identify optimal values for parameters for some user-task combination. The thresholds and measures used are based on an understanding of user needs. This may be flawed, but even so, the engineering design is more focused on user advantage than invention, and can calculate it.

On the IADO map, the only area that is most 'fully' design (hence [™]) combines a potentially blank canvas

with an explicit desire to meet understood user needs. Such design does not do well at CHI. For any non-trivial design, there are many understandings of needs and wants to be explained in terms of their field research grounding, many design decisions to be rationalised in terms of design purpose, and many demonstrated associations between design features and usage outcomes to explain in just 10 pages.

Full design currently fares less well than poorly grounded and often mistargeted premature optimisations, cool technical invention and now art-inspired probes. They are all good, but skip the hard part. Optimisation is most valuable in the context of a well structured and well grounded design. In turn, full design depends on craft materials (invented by others) and identification of what people could value and how (often first discovered by artists free of burdens of scientific methods). However, the value of inventions and art-inspired insights has to be realised through full integrating and holistic design, after which optimisation becomes worthwhile.

We thus do have a useful and fairly comprehensive toolbox of HCI methods, with much evidence of successful invention, art-mediated insights and engineered optimisations. However, we have a gaping hole for creation and maintenance of a focus on value throughout development. We need to express intended value, to form direct traceable and assessable associations to this value, and to evaluate its achievement. In some circles, 'value' is now coming to replace design as the defining term for full design. Thus the UK Design Council's Director of Design and Innovation, Richard Eisermann states [18]: "When I

talk about design, I try not to mention the 'd' word anymore. I try to talk about value".

Without tools for expressing intended value and tracking achieved value, evaluation remains hungry and impoverished. As Henry Lieberman argued [13], inventors can do well without empirical evaluation. There have been centuries of successful innovation without systematic empirical evaluation across the whole development lifecycle.

Art-based inquiries can only fail if they commit in advance to a definition of success. Without this, there is limited scope for evaluation. Like much contemporary art, technology probes have few well defined goals and expectations [10], but discoveries about usage have value. Some support formative evaluation and shape the next experimental probe. Others tell us something about the technology, others about people. From this perspective, the yield is much wider than evaluations focused on design and interaction quality.

Optimising approaches are well supported by experimental methods. Companies such as Amazon have now automated much of this with AB tests, allocating alternative designs to a sample of visitors and measuring resulting dependent variables. Aspects of first wave HCI experimental skill sets remain important and effective here. CHI conferences continue to include examples of psychological experiments being successfully applied to the optimisation of real world problems, for example, human interaction proofs during web email account creation (e.g., [4]).

Experiments outside of comprehensive design contexts are premature (note that [4] had extensive usage and security attack data from a massively used system). We do not know *what really matters* at the outset of invention or artistic creation. Inventors may think they do, while artist-designers may wait and see. HCI has paid limited attention to the top right space of full design, except when reviewers unfairly demand that Henry Lieberman and his ilk cross a diagonal to where they never meant to be.

HCI has thus drawn successfully on technical invention, experimental psychology, and more recently art to create, explore and optimise elements of designs. However, effective evaluation of designs should be in terms of purpose, rather than in terms of constructs that experimental psychologists know how to measure. To extend Whiteside and collaborators [20], targets chosen may not just be “ones users really want” but also ones that *full designers* have no interest in, lacking credible relations to design purpose. In fairness to evaluators however, getting fed with design purpose during evaluation planning requires a lot of hunting. Purpose evolves as designs develop. Few designers readily encapsulate their goals and aspirations, especially for complex interactive digital products and services. Business roles in a product team can be even less able to define measurable critical success factors!

Should I have started here?

Six pages in and I’m just about to talk about possible solutions. I could have started here, but when I’ve done this before, reviewers tell me that the problems that I’m trying to solve don’t exist. The purpose of the long preamble above is to argue that we do have a big problem. I’ve also learned to make it clear that HCI

isn’t a total disaster. When I have written previous such preambles, I get told that things aren’t that bad. Well of course they aren’t *that* bad, but they are *this* bad. I hope I’ve been clear this time on what *this* is.

Enriching Evaluation

Quite simply, we cannot fully fix evaluation until we have fixed full design. They are opposite sides of the same coin and we cannot expect to address either in isolation once the purpose of evaluation is to assess achievement of design purpose. This is impossible if design purpose is not clear in terms of the value that it intends to achieve in the world. Clear design purpose enables full evaluation, which in turn, enables clearer and more focused design. It’s a virtuous circle, but it can’t start until evaluation gets better fed by design. This, perversely, is why evaluation is currently so impoverished. At extremes, evaluation is expected to survive with no portions of design priorities. No wonder then that lack of value and relevance can result.

Laddering as HCI’s Giant Beanstalk

If you agree it’s bad, you may be interested in solutions. If not, perhaps you’ll just like what follows anyway. If you think it’s not bad, what follows should further reinforce my arguments.

If design is the creation of value, as leading experts claim, then we need to “create meaningful connections among people, ideas, art, and technology, shaping the way people understand their relationships with ... new products” [14]. With evaluation, connections become demonstrable. Designers need to create associations between product features (“art, and technology”) and personal values (“people, ideas”). Evaluators then assess the success of associations.

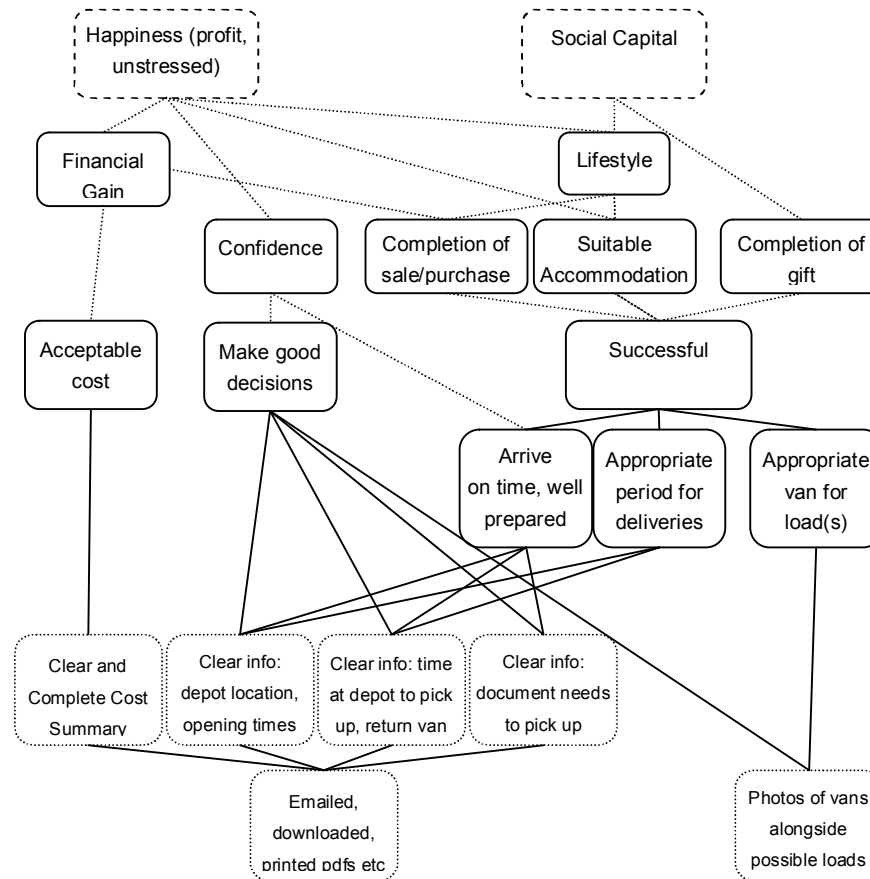


Figure 2: Worth Map for Van Hire

Frameworks for relating product attributes to personal values already exist in consumer marketing [2]. Laddering approaches build means-end chains from product attributes, via usage consequences, to personal values. There are differences over the cognitive

structure of means-end chains, which are based on consumers' beliefs and not usage studies, but HCI is free to make its own choices within marketing debates [7]. Means-end chains tend to converge on similar values; there are fewer of these than potential product attributes. Chains 'join' at points of convergent value to form a *hierarchical value model* [2]

By relaxing strong assumptions about consumer means-end chains, and by combining positive and negative associations (motivators and aversions) into a single model, I have created *worth/aversion maps* (W/AM) to capture desired associations (connections) between a design and its purpose of intended value. Figure 2 shows an example *worth map*. There are no negative associations down over to aversions, for reasons of space. For an example of a W/AM, see [7], as well as for a fuller discussion of the worth map in Figure 2. The lower two rows of boxes in Figure 2 are product attributes, with abstract attributes (qualities) above concrete product attributes (features). The two dashed boxes at the top are personal values. Boxes in between are usage consequences (functional, economic and psychosocial). Together with associations between them, worth map elements (or 'worthies'), express a design team's understanding of product success.

Worth maps unify the purpose of design and evaluation. The purpose of design can be represented as a set of worthies and associations between them, which should hold for a product to succeed. Together, these form a *worth system* that results from the human interactions which deliver intended associations. The purpose of evaluation is to test whether such associations would/do hold. The focus is likely to be on coherent worth *subsystems* rather than the whole map.

Related sets of associations will be evaluated together. The lines representing associations hide considerable complexity. Each is a model of human-computer interaction, where product use achieves desirable qualities that enable worthwhile usage consequences in specific usage contexts. Each association depends on several different human phenomena (e.g., behaviour, physical, economic, cultural and social contexts).

CHI 2.0: A Worth-Centred Menu

A W/AM represents a theory that can be tested and refined during development. The potential of Worth Maps (as W/AMs) is currently demonstrated by the range of new methods that can be based on them, as well as the way they can reform existing methods. Design and evaluations become much more tightly coupled, allowing the latter to feed off the former in ways that are not currently possible.

Throughout development, a W/AM represents the intended and achievable worth of a proposed design. Worth maps can be inspected for desirable qualities. Action can be taken to address weaknesses. The main iterative strategies are to add worthies (and to remove aversions from W/AMs), and to strengthen associations (remove or weaken in the case of negative associations in a W/AM). These generalise transformations of hierarchical value models for product improvement [2]. W/AMs can be inspected for these qualities:

1. Worth Spread
2. Worth Balance
3. Worth Expressiveness

4. Means-End Credibility
5. Worth Evaluability
6. Worth Achievement

These are now briefly described. Worth-based approaches tend to blur distinctions between requirements, design and evaluation. A W/AM can express a current understanding of user needs, wants and aversions (and for innovative designs, dreams and nightmares). It can also express how design decisions are focused on the achievement of worthwhile usage consequences. Lastly, it can record how well associations have really been forged, and the achieved worth that results. Clearly you can't get all of this onto Figure 2. For maps, there should be extensive survey data behind lines, labels and shading, and it should be clear how survey data connects with map features.

Worth Spread

W/AMs should cover as many positive values as possible that are important to individuals and collectives (encounters, institutions, kin, kind, locales [6]), and as few aversions as possible. Worth maps can be inspected using value lists, such as those covered in Value-Sensitive Design [9] those used in primed/prompted laddering such as VALS, LOV and RVS [21], or ones from theological and/or philosophical approaches that construct such lists. Both positive and negative values should be covered.

Worth spread can be a surprisingly simple but effective predictor of design iterations. A workplace information display with a largely functional design purpose was found to also support feelings of connectedness and

relatedness [16]. The extremely abstract motivational categories of existence, relatedness and growth (ERG Theory, [1]) suggest that an exclusive focus on *existence motivators* at work could be too narrow, and it was. The expressed worth in a W/AM for this Whereabouts Clock can thus be *spread* by adding psychosocial consequences and values for *relatedness* motivators. A new worth subsystem would start with existing product attributes, which could be improved and extended to strengthen associations for *relatedness*. A focus on positive motivators as well as aversions (as in much Value-Sensitive Design) broadens consideration of values in design.

Theories of motivation thus provide another source of high level *individual* motivators. Similarly, sociocultural approaches address *collective* motivations by 'reading' individual and collective values 'inscribed' in cultural forms in different worth arenas [6].

Worth Balance

By comparing W/AMs for different stakeholders, the relative impact of a design can be compared, prompting re-design where necessary to both reduce adverse impact on, and increase worth for, some stakeholders. This is a specific case of *raising a W/AM's centre of gravity*, since worthies get added towards the top and aversions get removed towards the bottom.

Worth Expressiveness

W/AMs reduce worthies to node labels, which cannot capture the full extent of intended worth. 'Survey data' is required to communicate 'the full worthy'. For example, direct quotes from interviewed stakeholders and photographs or video material can be organised

into *worth boards* to express intended worth, rather than just the 'mood' of a product [6].

Means-End Credibility

W/AMs suffer from two types of credibility problem. Firstly, diagrams force concrete product attributes to be generalised and grouped. In Figure 2, features are reduced to 'photos of vans alongside loads' and 'emailed, downloaded, printed pdfs etc', features that provide clear information (abstract quality) to support usage consequences of arriving on time and well prepared at the van depot, having arranged an appropriate van and hire period. Good decisions can be made on cost and van hire solutions. We should rightly challenge designers on any of such claims.

Credibility here requires separate detailed *worth tables* to model associations between concrete and abstract attributes. More generally, worth tables are similar in structure to web design *content matrices* and matrix methods from first wave HCI.

Secondly, means-end chains need to credibly extend beyond middling associations between concrete and abstract product attributes. Credibility along the whole means end-chain must be established. While this is ultimately a question of demonstrating achievement of claimed associations for all worth subsystems up a worthwhile ladder, much can be achieved via constructing narratives about envisioned usage.

Worth delivery scenarios take a 'happy ending' approach to scenario authoring, focusing scenarios on credible narratives that demonstrate how worth will be delivered (and aversions avoided) by a design with envisaged concrete and abstract attributes. Happy

endings can be based on worth board items. Such scenarios take their narrative 'skeleton' from means-end chains. They can be used to assess and improve credibility, viability and detail of claimed associations. They are essentially *Value Delivery Scenarios* [5] transferred into a W/AM based methodology.

Worth Evaluability

The complete topology of a W/AM is evaluable. Targets can be set for worthies. Associations can be tested. For product attributes, worth tables can include specific evaluation targets for abstract attributes (qualities) and can also include specific criteria to support a checklist approaches to feature inspection.

Planning empirical evaluation must systematically map worthies to evaluation criteria, which in turn are mapped to measures (with targets) and then instruments. It is important to measure above the level of product qualities and immediate functional or psychosocial consequences. These are measured as means to ends, but measures of high value terminal outcomes take precedence. Functional consequences should be measured by observing or instrumenting the world. Economic consequences require financial measures. Psychosocial consequences need research instruments from psychology and sociology.

As we move from measuring W/AM's central zones, more extensive *worth subsystems* are evaluated. Measuring quality in the world requires instrumentation of the world. Many outcomes cannot be measured during interaction. For example, in Figure 2, there are functional consequences of "successful delivery of load(s)", "arrive on time and well prepared", "appropriate period for deliveries" and "appropriate van

for load(s)" that can not be measured during web interactions while booking a van. These must either be measured through post-hire questionnaires via email, and/or via some form of instrumentation at the van hire depot. At this point we no longer just instrument interaction, but instead we must potentially instrument much within the scope of a sociodigital system. Such sociodigital systems become *self-instrumenting*, that is, they are designed to contain within them the means of their own evaluation.

Worth evaluability is thus assessed by examining mappings from worthies to evaluation criteria, and on to measures, instruments and targets. Again, this is all 'survey data' underpinning worth maps. At any time we can assess the quality of evaluation planning. If worthies cannot be measured, then the effectiveness of worth subsystems cannot be assessed. The cost of some instrumentation could be high, and development teams may have to decide to do without. Such decisions determine the *evaluability* of a worth map. Again, it is design that feeds evaluation. If worthies cannot be operationalised for measurement, they cannot be fully evaluated. Associations within worth subsystems can be assessed to some extent by usage studies, but as with all unfocused user testing "it's chancy ... you can't depend on it" [13].

Worth Achievement and Total Iteration Potential

The above methods are all analytical, although *worth evaluability* bridges to empirical evaluation. The worth-based rendezvous between design and evaluation however means that much of the cooking to feed evaluation is complete before empirical studies start. W/AMs provide a framework for all evaluation

reporting, with empirical data as further 'survey data' that can be directly associated with worth subsystems.

Worth-based empirical evaluation validates W/AMs and not just products. Evaluation can be either summative (through instrumented worthies), or formative, refocusing design purpose through iterating worth subsystems within a development process with *total iteration potential* [7]. By iterating W/AMs, design purpose, user research and evaluation planning are all iterated. Iteration is no longer focused exclusively on design features [5].

The worth-based rendezvous between design and evaluation is further advanced by self-instrumentation, which inevitably extends designs to bring consequences beyond those previously in a W/AM. For example, instrumenting student interest in courses on a university web site requires additional interactive content that extends product attributes [5]. This extends both design (concrete attributes) and usage consequences. If prospective students can be persuaded to register, a worth subsystem can be added to strengthen associations between usage consequences of "interest in a course on offer", "students apply for a course with support of parent and/or advisers" and "students apply for course that they will enjoy and complete". This will add further concrete attributes that implement a 'sales pipeline' to improve conversion from initial interest to applications.

Self-instrumentation paradoxically fixes some problems before they could be found by supporting creative and innovative value-adding design extensions as a consequence of planning evaluation measures.

Existing empirical approaches such as probes can be used as W/AM extension methods, where the aim is to use discoveries to add worthies and new associations. *Technology probes* [10] expose usages and valuable outcomes that a design team may not have envisaged. More generally, probes provide access to user appropriations [19], which by their nature create new usage consequences without changing concrete product attributes, although abstract attributes may be enhanced or even created.

In worth-based assessment, we empirically evaluate to test design theories, and not just to identify product defects. After all, the source of product defects are our design theories (however implicit), and not the product itself, which, after all, is only a human artefact, albeit perhaps poorly understood by those who only make.

CHI 2.0: Let the mash-up begin

No discipline has a monopoly on understanding human worth. Nor can a discipline be involved in HCI 'on its own terms'. Each discipline has assumptions and values, and there is no realistic chance of integrating all perspectives in a liberal inter-disciplinary melting pot. A post-disciplinary approach is needed to unlock disciplinary shackles and reject superficial postures of multi-disciplinary cooperation [15]. This is relevant to the debate that Paul Dourish has started [8]. His implicit multi-disciplinary view of HCI, with disciplinary contributions judged on their own terms, offers little worth to designers, who will borrow what they want as they want from existing disciplines. For example, economists have many theories on how value (as worth and/or price) arises. Subjective theories view value for money as a matter of individual judgement. These

appear to be most relevant to HCI, and are most likely to be borrowed, regardless of economists' views.

Designers have no need to apologise for 'selective deployment' from an appropriated 'HCI arsenal' [8]. Designers must appropriate ('borrow'), but there will be gaps in current theoretical understandings of worth. These must be filled by innovating through reflective practice. In post-disciplinary worth-based HCI, it is for designers to borrow and actively adapt what they want and to then fill the gaps, *based on the needs of the design process alone*. Thus, while my approach to categories and arenas of worth [6] may well 'upset' many from academic disciplines who study collective behaviour, breached disciplinary standards *do not matter* if 'oversimplification' delivers of demonstrable benefits for creativity, design and innovation. CHI is our bus, and there's no place for back seat drivers.

Designers and evaluators are on largely their own. They cannot expect to import 'results' uncritically and passively from disciplines of individual and collective worth. Dourish correctly argues that there can be no implications for design, but not just for ethnography: there have never been direct links from psychology to design either. There are implications, and links to, *existing* designs, and this is not splitting hairs. Different disciplines and theories make different contributions to our understanding of W/AM associations, but do not give rise directly to designs. That is a job for full-time full designers.

The challenge facing worth-based design is thus to develop new methods that combine design and evaluation, borrowing on our own terms from a wide range of scientific, management and cultural disciplines

and practices. Such a process will not have an end point when methods can be summatively validated. New methods must be continuously developed and refined in the context of full designs, which places assessments of use beyond 10 page CHI papers.

Archived 'results' have little to contribute to the development of full scale design and evaluation methods. With the social networking of Web 2.0, we now have a better medium to share experiences and examples, working collaboratively to improve methods. While CHI waits patiently for a properly validated third wave, CHI 2.0 could take on a life of its own without submission deadlines and presentation slots. There'll be much pot luck, but bring what you can. Evaluators are still very hungry for design purpose, and when they are better fed, they'll be much better company. Even usability gurus may discover their softer side!

Acknowledgement

This work is supported by a Fellowship from NESTA (www.nesta.org.uk/about/directory/index.aspx?a=C&p=3)

References

- [1] Alderfer, C. *Existence, Relatedness, and Growth*. Free Press, 1972
- [2] Aschmoneit, P. and Heitmann, M. "Consumers cognition towards communities: Customer-centred community design using the means-end chain perspective" *in Proc. 36th Hawaii Int. Conference on System Sciences*, IEEE, 2003, 216
- [3] Bødker, S. "When Second Wave HCI meets Third Wave Challenges," *Proc NordiCHI 2006*, ACM, 1-8.
- [4] Chellapilla, K., Larson, K., Simard, P., and Czerwinski, M. "Designing human friendly human interaction proofs" *in Proc. CHI '05*. ACM, 711-772.

- [5] Cockton, G., "A Development Framework for Value-Centred Design," in *CHI 2005 Extended Abstracts*, ed. C. Gale, ACM, 1292-1295.
- [6] Cockton, G. "Designing Worth is Worth Designing," in *Proc NordiCHI 2006*, ACM, 165-174.
- [7] Cockton, G. "Putting value into e-valuation" in [12]
- [8] Dourish, P. "Implications for design" in *Proc. CHI 2006*, 541-550.
- [9] Friedman, B. and Kahn, P., "Chapter 61: Human Values, Ethics and Design", in *The Human-Computer Interaction Handbook*, eds. J. Jacko and A. Sears, 1171-1201, Lawrence Erlbaum Associates
- [10] Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., Beaudouin-Lafon, M., Conversy, S., Evans, H., Hansen, H., Roussel, N., and Eiderbäck, B.. "Technology probes: inspiring design for and with families," in *Proc. CHI 2003* ACM, 17-24
- [11] Iivari N. "Usability specialists : 'A mommy mob', 'realistic humanists' or 'staid researchers'? An analysis of usability work in the software product development," in *Proc. INTERACT 2005*, Springer LNCS 3585, 418-430
- [12] Law, E. L-C, Hvannberg, E.T. and Cockton, G. (eds) *Maturing Usability: Quality in Software, Interaction and Value*, to appear 2007, Springer.
- [13] Lieberman, H. *The Tyranny of Evaluation*, paper presented at CHI 2003 Fringe, available at <http://web.media.mit.edu/~lieber/Misc/Tyranny-Evaluation.html>, last accessed 1/2/07
- [14] Mok, C. *Designing Business*, Hayden Books, 1996.
- [15] Sayer, A., "Long Live Postdisciplinary Studies! Sociology and the curse of disciplinary parochialism/imperialism" British Sociological Association Conf, 1999
- [16] Sellen, A., Eardley, R., Izadi, S., and Harper, R.. "The whereabouts clock: early testing of a situated awareness device" in *CHI '06 Extended Abstracts*, ACM, 1307-1312
- [17] Sklar, A. and Gilmore, D. "Are you positive?" *interactions* 11, 3 (May. 2004), 28-33
- [18] Vanderbeeken, M, *Experientia interviews Richard Eisermann*, (12th Jan 2006) /www.experientia.com/interviews/eisermann, last accessed 1/2/07.
- [19] Volda, A. and Mynatt, E. D. "Six themes of the communicative appropriation of photographic images," in *Proc CHI 2005*, ACM, New York, 171-180
- [20] Whiteside, J., Bennett, J., and Holtzblatt, K., "Usability engineering: Our experience and evolution," in *Handbook of HCI*, 1st Edition, ed. M. Helander., North-Holland, 791-817, 1988
- [21] Wong, B. and Jeffery, R. *A Quantitative Study on the Role of Cognitive Structures in Software Quality Evaluation*, Tech. Report 02/3, National ICT Australia ESE, 2002, www.caesar.unsw.edu.au/publications/pdf/Tech02_3.pdf, last accessed 07/01/07
- [22] Zhai, S. *Evaluation is the worst form of HCI research except all those other forms that have been tried*, /www.almaden.ibm.com/u/zhai/papers/EvaluationDemocracy.htm, accessed 1/2/07