Comparative Appraisal of Expressive Artifacts

Melanie Feinberg
School of Information
The University of Texas at Austin
1616 Guadalupe St., Suite 5.202
Austin, TX 78701-1213 USA
feinberg@ischool.utexas.edu

Abstract
This paper describes a form of comparative, structured appraisal of expressive artifacts that adds to the existing repertoire of HCI assessment techniques. Comparative appraisal uses a situationally defined procedure to be followed by multiple assessors in examining a group of artifacts. The conceptual basis for this method is drawn from writing assessment.

Author Keywords
Evaluation; criticism; assessment

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

CHI'13, April 27 – May 2, 2013, Paris, France.

Introduction
Research that explores the interactive artifact as a cultural form is gaining traction in HCI. Such research highlights the qualities of interaction as generator of aesthetic experience, and the software artifact as a means of shaping that experience [1, 16]. HCI research has addressed the design of artifacts that produce different forms of expression and the development of environments through which these artifacts can be created by others [21, 3, 12]. For example, Gaver et al's Prayer Companion was designed to enrich the spiritual activities of cloistered nuns by unobtrusively displaying brief messages of various sorts via a custom device [10]. The Prayer Companion was not designed to solve a problem but to contribute an interpretively flexible extension of the nuns' prayer experience.

The means and measures appropriate for evaluating task-focused HCI research are less suited to the consideration of artifacts designed for experiential, rhetorical, and other expressive aims [12, 1]. One alternative to traditional performance-based evaluation has focused on incorporating reflective elements into the design process itself, as a resource for the evolution of ideas and prototypes. These reflective design approaches interrogate in-progress designs through the exploitation of expert judgment that resides within a skilled design community, sometimes in conjunction with the reflections of potential users [21].

Another mode of assessment has looked to the humanities. Humanistic criticism produces illuminating interpretations of creative works by employing intricate theoretical frameworks in conjunction with close readings of selected examples [1]. In HCI, research inspired by humanistic criticism has introduced particular theoretical orientations to the field, such as feminism and critical theory, and has proposed how such frameworks can be used to interpret existing artifacts or potentially generate new ones [2]. Complementary work has developed critical vocabularies specific to the HCI context [12, 15].

Design reflections and humanistic criticism typically focus on unique qualities of the examples they analyze, producing interpretations that reveal previously unarticulated properties. Sometimes, though, we may want to compare expressive artifacts in a more systematic way, along similar dimensions. While such sorting is a feature of experimental evaluation, the characteristics of interest for expressive artifacts may be complex ideas, such as suppleness and defamiliarization, that do not lend themselves to measurement [12, 3]. Moreover, the identification of these concepts requires interpretive judgment, which in turns relies on knowledge and skill in the assessor. This judgment goes beyond user preferences that emerge via crowdsourced ratings [as in, for example, 13].

This paper describes a form of comparative, structured appraisal that supplements existing assessment techniques. This approach resembles experimental evaluation in using a set of procedures to be followed by multiple assessors in examining a group of artifacts. However, this approach is also grounded in the recognition that expressive qualities are not conventionally measurable and that absolute agreement between assessors is neither possible nor desirable. The conceptual basis for this comparative method is adapted from writing assessment. Instructors and researchers in composition and rhetoric have long struggled with providing fair, accurate assessments of student writing proficiency while acknowledging that characteristics of good writing are difficult to define precisely, impossible to measure quantitatively, and reliant upon an indeterminate range of contextual factors [4, 18].

In the next section, I summarize the motivating scenario that provoked my investigation into comparative appraisal and present several additional use cases. I then describe how writing assessment provides a conceptual grounding for developing a project-specific comparative appraisal procedure. Finally, I summarize the comparative appraisal developed for the motivating scenario.

This paper's contribution lies in presenting the goals, justification, and utility of this form of assessment as demonstrated through a particular case study, and not in the particular approach used in the motivating scenario. The argument presented here is meant to inform the fashioning of project-specific appraisal methods that are tailored to their contexts.

Motivating scenario and additional use cases

The need for comparative appraisal arose in an interdisciplinary project that sought to translate the insight of humanistic criticism to the realm of design. An initial study used a humanities-based approach to explore what makes personal digital collections (shared sets of resources such as Pinterest boards and YouTube

playlists) interesting as forms of creative expression [8]. This initial work proposed a set of three expressive qualities that personal collections might exhibit: an eclectic purpose, an authorial voice, and emotional intimacy. A subsequent study involved a lab experiment to see whether exposure to collections that embodied all three of these qualities would affect the process or product of collection design [9]. Working within an easy-to-use digital video library environment, participants created personal collections using a library of source material focused around a particular theme. After creating initial collections, participants interacted with example expressive collections, created by the researchers to enact all three qualities under investigation. Participants compared their designs to the examples. Then participants created a second collection using another source library.

This experimental protocol required a way to systematically compare participant collections to the expressive examples and to each other, to determine if interacting with the examples had an effect on subsequent designs.

Additional use scenarios

As another potential use case, consider design research that attempts to identify and exploit particular interactive qualities, such as Isbister and Hook's work on suppleness [12]. Isbister and Hook describe the quality of suppleness and characterize how several design prototypes enact this quality. As a complement to the critical analyses that they employ, it could be productive to see how several assessors describe, across the range of prototypes, the manifestation of factors that Isbister and Hook suggest as contributing to suppleness, which include subtle social signals, emergent dynamics, and moment-to-moment experience. Both the agreement and disagreement of assessors regarding the strength of these factors and each factor's contribution toward ultimate suppleness would provide useful input for further investigation into the nature of suppleness and factors that produce it.

A complementary use case involves systematic assessment of design alternatives designed to enact certain qualities. For example, Petrelli, et al proposed a set of four ideas to exhibit "playfulness and engagement across generations" in the context of Christmas celebrations [20]. While these designs were discussed by focus groups, a comparative appraisal could additionally interrogate the degree to which the selected qualities appear in each prototype, as well as identifying factors that contribute to the generation of each quality. The information generated from such a comparative appraisal could help select prototypes for continued development and also help researchers refine their ideas of the interactive qualities being enacted.

A third use scenario involves the design of authoring environments for users to develop expressive artifacts. For example, Likarish and Winet describe a collaboratively authored Twitter novel created as part of a public art project [14]. The completed novel exhibited a polyphonic voice, which made it seem incoherent, and lacked interaction between the characters. Likarish and Winet propose writing tools to facilitate increased consistency of voice and increased character interaction in collaborative fiction. A comparative appraisal to assess the products created with such tools would help to characterize the tools' effects.

Lessons from writing assessment

To develop an appraisal procedure for the motivating scenario, I turned to writing assessment, which wrestles with similar situations: how to assign composition students to an appropriate course level, for example, when students may approach the writing of sample essays using different but equally acceptable strategies, and where the notion of acceptability itself may be difficult to define.

To be clear, writing assessment is not criticism. Criticism is a form of research, and it relies upon skilled interpretive expertise in conjunction with a grounding in appropriate literature. Such criticism has traditionally been intended to produce new scholarly knowledge, not to provide a basis for discriminating between potential designs as part of an ongoing project. The science-based constructs of reliability and validity are meaningless in the critical context: the goal of criticism is to illuminate new conceptual space, not to prove or disprove hypotheses.

In contrast, writing assessment is a pragmatic activity focused on making decisions; it is not itself research. An assessor in a university's writing program, for example, might determine whether a student's portfolio should pass or fail the university writing requirement. Assessors are trained to identify criteria employed in a particular assessment; they need not be scholars.

Writing assessments must be consistent enough across multiple raters to ensure confidence in decisions such as passing or failing. Accordingly, reliability and validity have been employed in this domain. However, their meaning and the nature of their relevance has been debated for this context. Indeed, the literature of writing assessment has been characterized as a progressive conflict between reliability and validity [4, 6, 18]. These debates inform my own approach to comparative appraisal.

While indirect quantitative testing methods, such as multiple-choice examinations of grammar mechanics, might be statistically reliable, writing teachers have long contended that such methods do not achieve face validity as a determination of writing ability; a student can master the rules of grammar and yet not be able to write proficiently or persuasively [6]. However, experts judge writing samples differently, as famously demonstrated in a study conducted by the Educational Testing Service (ETS) in 1961 [7]. 300 writing samples written by college students were sent to 53 experts in a variety of fields, who rated the samples and commented upon strengths and weaknesses. Agreement was dismal: 94 percent of the essays received at least seven different grades out of nine possibilities. From this set of varied assessments, the ETS researchers analyzed rater comments to derive five broad areas that captured most criteria variously employed by the raters: Ideas, form, flavor (style), mechanics, and wording [7]. To decrease variability of the sort described in the ETS report, writing assessment researchers developed holistic scoring methods based on standardized rubrics that formalize a small set of generalized criteria such as those isolated in the ETS study, supported by rater training sessions in which applying the rubric consistently is emphasized [11]. In the U.S., such methods have been widely adopted for both national (such as Advanced Placement exams) and institutional testing purposes [4, 6, 18].

However, while the formalization of assessment criteria via standardized rubrics increased rating consistency, concerns about test validity continued. Moss notes that reliability decreases when assessors examine portfolios of student work, instead of single test essays, because portfolio samples are created under different circumstances, unlike test essays that respond to a single prompt [17]. And yet surely, Moss contends, the evidence provided through the "complex, authentic tasks" represented in a portfolio is more indicative of writing ability than a context-stripped essay from a test. In their courses, instructors teach how to compose appropriate written material for different contexts, because they believe, as a core value, that good writing responds to a situation. Yet assessment protocols have devalued this skilled expertise in favor of techniques that can be implemented widely and consistently. Accordingly, Moss asserts that focusing on reliability in writing assessment can impede validity, suggesting that disagreement between raters might be an opportunity for productive dialogue regarding assessment criteria and implementation—in other words, a means through which the ultimate validity of the assessment instrument can be solidified [17].

Various researchers have extended this argument, emphasizing the pedagogical poverty of context-independent assessment criteria and the need to judge writing according to local values (e.g., according to the instructional philosophy of a particular composition department). The assessment process, including development and implementation of localized procedures, becomes a means of defining, debating, and articulating those values for a particular instructional community [11, 4]. While some proposals for localized assessment argue for getting rid of rubrics entirely, contending that they thwart the recognition of imaginative solutions to writing problems, others retain the structure of rubrics in a more flexible, context-specific manner [5]. But the point of the rubric becomes less to assign points or grades consistently and more to structure a principled conversation about the work, either between multiple assessors or between assessors and students.

Parkes proposes that reliability of such localized assessment procedures be formulated as a type of argument [19]. For each assessment situation, the most applicable values associated with reliability (dependability, accuracy, and so on) are selected as appropriate for the assessment purpose, along with a proposed level of reliability for the situation (accuracy may need to be high if an assessment is a graduation requirement, but lower if the assessment is used for class placement). The assessment designer musters evidence to demonstrate that the procedure adheres to the defined reliability construct [19].

In developing a comparative appraisal procedure, then, the literature of writing assessment suggests that:

- The criteria being assessed should be grounded in project-specific goals and values.

- A systematic procedure and set of assessment criteria can direct assessors' attention consistently on the artifacts being examined; however, the aim should center on consistent focus, rather than consistent ratings (that is, disagreement can be as informative as agreement).

- Reliability and validity must be confronted; their meaning cannot be assumed, but neither can their potential relevance be dismissed. Instead, the designer of a comparative appraisal formulates an argument that defines validity and reliability appropriately for the situation and that provides evidence to support the proposed definitions.

Comparative appraisal procedure for motivating scenario

This section presents an extended example of a comparative appraisal procedure that responds to the motivating scenario. I begin by describing the values addressed through the appraisal and its ultimate goals. I then summarize procedure components and provide an argument to demonstrate the procedure's reliability and validity for the situation of its use. Again, while the appraisal procedure itself is deeply enmeshed in the motivating scenario and cannot be merely exported to other research contexts, its goals, justification, and subsequent implementation can serve to inform the development of similar protocols.

Appraisal goals and values

In the localized approach to writing assessment, evaluative criteria are generated based on the values of the immediate instructional community as to what constitutes good writing. For the motivating scenario, criteria were generated based on the proposed values examined in the research project: the three expressive qualities defined in [8] and the overall expressiveness potentially enabled through the synthesis of those characteristics. While this may seem like an obvious decision, the larger point is that *any* comparative appraisal relies for its conceptual basis upon project-specific criteria. The assessor's evidence for determining the relative presence of the appraisal criteria is based in the mechanisms of expression appropriate to the specific artifact at hand. For the motivating scenario, mechanisms include the selection, description, and arrangement of items in a personal digital collection.

Localized modes of writing assessment also emphasize the rubric as a form of procedural infrastructure to systematically focus an assessor's attention in particular ways, and accordingly downplay the rubric as a means of generating reliably consistent scores between assessors. Similarly, in the comparative appraisal procedure developed for the motivating scenario, the presence of a certain number or type of these mechanisms does not mandate a particular judgment. The goal is to consistently direct the attention of each assessor on the work under review in similar way, and not to create a formalized scale that ensures consistent ratings across multiple assessors. The ultimate assessments are open to the possibility of principled interpretive differences and yet are still comparable across defined dimensions. Additionally, the artifacts being appraised are not being "graded" or described as holistically good or bad. The appraisal only compares perceived differences in the strength in which the particular characteristics of interest appear.

Procedure components

For the motivating scenario, the artifact being assessed is the personal digital collection. For each collection, assessors perform the same tasks for of each of the three expressive qualities identified in [8]: an original purpose, authorial voice, and emotional intimacy. These tasks involve describing how the quality is exhibited

through the collection, rating the strength of the quality, and describing how each mechanism through which expression is generated—selection, description, and arrangement of resources—contributes to the manifestation of the quality. A worksheet documents each task and provides for standardized recording.

Assessors performed the following tasks for each expressive quality:

1. Describe, in free text, the way that the quality is enacted through the collection.
2. Provide a rating on a scale of 1 to 10 to express the strength of that quality in the collection.
3. According to a brief coding scheme (less than ten categories) developed through preliminary review of the collections to be assessed, record all the instances in which selection of resources contributed to the manifestation of the quality.
4. Using another brief coding scheme, record all instances in which the description of resources through labels or annotations contributed to the manifestation of the quality.
5. Describe, in free text, contributions that the arrangement of resources (such as the order of items) makes to the manifestation of the quality. (This mechanism does not employ coding categories because there was less regularity in its employment across collections.)
6. Describe, in free text, any contribution resulting from the integration of three expressive mechanisms—selection, description, and arrangement—to the manifestation of the quality.

After assessing the manifestation of each expressive quality according to these defined tasks, the assessor provides an overall expressiveness rating on a scale of 1 to 10 along with a brief explanation of that rating. (Overall expressiveness is not a simple average of the three expressive qualities.)

Prior to beginning the appraisal, assessors discussed a draft worksheet to promote shared understanding of appraisal elements: expressive qualities under examination, mechanisms that work to produce the qualities, codes for various forms of selection and description, and so on. After revision of the worksheet, each assessor conducted several preliminary appraisals, which were then discussed to resolve discrepancies in how assessors understood appraisal elements (and not to force agreement on specific explanations or ratings). As the appraisal continued, regular discussions were held, and individual assessors prepared for these by writing memos in which they explored their rationale for rating collections differently. After completing preliminary appraisals of all collections, assessors internally harmonized their ratings, making adjustments as necessary to ensure that their own idea of what constituted a 3 or an 8 was consistent over the set of items to assess, even as their evidence for each rating might differ for each collection.

Reliability and validity argument for comparative appraisal procedure

In the literature of writing assessment, researchers identified problems with validity when students were asked to produce assessment materials that were not congruent with what instructors valued as good writing [11, 17]. For example, writing instructors might believe that good writing requires revision, and yet students would write under timed test conditions for assessment.

The comparative appraisal procedure as created for the motivating scenario avoids these problems and achieves construct validity. First, the study participants produced precisely the same materials, personal digital collections, as those examined in the first study, [8], that identified the expressive qualities. Additionally, the example and participant collections were produced in the same manner, using the same materials. Second, the characteristics of interest are directly examined in the appraisal procedure, not via indirect substitutes. The procedure looks at each expressive quality separately and provides three complementary means of registering that characteristic's presence in the collection being assessed: through a holistic numerical rating, through a holistic text explanation, and as specifically manifested through each of the three expressive mechanisms appropriate for collections: selection, description, and arrangement. Identification of selection and description contributions to each quality is systematized with defined coding categories. This three-stage process enabled us to see if a quality's manifestation is due to some previously unidentified mechanism in addition to selection, description, and arrangement: if the quality's strength is given a high rating, and yet neither selection, description, nor arrangement contributes to the presentation of that characteristic, then we have learned that the theoretical construct underlying the study is insufficient to explain the observed phenomena. Similarly, the overall expressiveness rating and explanation are separate from the assessments of the three identified expressive qualities. If collections are consistently rated more highly or poorly than the ratings for their particular qualities, then we may be able to identify additional contributors to expressiveness, or to determine that some of the previously identified qualities are more or less important than others.

In terms of a reliability argument as articulated by Parkes, the purpose of the comparative appraisal procedure is to sort the collections, both participant and example, into ranges that represent different levels of expressiveness [19]. The appraisal procedure is not intended to *explain* the differences between ranges (that is, how a collection in the 8-10 range is different from a collection in the 1-3 range) or to illuminate the unique qualities of each collection in the manner of criticism, although the appraisal procedure does provide a means for identifying complementary close readings that might produce such explanations. Accordingly, the primary value enacted through this comparative appraisal procedure is consistency within the assessments contributed by a particular assessor. It is important that each rater be confident that, say, all of the 2s for overall expressiveness and for each individual characteristic are equivalent, although each 2 might be placed in that category for different reasons, and that the relative distance between a 2 and a 6, for example, is clear in the assessor's mind. Accordingly, a secondary value is coherence of explanation in each assessor's rationale for making appraisal decisions. Another secondary value is consensus between assessors on the meanings of the constituent concepts of the appraisal and on the goals of the procedure itself. The overall tolerance required for any particular appraisal is relatively low across assessors, because we are interested only in sorting into ranges, and this sorting is not designed to be an explanation of anything in itself, but only the means through which both trends and discrepancies can be characterized and explained via other means (such as close readings).

Utility of the appraisal findings does not depend on agreement across assessors for particular judgments. While relative agreement regarding placement into ranges may provide useful information, discrepancies across assessors also provide useful information. As evidence for reliability, several elements contribute to the primary value of consistency within raters. For each appraisal, an assessor provides multiple forms of judgment: numerical ratings, explanations of these ratings, and systematic identification of elements that contribute to the production of each quality (either by codes or free text). These multiple forms of judgment constitute internal checks on the assessor, ensuring a well-developed rationale for each appraisal. Moreover, the final harmonization process ensures that shifts in how judgments are applied over the length of the appraisal procedure are identified and adjustments made. The value of coherence is achieved through the writing of text explanations to supplement other forms of judgment, and through the discussions conducted throughout the process. While these discussions are not meant to persuade any assessor to change a reasoned opinion, they do require assessors to express their rationale cogently in language that others can understand, which can sometimes reveal flaws in one's initial interpretive logic. The value of consensus is also produced through discussions, in particular the initial norming sessions where the appraisal worksheet is debated, and where preliminary assessments are shared and questioned to increase mutual understanding of the constituent concepts and goals.

In sum, this section demonstrates how the comparative appraisal procedure developed for the motivating scenario achieves validity and presents a case through which a limited form of reliability is claimed to be necessary. In providing an example of such an argument, this section shows the process through which similar arguments might be made for any such appraisal, as developed for different project situations.

Conclusion

The comparative appraisal method developed for the motivating scenario was used in both [9] and in a subsequent experiment. It has proved successful as a key component of our data analysis, as it facilitates systematic comparison of the expressive artifacts created in our study while remaining sensitive to the complex, subtle nature of the qualities being investigated. In [9], disagreement between the two assessors for each artifact was minimal, as far as the project goal of sorting into ranges was concerned, and demonstrated a large difference between participant and example collections, before and after the experimental intervention. With confidence in this assessment, we were then able to focus on isolating, via the close reading of both individual collections and participant interview comments, reasons for these differences. In the follow-up experiment, which used a physical environment instead of a digital one, general agreement between the three assessors for each artifact was disrupted by mild disagreement for certain collections. Discussions of these disagreements were particularly insightful, as they revealed new ideas about how the qualities being investigated might interact. (In short, some assessors expected all qualities to smoothly integrate, but other assessors proposed that conflicts between qualities could productively affect overall expressiveness.)

Our experiences illustrate both the definite utility and associated limitation of comparative appraisal: its

findings provide a solid basis for comparison, but they do not in themselves explain observed differences. However, both the systematic procedure, as well as the assessments themselves, can suggest a path toward constructing such explanations.

References

1. J. Bardzell. Interaction criticism: an introduction to the practice. *Interacting with Computers* 23, 604–621, 2011.
2. S. Bardzell. Feminist HCI. *Proceedings of ACM CHI 2010,* 1301–1310, 2010.
3. G. Bell, et al. Making by making strange: defamiliarization and the design of domestic technologies. *ACM Transactions on Computer-Human Interaction* 12(2): 149–173, 2005.
4. B. Broad. *What we really value: beyond rubrics in teaching and writing assessment.* Logan, UT: Utah State University Press, 2003.
5. B. Broad. Introduction. *Organic writing assessment: dynamic criteria mapping in action.* Logan, UT: Utah State University Press, 2009.
6. D. Charney. The validity of using holistic scoring to evaluate writing: a critical overview. *Research in the Teaching of English* 18(1): 65–81, 1984.
7. B. Diederich, et al. Factors in judgments of writing ability. Princeton, NJ: ETS report number RB-61-16, 1961.
8. M. Feinberg. Personal expressive bibliography in the public space of cultural heritage institutions. *Library Trends* 59(4): 588–606, 2011.
9. M. Feinberg, et al. Understanding personal digital collections. *Proceedings of ACM DIS 2012,* 200–209, 2012.
10. B. Gaver, et al. The Prayer Companion. *Proceedings of ACM CHI 2010,* 2055–2064, 2010.
11. B. Huot. Toward a new theory of writing assessment. *College Composition and Communication* 47(4): 549–566, 1996.
12. K. Isbister and K. Hook. On being supple. *Proceedings of ACM CHI 2007*, 2233–2242, 2009.
13. B. Lee, et al. Designing with interactive example galleries. *Proceedings of ACM CHI 2010,* 2257–2266, 2010.
14. P. Likarish and J. Winet. Exquisite corpse 2.0: qualitative analysis of a community-based fiction project. *Proceedings of ACM DIS 2012,* 564-567, 2012.
15. J. Lowgren and E. Stolterman. *Thoughtful interaction design.* Cambridge, MA: MIT Press, 2004.
16. J. McCarthy and P. Wright. *Technology as experience.* Cambridge, MA: MIT Press, 2004.
17. P. Moss. Can there be validity without reliability? *Educational Researcher* 23(2): 5–12, 1994.
18. P. O'Neill, et al. *Guide to college writing assessment.* Logan, UT: Utah State University Press, 2009.
19. J. Parkes. Reliability as argument. *Educational Measurement: Issues and Practice,* Winter 2007.
20. D. Petrelli, et al. Digital Christmas: an exploration of festive technology. *Proceedings of ACM DIS 2012,* 348-357, 2012.
21. P. Sengers and B. Gaver. Staying open to interpretation. *Proceedings of ACM DIS 2006,* 99–108, 2006.