Comparative Appraisal: Systematic Assessment of Expressive Qualities

Melanie Feinberg
School of Information
The University of Texas at Austin
Austin, TX USA
feinberg@ischool.utexas.edu

ABSTRACT
Clifford Lynch describes the value of digital libraries as adding interpretive layers to collections of cultural heritage materials. However, standard forms of evaluation, which focus on the degree to which a system solves problems, are insufficient assessments of the expressive qualities that distinguish such interpretive content. This paper describes a form of comparative, structured appraisal that supplements the existing repertoire of assessment techniques. Comparative appraisal uses a situationally defined set of procedures to be followed by multiple assessors in examining a group of artifacts. While this approach aims for a goal of systematic comparison based on selected dimensions, it is 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 drawn from the literature of writing assessment.

Categories and Subject Descriptors
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous

General Terms
Design

Keywords
Evaluation; criticism; assessment

1. INTRODUCTION
In 2002, Clifford Lynch suggested that the "aggregation of materials in a digital library can be greater than the sum of its parts" [21]. For Lynch, a digital library does not just make "raw material" (which Lynch calls "collections") available; a digital library adds "layers of interpretation" that weave together items perhaps held by multiple institutions, enacting a coherent, purposeful perspective upon the assembled contents. Lynch contends that this synthetic, expressive activity might form the substrate onto which intellectual communities would coalesce around digital libraries.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.

In 2011, as social media tools began to proliferate, Marty and Kazmer considered the efforts of cultural heritage institutions to focus social media toward the co-construction of knowledge with their user communities [22]. While Lynch was referring to the interpretive digital library as catalyst for a scholarly community, and Marty and Kazmer were imagining a wider public audience, the broader vision is similar: the deployment of collection materials as expressive elements to structure an array of interpretive possibilities. Tools like Storyspace, which facilitates the development of narratives that join museum objects in interpretively illuminating ways, are emerging to support the realization of such ideas [30].

In commenting about potential difficulties regarding the sustainability of such interpretive layers for digital libraries, Lynch pinpoints a complex problem for these initiatives: how to determine when an interpretation is interesting, when it requires updating, and when it has perhaps become outmoded or no longer viable. The question of what makes an interpretive element interesting, or provocative, or thoughtful, among myriad potential expressive qualities, and relatedly what makes one such element more or less interesting (or provocative, or thoughtful, or any such quality) than another, is not easily answered. While standard evaluative techniques such as user surveys and usability tests are undeniably helpful in determining how and to what extent digital libraries address user information needs, they are less appropriate when considering questions like "How does each set of metadata elements explore the subtleties of cultural interplay in these art objects?" or "How strongly does the unique perspective of the author appear in these user-contributed personal collections?" While user opinions can certainly be generated to comment upon such areas, it is difficult to use these opinions for systematic comparison along defined dimensions, because user opinions may all differ in how they define a unique authorial voice, or any other expressive quality. Additionally, qualities such as cultural interplay and authorial voice are not measurable in conventional ways, unlike performance metrics and similarly quantifiable characteristics, such as precision and recall or task completion times.

In this paper, I describe a form of comparative, structured appraisal, focused on expressive qualities, that supplements the existing repertoire of assessment techniques. The approach that I describe resembles experimental evaluation in using a standard 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. Accordingly, the mode of appraisal that I propose is situationally defined for particular contexts. This paper provides an example of such an appraisal as developed for a particular project.

The conceptual basis for this comparative method comes from the literature of writing assessment. Instructors and researchers in the

field of composition and rhetoric have long struggled with the task of providing fair, accurate assessments of student writing proficiency while acknowledging that the characteristics of good writing are difficult to define precisely, impossible to measure quantitatively, and reliant upon an indeterminate range of contextual factors [5, 26]. Additionally, while writing assessments follow similar principles in their design, they must flexibly account for different pedagogical goals, instructional situations, and particular values in terms of what constitutes writing skill.

In the next section, I summarize the motivating scenario that provoked my investigation into comparative appraisal and present several additional use cases. Next, I situate comparative appraisal in the context of other forms of assessment for both information systems and human-computer interaction. Following this extended project rationale, I 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 and its implementation in two related studies.

This paper's contribution lies in presenting the goals, justification, and utility of this form of assessment as demonstrated through a 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.

2. PROJECT RATIONALE: 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 digital library 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 [9]. This foundational work proposed a set of three more specific expressive qualities that personal collections might exhibit: an original 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 [10]. Working within an easy-to-use digital video library environment, the Open Video Digital Library Toolkit, participants created personal collections using a library of source material focused around a particular theme [14]. 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 protocol required a way to systematically compare participant collections to the examples and to each other, to determine if interacting with the examples had an effect on subsequent designs (see Figure 1). To accomplish this goal, we had to determine how, and to what extent, a collection demonstrated the expressive qualities of original purpose, authorial voice, and emotional intimacy. We needed to establish whether participant collections showed, for example, a stronger voice after exposure to the expressive examples, or whether this exposure had no clear effect.

2.1 Additional use scenarios

Additional use cases appear in design-focused research that attempts to identify and exploit particular expressive qualities. For example, in the City of Lit digital library described by Hsieh and colleagues, undergraduate literature students contribute to a collection focused on Iowa City authors and locations. Initial evaluations of the system focused on students' perceptions of usability and on the system's utility for their own learning [15]. A comparative appraisal of students' contributions, perhaps focused around qualities such as comprehensiveness and local color, would provide another stream of information to support continued system development.

Another scenario involves systematic assessment of design alternatives designed to enact certain qualities. In an example from HCI design research, Petrelli and colleagues proposed a set of four ideas to exhibit "playfulness and engagement across generations" in the context of Christmas celebrations [28]. 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 collections or other artifacts that highlight expressive qualities. Storyspace, created by Wolff and colleagues, enables museum curators to assemble museum database records that combine to tell a coherent story [30].

Storyspace uses a sophisticated ontology that can propose alternate narrative structures for a particular object group. A structured comparison of such alternatives generated for particular collections would provide another means of assessment for the Storyspace ontology, in addition to the planned elicitation of feedback from curators using the prototype system. As another example, Likarish and Winet describe a collaboratively authored Twitter novel created as part of a public art project [19]. The completed novel exhibited a polyphonic voice, which made it seem incoherent, and the novel also 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.

3. PROJECT RATIONALE: LIMITATIONS OF CURRENT ASSESSMENT TECHNIQUES

In this section, I describe how existing approaches to assessment do not adequately address the needs of the motivating scenario and accompanying use cases. First, I consider traditional means of evaluation focused around solving user problems, and then I examine alternative means of assessment developed by design-oriented human-computer interaction research.

3.1 Traditional evaluation: measuring effectiveness and efficiency in solving identified problems

Surveying the state of the art in digital library evaluation, Fuhr and colleagues define a three-part model that incorporates

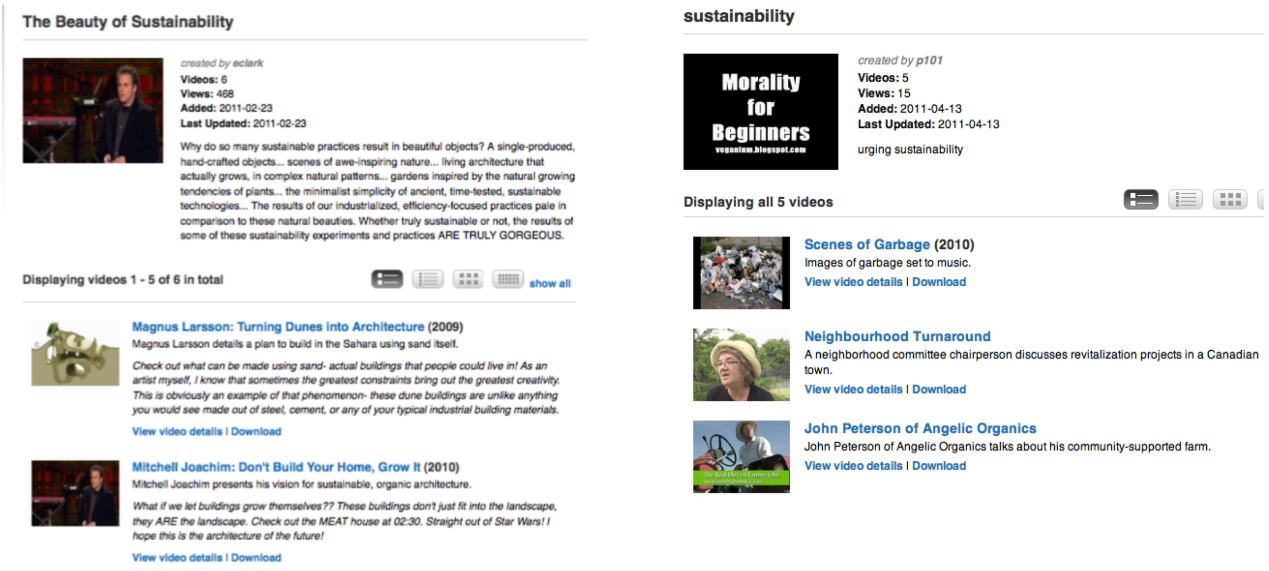

Figure 1: In the motivating scenario, we needed to compare qualities such as original purpose between personal digital collections like these. The left-hand collection snippet is an example created by researchers; the collection snippet on the right was created by a study participant. (The title and text paragraph that precede the video list are collection-specific annotations. In the video list, the plain text comes from the system and is not created by the collection author; the italic text is a collection-specific annotation.)

usability, usefulness, and performance [11]. Performance evaluation focuses on measurement of system properties, such as precision and recall for resource retrieval, usage rates for individual resources, quality of multimedia playback, and so on. Usability involves the ease, speed, and satisfaction with which users can perform tasks, and usefulness describes the extent to which the library's content is appropriate for users' information needs.

The evaluation model developed by Fuhr and colleagues, which relates system, content, and user elements, adopts a standard problem-solving orientation, in which a digital library's purpose is oriented around facilitating an identified goal, typically that of satisfying information needs. Although the composition of such models may vary, they tend to maintain this problem-solving focus. Khoo and MacDonald's evaluation model, for example, incorporates broader organizational components to consider the goals of the sponsoring institution and how those goals are facilitated [17]. While this model includes additional elements, its object remains the degree to which a digital library solves identified problems, or produces established outcomes (for example, does item metadata make it easier and faster for users to find what they are looking for?).

A task-based focus for evaluation is of course valuable and necessary. The fulfillment of identified user needs through digital library services, and the accomplishment of other institutional goals (such as increasing a user base) is undeniably crucial, and such models support this form of assessment. However, interpretive elements and the expressive qualities associated with them don't fit neatly into models oriented around problems and solutions, or information needs and the satisfaction of those needs. A metadata element set that enacts relationships to reveal cultural interplay between museum objects achieves its interest partly from illuminating the materials in unanticipated ways, in addition to supporting existing, recognized information needs. Similarly, a personal digital collection of diverting novels for long airplane trips that gains its character from a unique, engaging authorial voice might solve no identified problems but still represent a compelling interpretive layer that enriches a digital library experience. Expressive qualities such as cultural interplay and authorial voice speak to the transformative potential of digital library design as an intellectual or aesthetic experience, rather than support for existing tasks. It makes sense to ask users how a system supports the resolution of their information needs, because users are experts in their own needs. It is less clear that user surveys or interviews could inform confident judgments of expressive qualities, without instructing users about what those qualities are and how they might manifest in the digital library. (This doesn't mean than the effects of these qualities wouldn't be perceived by users, just that users might not have the vocabulary or expertise to productively articulate such effects for purposes of comparison and appraisal.)

3.2 Alternate modes of assessment from HCI research

Design-oriented HCI research has begun to explore the interactive artifact as a cultural form. Such research highlights the qualities of interaction as generator of aesthetic experience, and the software artifact as a means of shaping that experience [1, 23]. For example, the Prayer Companion was created to enrich the spiritual activities of cloistered nuns by unobtrusively displaying brief informative messages, including notifications of current events, via a small custom device [12]. The Prayer Companion was not designed to solve a problem but to contribute an interpretively flexible extension of the nuns' prayer experience.

In HCI, various alternatives to experimental evaluation have been proposed to support principled assessment of such expressive interactive artifacts. One alternative has focused on incorporating reflective elements into the design process itself, as a resource for the evolution of ideas and prototypes. These reflective 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 [13, 29].

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, 2]. 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 help interpret existing artifacts and generate new ones [3, 4]. Complementary work has developed critical vocabularies specific to the HCI context [20].

Design reflections and humanistic criticism typically focus on unique qualities of the examples they analyze, producing interpretations that reveal previously unarticulated properties. For example, a critical reading of the maeve database pinpoints the integration of content and navigation as a distinctive element of the interactive experience [2]. In our case, however, the needs of the motivating scenario required the comparison of expressive qualities in a more systematic way, along consistent dimensions. And yet the identification of expressive qualities does involve interpretive judgment, which in turn relies on a certain level of knowledge and skill in the assessor. This judgment goes beyond user preferences that emerge via crowdsourced ratings [as in, for example, 18]. In short, we needed to develop a new means of assessment, in which informed judgment is deployed in a controlled, systematic way across multiple exemplars.

4. LESSONS FROM WRITING ASSESSMENT

To develop such 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 inquiry in itself, 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 the university writing requirement. While assessors are trained to identify criteria employed in a particular assessment, and while they typically have knowledge and expertise in writing, 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 [5, 7, 26]. 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 [7]. However, experts judge writing samples differently, as famously demonstrated in a study conducted by the Educational Testing Service (ETS) in 1961 [8]. 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 [8]. 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 [16]. In the U.S., such methods have been widely adopted for both national (such as Advanced Placement exams) and institutional testing purposes [5, 7, 26].

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 [24]. 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 [24].

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). Accordingly, the assessment process, including development and implementation of localized procedures, becomes a means to articulate those values for a particular instructional community [16, 5]. 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 [6]. But the point of the rubric becomes less to assign points or grades consistently and more to structure a principled conversation about good writing, 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 [27]. 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 [27].

To summarize, in developing a comparative appraisal procedure to respond to situations such as that represented in the motivating scenario, 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.

One might note a certain level of similarity between a writing assessment that examines certain characteristics of a work according to criteria and procedures made systematic via a rubric and the popular usability inspection method of heuristic evaluation, as initially defined by Nielsen and Molich [25]. In heuristic evaluation, a small group of usability evaluators reviews an interface for problems, as identified according to an agreed-upon list of usability principles. From a procedural perspective, heuristic evaluation does resemble the localized expression of writing assessment that I have described here: a set of trained assessors examines an artifact according to identified criteria. However, heuristic evaluation operates under a problem-solving orientation similar to most traditional software evaluation modes. In heuristic evaluation, assessors are specifically looking for "problems" as defined according to generally applicable principles (that is, that would be considered problems in all software systems). The goals and values of heuristic evaluation are universal, not local, and the review is focused on problems to be fixed, not on characterizing the extent to which criteria manifest and the (perhaps unanticipated) effects thus generated. Moreover, the assessors in heuristic evaluation are expected to agree; while multiple assessors are used to ensure that more problems are identified, the sense is that individual assessors merely miss some issues, instead of having principled agreements about what constitutes an issue. Finally, the procedure of a heuristic evaluation is not adapted based on situationally specific needs for reliability and validity. Accordingly, the comparative appraisal method developed for the motivating scenario, described in the next section, differs from heuristic evaluation in all three of these ways: its criteria are based upon locally determined goals and values, the assessors are not expected to agree, and its procedures align with situationally determined requirements for reliability and validity.

5. EXAMPLE COMPARATIVE APPRAISAL PROCEDURE FOR MOTIVATING SCENARIO

This section presents 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. While this example appraisal procedure 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, for use cases such as those described earlier in the paper.

5.1 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 [9] 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. Moreover, the assessor's evidence for determining the relative presence of 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 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 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 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.

5.2 Procedure components

For the motivating scenario, the artifact being assessed is the personal digital collection (see Figure 1 for examples). For each collection, assessors perform the same tasks for each of the three expressive qualities identified in [9]: 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.

For each collection, assessors performed the following tasks for each of the three expressive qualities (purpose, voice, and emotion):

1. Describe, in free text, the way that the quality is enacted through the collection.
2. On a scale of 1–10, rate 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 provided 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, providing for the sum to be either more or less than the parts.)

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.

5.3 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 [16, 24]. 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, [9], 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 relatively rough ranges that represent different levels of expressiveness [27]. The goal is to merely confirm a difference between, say, an 8 and a 2, and not to draw conclusions about the difference between a 4 and a 5. Additionally, 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.

6. IMPLEMENTATION OF COMPARATIVE APPRAISAL IN TWO STUDIES

The comparative appraisal method developed for the motivating scenario was used in both [10] 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.

6.1 Initial implementation of comparative appraisal: the utility of assessor agreement

After the laboratory sessions for [10] were completed, it was apparent that the participant collections used many fewer descriptive elements (titles and annotations) than the expressive examples did. Accordingly, we assigned three assessors for the example collections, where greater complexity seemed more likely to lead to potential difference of opinions, and two for each participant collection. We also decided to focus on the more complex examples in the initial norming sessions where preliminary appraisals were discussed. As described in the previous section, while it was not important to achieve agreement between assessors, it was important for the logic used by a single assessor to be consistent within and across appraisals. This was initially confusing to some assessors; it was crucial to emphasize that there was no pressure to change assessments, only to justify them.

Additionally, it was necessary to clarify that the appraisal procedure was not meant as a vehicle to perform an abbreviated mode of criticism, that is, to explain the unique properties of a collection in an original and comprehensive way. For example, while an assessor might have had a personal reaction against a brash authorial persona expressed through a particular collection, an exploration of that reaction would not align with the goals of the appraisal; the appraisal only examined the strength and manner in which an authorial voice was presented, not the effect of that characteristic on the assessor. An appropriate appraisal in this case might claim that the use of capital letters, short phrases, and evocative wording choices in annotations produced a strong authorial voice. However, a remark that the assessor found this authorial personal distasteful would not be germane to the appraisal's goals.

Ultimately, in this study, disagreement between the assessors for each artifact was minimal. The appraisal found a large difference between participant and example collections, before and after the experimental intervention. In short, participant collections demonstrated expressive qualities much less strongly than the examples, and interacting with the examples did not have a clear effect.

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. It is important to emphasize that while the comparative appraisal was a vital element in our data analysis, it did not in itself explain differences between participant and example collections. However, the findings from the appraisal enabled us to focus our subsequent efforts and construct such explanations. Briefly, our finding was that participants, despite understanding that their collections created for the study were meant for a public audience, and not for personal information management, were nonetheless building collections based on a personal information management orientation, instead of creating the collections as public expression. A significant manifestation of this personal information management approach was that participant collections used very few annotations and titles, in contrast to the example collections. While this difference in descriptive metadata was immediately apparent to us, a cogent explanation emerged only following the appraisal process and subsequent closer examination of certain collections in conjunction with interview commentary.

6.2 Second implementation of comparative appraisal: the utility of assessor disagreement

In a follow-up experiment (the findings are not yet published), we attempted to reorient participants to a mode of creative expression, instead of personal information management, by creating a test condition that explicitly separated out the mechanisms of selecting, arranging, and describing resources as distinct tasks. To keep the experiment focused on task structure, we used a physical environment instead of a digital one: participants created their collections with notecards and bulletin boards, instead of a digital library environment (see Figure 2 for an example).

In this experiment, the participant collections were much more complex (both for the test condition and the control, which did not include separate subtasks) than the previous study. Accordingly, three assessors performed appraisals on all participant collections as well as all example collections. It was especially helpful for assessors to write internal memos that made explicit rationale for

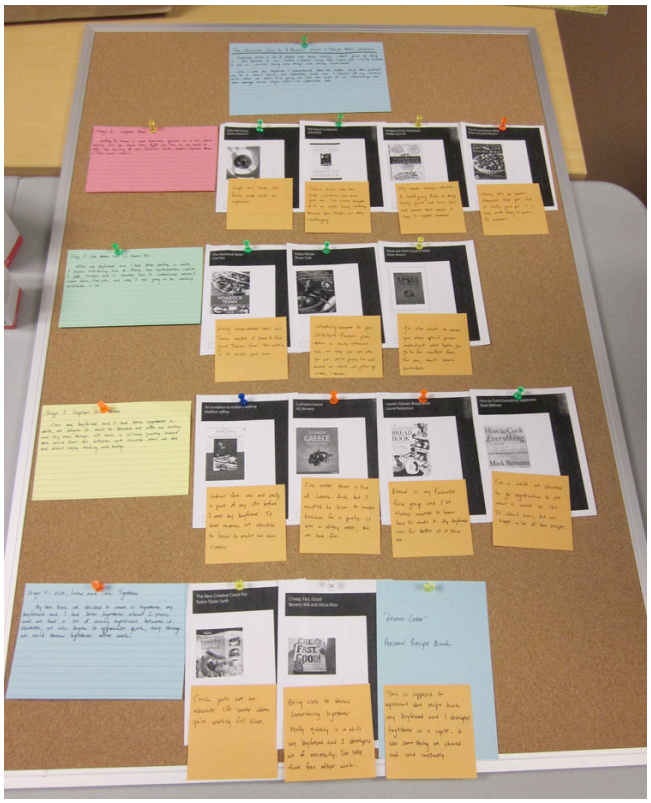

Figure 2: Participant collection from the second experiment. The white slips represent cookbooks selected for the collection; the colored notes are participant annotations.

high, low, and middle ratings. For example, one assessor wrote a memo that identified the role of cohesion, particularly regarding the characteristic of original purpose, in her judgments. This assessor realized that she consistently found collections where all the items adhered to a clearly defined purpose to be more expressive overall, and that she found digressive annotations distracting.

In this second study, findings from the appraisals showed two things. First, the participant collections as a group were much more expressive than the first study, although not quite as expressive, in terms of general ranges, as the examples. Second, there was no clear difference between the two test conditions, using a structured task to create the collections, or using an unstructured task.

As with the first experiment, the appraisal findings were necessary but not sufficient. They demonstrated differences (and lack of difference) between groups, but did not provide explanations. Despite the much greater diversity of collection expressiveness in the second study, differences in numerical ratings between assessors were modest: 1 to 2 points overall between all three assessors for the vast majority of cases. In four collections (out of 24 total), there was a discrepancy of 3 points in overall expressiveness (and none larger than 3). These discrepancies, however, were extremely productive in gaining insight from the data. In examining assessors' rationale for the divergent appraisals, it became clear that cohesion, as identified in one assessor's memo, was an important factor for all assessors but sometimes in different ways. For the first assessor, cohesion of purpose was always the most important. For the two other assessors, a fractured sort of cohesion was sometimes acceptable, as when digressive annotations for collection items contributed to the quality of authorial voice at the expense of purpose. This identification of cohesion as a primary contributor to appraisal then turned our attention to the means through which cohesion was enabled, which led us to focus on the contributions of the study's material conditions (the use of physical items) as a means of framing the collection as a creative work. The framing process became the focus of this study's findings.

7. Conclusion

Our experiences illustrate both the utility and limitation of comparative appraisal: its findings provide a solid basis for comparison, but they do not in themselves explain observed differences. However, the systematic procedure and focused attention of the appraisal, as well as assessment findings, can suggest a path toward constructing such explanations.

In summary, comparative appraisal is not a lightweight version of humanities-style criticism, nor is it a less strict means of traditional evaluation, and it is not meant to replace either. It performs a separate, useful function, just as writing assessment does: to develop criteria and associated processes to systematically sort and relate alternatives. Given the interpretive richness of the materials under consideration, any comparative appraisal must be explicitly devised to accommodate the particularities of the research or practice situation: the goals of the appraisal, the values that inform it, and the artifacts being appraised will need to be taken into account when determining what to assess, the form in which assessments should be conveyed, and the appropriate evidence to support assessment. While the localized approach advocated here is not simple, it is nonetheless tractable, and it achieves a necessary and useful balance between existing assessment modalities.

8. ACKNOWLEDGMENTS

This work was partially funded by the John P. Commons Teaching Fellowship and the Alumni Fellowship from the School of Information at the University of Texas at Austin. Many thanks to Gary Geisler, Eryn Whitworth, Ramona Broussard, Emily Clark, Eliot Scott, and Rachel Appel, my collaborators in the motivating studies. Thanks as well to Ciaran Trace for helpful comments on an earlier draft.

9. REFERENCES

[1] J. Bardzell. Interaction criticism: an introduction to the practice. *Interacting with Computers* 23, 604–621, 2011.

[2] J. Bardzell, J. Bolter, and J. Lowgren. Interaction criticism: three readings of an interaction design, and what they get us. *Interactions,* 32–37, 2010.

[3] S. Bardzell. Feminist HCI: Taking stock and outlining an agenda for design. *Proceedings of ACM CHI 2010,* 1301–1310, 2010.

[4] S. Bardzell, et al. Critical design and critical theory: the challenge of designing for provocation. *Proceedings of ACM DIS 2012,* 288–297, 2012.

[5] B. Broad. *What we really value: beyond rubrics in teaching and writing assessment.* Logan, UT: Utah State University Press, 2003.

[6] B. Broad. Introduction. *Organic writing assessment: dynamic criteria mapping in action.* Logan, UT: Utah State University Press, 2009.

[7] 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.

[8] B. Diederich, J. French, and S. Carlton. Factors in judgments of writing ability. Princeton, NJ: ETS report number RB-61-16, 1961.

[9] M. Feinberg. Personal expressive bibliography in the public space of cultural heritage institutions. *Library Trends* 59(4): 588–606, 2011.

[10] M. Feinberg, G. Geisler, E. Whitworth, and E. Clark. Understanding personal digital collections: an interdisciplinary exploration. *Proceedings of ACM DIS,* 200–209, 2012.

[11] N. Fuhr, et al. Evaluation of digital libraries. *Int'l Journal of Digital Libraries* 8: 21-38, 2007.

[12] B. Gaver, et al. The Prayer Companion: openness and specificity, materiality and spirituality. *Proceedings of ACM CHI 2010,* 2055–2064, 2010.

[13] B. Gaver. Cultural commentators: Non-native interpretations as resources for polyphonic assessment. *International Journal of Human-Computer Studies* 65: 292–305, 2007.

[14] G. Geisler. Open Video Digital Library Toolkit software. Documented at http://www.open-video-toolkit.org/

[15] H. Hsieh, et al. The "City of Lit" digital library: a case study of interdisciplinary research and collaboration. *Proceedings of JCDL 2012,* 203-212, 2012.

[16] B. Huot. Toward a new theory of writing assessment. *College Composition and Communication* 47(4): 549–566, 1996.

[17] M. Khoo and C. MacDonald. An organizational model for digital library evaluation. *Proceedings of TPDL 2011,* 329-340, 2011.

[18] B. Lee, et al. Designing with interactive example galleries. *Proceedings of ACM CHI 2010,* 2257–2266, 2010.

[19] 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.

[20] J. Lowgren and E. Stolterman. *Thoughtful interaction design.* Cambridge, MA: MIT Press, 2004.

[21] C. Lynch. Digital collections, digital libraries, and the digitization of cultural heritage material. *First Monday* 7(5): 2002. (Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/949/870)

[22] P. Marty and M. Kazmer. Introduction to understanding users. *Library Trends* 59(4): 563–567, 2011.

[23] J. McCarthy and P. Wright. *Technology as experience.* Cambridge, MA: MIT Press, 2004.

[24] P. Moss. Can there be validity without reliability? *Educational Researcher* 23(2): 5–12, 1994.

[25] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. *Proceedings of ACM CHI 1990,* 249-256, 1990.

[26] P. O'Neill, C. Moore, and B. Huot. *Guide to college writing assessment.* Logan, UT: Utah State University Press, 2009.

[27] J. Parkes. Reliability as argument. *Educational Measurement: Issues and Practice,* Winter 2007.

[28] D. Petrelli, et al. Digital Christmas: an exploration of festive technology. *Proceedings of ACM DIS 2012,* 348-357, 2012.

[29] P. Sengers, et al. Reflective design. *Proceedings of AARHUS 2005,*49-58, 2005.

[30] A. Wolff, et al. Storyspace: a story-driven approach for creating museum narratives. *Proceedings Hypertext 2012,* 89-98, 2012.