Characterizing Web-Based Tutorials: Exploring Quality, Community, and Showcasing Strategies

Matthew Lount
University of Manitoba
Winnipeg, MB, Canada
matthew.lount@gmail.com

Andrea Bunt
University of Manitoba
Winnipeg, MB, Canada
bunt@cs.umanitoba.ca

ABSTRACT

End-user authored tutorials found on the Web are increasingly becoming the norm for assisting users with learning software applications, but little is known about the quality of these tutorials. Using quality metrics derived from previous work, we perform a usability expert review on a sample of Photoshop tutorials, a popular image-manipulation program with one of the largest showings of web-based tutorials. We also explore how the characteristics of these tutorials differ across four tutorial sources, representing those that are, i) written by a close-knit online community; ii) written by expert users; iii) most likely to be found; and iv) representative of the general population of tutorials. Our analysis reveals that expert users generally write higher quality tutorials, and that many of the tutorials in our sample suffer from some important limitations, such as lacking attempts to help users avoid common errors. We also find that a single five-star rating system did not sufficiently distinguish quality between the tutorials. Building on this later finding, we propose and evaluate a rating approach based on multiple criteria, finding strong initial support for such an approach.

Categories and Subject Descriptors

H.5.2 User Interfaces: Training, help, and documentation

General Terms

Documentation, Design, Human Factors.

Keywords

Web-Based Tutorials; Usability Expert Review; Software Learnability; Categorical Rating Schemes

1. INTRODUCTION

As feature-rich applications continue to offer an increasing number of commands, attaining expertise in the use of the application becomes increasingly difficult. At the same time, there has been a growing trend for internet sites to allow users to post their own content. As a consequence, there are numerous online communities that allow people to share tutorials for using applications. Thus, rather than consulting help manuals or documentation designed in-house, users now often turn to search engines (e.g., Google) for help [7], and in doing so, are inevitably pointed to tutorials authored by other end-users.

This movement from in-application documentation and professionally authored manuals to end-user generated tutorials raises a number of important questions about their utility as a help resource, including: (1) What is the quality of the tutorials that are available for users on the web? (2) Does the quality of tutorials vary according to characteristics of the authoring community? (3) Are currently used mechanisms for highlighting tutorial quality (i.e., five-star rating schemes, and the number of views) effectively distinguishing between the available tutorials? In this paper, we explore these questions to shed light on existing authoring practices and the current quality of available documentation. We also aim to provide concrete data with which to ground the design of tutorial authoring tools and tutorial selection mechanisms, both of which have received significant attention in the research literature in recent years (e.g., [6, 7, 10, 12, 16]).

To characterize current web-based tutorials, we performed a usability expert review. To assist in this review, we first collated a series of quality metrics from the literature on effective tutorial design. We then applied these metrics to 154 tutorials, sampled across four different tutorial sources: i) a close-knit online authoring community, ii) those written by expert users, iii) popular tutorials retrieved through Google Search (using the CUTS method [9]), and iv) a set of tutorials representative of the general population of tutorials. In sampling existing tutorials, our methodology shares similarities with an observation study, with the aim of understanding current practices. We focused our exploration on text- and image-based tutorials for Photoshop, an image-manipulation application with one of the largest showings of online tutorials (e.g., over 18,000 on www.tutorialized.com). While video tutorials are becoming increasingly popular, we felt that text- and image-based tutorials represented a good starting point for exploration in this space. For example, in comparison to text- and image-based tutorials, there is limited information in the research literature on what makes for an effective video tutorial, making it difficult to collate a series of established quality metrics.

Our review indicates that most tutorials in our sample adhere to many established guidelines, providing initial evidence that these resources are useful forms of application learning. At the same time, there are a number of ways in which the tutorials are falling short, suggesting concrete opportunities for increased authoring support. We also found a number of clear differences in the tutorial sources, with a tutorial source contributed to by expert Photoshop users having a number of advantages over the remaining sources.

In terms of showcasing mechanisms, we found that five-star ratings tend to cluster around the middle of the scale regardless of tutorial quality, limiting their effectiveness as a discriminator in this particular domain. Motivated by this finding, we developed a set of rating categories that function as an alternative to the single

SIGDOC 2014, September 27–28, 2014, Colorado Springs, CA, USA.

overall rating, and conducted a user study to examine their desirability and effectiveness. Our study provides initial evidence that users do take multiple factors into consideration when assigning a rating, suggesting that their feedback would better captured through multiple categories than a single overall assessment.

This paper makes three primary contributions. (1) Informed by prior work, we propose a set of metrics to assess the quality of image- and text-based tutorials. (2) Applying these metrics to characterize a sample of Photoshop tutorials on the web, we highlight opportunities to design tools to better support tutorial authors. (3) Based on the limitations of a single 5-star rating scheme uncovered in our analysis of web-based tutorials, we propose and provide initial validation of a categorical rating approach.

2. RELATED WORK

We focus our coverage of related work on two areas. We first touch upon prior work on creating effective instructional materials, revisiting this work in more detail when we define the quality metrics used in our usability expert review of web-based tutorials. We then turn to research on systems that support tutorial authoring and consumption.

2.1 Designing Effective Instructional Materials

There is a large body of work on how to best instruct novice users in learning systems. Much of this work has focused on effective media use. For example, Booher [3] and Stern [24] both advocated combining text and images, with Stern adding that voice was not good at disseminating procedural instructions. More recently, Palmiter and Elkerton compared video instruction with more traditional text- and image-based tutorials [22]. They found that video has the potential to speed task completion, but only when there is a close match between the task and the video. They also note that traditional tutorials seem to produce better long-term learning. Recent evidence has shown that this is not always the case. ToolClips [11], which integrates short video clips into tooltips, has been shown to be a highly effective learning aid, even over the long term.

With a better understanding of what media work best for instruction, more recent work has focused on a fine-grained examination of how to design instructions (e.g., [2, 5, 13, 14, 21]). We use this body of work to define our quality metrics, which we describe in the "Defining Quality Metrics" section.

2.2 Supporting Tutorial Authoring and Consumption

Lately, there has been increasing interest in facilitating the process of creating tutorials. One approach has been to leverage application logging. For instance, Grabler *et al.* [10] created a system that automatically authors step-based tutorials by combining command-usage recordings with image-recognition software. MixT similarly uses demonstration to automatically generate tutorials, mixing videos into each of the textual steps [6]. FollowUs records not just the author's demonstration, but also the tutorial users' actions, and presents these videos to future users [18]. Finally, Kim *et al.* [14] designed a system that enables crowd workers to segment video tutorials into key steps.

Alternative forms of tutorials have also been investigated. For example, Chronicle [12] allows users to explore a video recording of the creation of an image file, providing tools to locate points of interest within the history. Fernquist *et al.* [8] created Sketch-Sketch Revolution, a system that goes beyond task-centric tutorials by also trying to assist with and teach fundamental drawing concepts, and adapts its assistance methods to the skill level of the user. Tapp Cloud, created by Laput *et al.* [19], turns pre-existing tutorials into macros, allowing users to quickly complete the tutorials' steps.

Finally, given the sheer volume of web tutorials available, work has also focused on how to assist users in deciding which tutorials will best support their goals. Ekstrand *et al.'s* [7] help system integrates lists of interface components into Google search results. In a similar fashion, the Delta [16] system displays tutorials' commands, final images, and numbers of steps. To complement these types of command-oriented approaches, Bunt *et al.* [4] proposed also displaying a summary of community feedback via tagged user comments.

While the above discussion highlights that there is a wide body of literature on supporting tutorial authoring and use, little is known about the current state of online tutorials. One exception to this comes from Lafreniere *et al.* [17] who examined tutorial comment sections to understand the manner in which people use tutorials. To our knowledge, this paper represents the first attempt to systematically characterize web-based tutorials themselves.

In the upcoming sections we describe our method for reviewing and characterizing online tutorials. We begin by describing properties of the tutorial sources that we considered in this work. We then examine tutorial showcasing techniques utilized by these sources to illustrate why we were not able to use user-contributed data, such as ratings, to guide our discussion of tutorial quality. Next, we define a set of metrics for assessing tutorial quality, and describe how we applied them to a sample of on-line tutorials.

3. TUTORIAL SOURCES

In this section, we describe the properties of the four sources of Photoshop tutorials included in our analysis: an application-centered community, a tutorial aggregator, a tutorial factory, and popular task-based tutorials.

3.1 Application-Centered Community

An application-centered tutorial community is typically a tight-knit group of application-specific enthusiasts who share their knowledge and skills with other members of the community, and often the public at large. Members of these communities range in experience from new users, to those who use the application as part of their professions, with tutorials being written in their spare time. The tutorials are not offered for a price, although authors may be profiting indirectly through related services. These websites are generally more than simply lists of tutorials, also providing forums, member services, and private messaging facilities. The application-centered community that we used was Renderosity (www.renderosity.com).

3.2 Tutorial Aggregator

Tutorial aggregator sites typically do not host the actual tutorials, but link to as many as they can, and generally for a variety of applications. This means that members of these communities do not necessarily write the tutorials that they post, but that the sites have a wide sampling of the tutorials that are available across the web. It is unclear if community members are paid to collect tutorials, or if they collect them for more altruistic reasons. Interactions between member users are limited; the term community is used very loosely for this type of site. We chose

Tutorialized (www.tutorialized.com) as our tutorial aggregator, which currently has links to over 18,000 Photoshop tutorials.

3.3 Tutorial Factory

A tutorial factory is a site run by a company that pays people to author tutorials. Our observations of these sites indicate that the authors are typically expert users of an application, either involved in developing the system itself, or in using the application in their daily lives. For this type of tutorial community, we chose PhotoshopTutorials (www.photoshoptutorials.ws), where the authors are paid between $150 and $300 per tutorial, and include short biographies discussing their credentials and expertise.

3.4 Popular Task-Based Tutorials

For the fourth source, we used the CUTS technique [9, 17] to generate the three most common "how to" searches pertaining to Photoshop. Searching for tutorials related to these queries provides a source of popular task-focused tutorials. This source is thus meant to represent the tutorials that an average user is most likely to view when searching for tutorials using Google. We elaborate on our use of CUTS when describing our data collection method.

4. ANALYZING SHOWCASING STRATEGIES

The goals of our study are threefold: i) to analyze the quality of online tutorials, ii) to examine differences in quality between tutorial sources and iii) to understand how effective current showcasing methods are at highlighting differences in quality. As indicated earlier, we focused this initial exploration on Photoshop tutorials, given their prevalence on the Web. We began by looking at data for the showcasing methods themselves to get a sense of their ability to act as a measure of quality.

Of our four tutorial sources, Tutorialized was the only site to post such data, which included the average rating, number of votes, number of views, and date posted. We used an automated system to collect values for each of these measurements from all tutorials available on the site.

Our first point of analysis was to examine the average rating displayed to users for all tutorials on Tutorialized; Tutorialized rounds the rating down to the next half star on a five-star scale. Table 1 shows that out of the 18,133 ratings we collected, 16,632 (91.7%) were either 2.5 (57.2%) or 3.0 (34.5%); another 707 (3.9%) had the rating of 2.0, resulting in 95.6% of the ratings being clustered into three of the nine possible values. This implies that either there is nearly no diversity of quality, or that the displayed (i.e., averaged) ratings are not functioning as accurate discriminators.

Table 1: Votes by Ratings

Visual Rating	% of Tutorials n = 18,133	Median Number of Votes	Inter-Quartile Range of Votes
1.0	1.1	1	0
1.5	0.2	1	10
2.0	3.9	23	43
2.5	57.2	67	103
3.0	34.5	50	131
3.5	0.7	22	29
4.0	1.2	1	0
4.5	0.1	2	1
5.0	1.1	1	0

The number of votes is intended to add confidence to a tutorial's rating, but, as can be seen in Table 1, the vast majority of the tutorials with ratings other than 2.5 and 3.0 have very few votes. We also found that the number of votes is linearly influenced by the date that the tutorial was uploaded ($p < 0.001$, $R^2 = 0.493$). From this, we infer that the number of votes is at least partially representative of how old the tutorial is, and that the more people who vote, the more central the rating becomes.

The number of views also appeared to provide limited discriminating information, as these values are compounded by the amount of time the tutorials are online, activity on the site, and the likely inaccurate star ratings.

5. DEFINING QUALITY METRICS

Given that user-contributed quality assessments do not appear to represent an accurate estimate of tutorial quality, our next step was to define our own set of quality metrics. As a starting point, we examined the literature on teaching computer skills through procedural instructions, deriving a number of metrics from what the authors deemed important. We then augmented this set by interpreting comments posted by users about the tutorials. Below, we expand on how we derived these metrics, which are summarized in Table 2.

5.1 Metrics from Related Work

Booher [3] and Stern [24] found that a combination of printed words (M1) and images (M2) were required for tutorials to be effective, finding that images allowed for speed, while text was necessary for accuracy.

Based on their survey of prior studies ([2, 3, 13, 14, 21]), Grabler *et al.* [10] indicated that tutorials should use numbered steps (M3–M5), and combine text descriptions of the actions to take with screenshots of the results of completing the step (M6–M8). They specified that screenshots should contain relevant interface widgets (M9), with any needed parameters either filled in (M10), or specified in the description of the steps (M11). The authors also advised that annotating images increases understanding (M12–M13), specifically suggesting the use of arrows and highlighting to denote areas of interest. Finally, they suggested that repetitive steps should be condensed through references to past steps (M14).

Carroll [5] ran a series of studies on users' practices when learning new software. He found that helping users with potential errors should be included in any instructional material (M15). He also describes how users have a tendency to not read material fully, and to follow tutorials without understanding why they are doing things. Consequently, he emphasizes the value of explanations which describe *why* steps are being done (M16).

In their formative study of the Delta tool, Kong *et al.* [16] found that when searching for a tutorial, users emphasized the importance of knowing which commands are used in the tutorial (M17–M19), and being able to ascertain the results of following the tutorial (M20–M21).

Table 2: Metrics. M1–M21 came from related work. M22–M24 are based on user comments posted at the end of tutorials.

ID	Metrics and Sources	Additional Collection Information
M1	Number of images [3,24]	Automated collection. Manually removed ad images
M2	Number of words [3,24]	Automated collection. Manually removed unrelated headers and ad text
M3	Is step-based [10]	Delimits segments using white space, images, numbers, or bullets
M4	Has numbered steps [10]	Objective
M5	Number of steps [10]	Objective (given M3)
M6	Has end of step images [10]	Has screenshots of relevant interfaces or expected workspace at the end of some (other than the last) steps
M7	Number of end of step images [10]	Objective (given M6)
M8	Number of textual references to images [10]	Objective
M9	Number of images with tool palettes [10]	Any image of a toolbar or palette with a tool selected; unselected toolbars did not count
M10	Number of images with parameters [10]	Images with any values filled into text boxes, check boxes, etc.
M11	Number of parameters in text [10]	In-text values for text boxes, check boxes, etc.
M12	Number of images with annotations [10]	Objective
M13	Number of annotations in images [10]	Objective
M14	Repetitive steps condensed [10]	References past steps from similar steps
M15	Number of tips and hints [5]	Text that describes common problems and what can be done to mitigate those problems
M16	Number of explanations of why steps are conducted [5]	Text that describes why a step was taken (not just how to do it)
M17	Number of textual references to shortcuts [16]	Objective
M18	Number of textual references to menus [16]	Objective
M19	Number of textual references to tools [16]	Objective
M20	Presence of final image [16]	Objective
M21	Preview of final image [16]	Final image appears at the start of the tutorial, not just at the end
M22	Presence of source files	Includes a list of source files (just a starting image, if no other resources are used) and where to get them
M23	Presence of original image	Objective
M24	Specified version	Objective

5.2 Metrics from User Comments

The comments that most sites allow readers to post at the end of a tutorial provide additional information on what users find particularly useful about tutorials [17]. For instance, users often would ask for the location of the source files that are used in the tutorial (M22–M23). Users also often questioned which version of Photoshop was being used (M24), as each version results in significant changes being made to the software's functionality and interface.

6. DATA COLLECTION METHOD

For three of our tutorial sources, we randomly sampled a subset of the available tutorials: 20 from the 493 available on PhotoshopTutorials, 20 from the 126 available on Renderosity, and 100 from the 18,133 available on Tutorialized. Given that Tutorialized attempts to gather as many tutorials as it can, we felt that it should contain a representative range of tutorials available online.

We then supplemented this sample by applying the CUTS method in the manner described by Lafreniere *et al.* [17]. In particular, we used CUTS to select three common "how to" searches pertaining to Photoshop: i) "how to cut out an image in Photoshop", ii) "how to feather in Photoshop", and iii) "how to make a mix tape cover in Photoshop". For each of the above searches, we selected the first five unique tutorials that Google returned. We sampled a total of 14 tutorials using this approach since the third query returned only four unique tutorials within the first five pages of results. This brought our total sample size up to 154.

The first author manually examined the content of each tutorial in our sample and collected values for each of the 24 metrics listed in Table 2 (each tutorial requiring approximately one hour to code). Collection for values for metrics M1 and M2 was automated, while effects of ads were removed manually. Table 2 lists the rules followed coding the tutorials involved any subjectivity.

Table 3: Descriptive Statistics, Main Effects and Post-Hoc Analysis for the Quantitative Metrics. All metrics are normalized by step, except M5 and M17–M19. Shaded cells indicate instances where one community is significantly better than all the others.

Metrics	Overall Median (IQR)	Main Effect		Median (IQR)				Pair-wise Significance (p < 0.05)
		H	p <	Photoshop-Tutorials (P)	Renderosity (R)	Google (G)	Tutorialized (T)	
Images (M1)	1.35 (0.96)	23.304	0.001	2.07 (2.39)	1.00 (0.63)	1.07 (0.30)	1.50 (0.79)	P > R, P > G, T > R
Words (M2)	43.37 (45.92)	12.716	0.01	68.40 (85.47)	46.91 (126.42)	44.65 (36.28)	37.08 (38.17)	P > T
Steps (M5)	1.35 (0.96)	12.512	0.01	14.50 (21.50)	12.00 (15.25)	7.00 (8.00)	7.00 (8.00)	P > T
End of Step Images (M7)	0.67 (0.51)	18.991	0.001	0.87 (0.28)	0.69 (0.49)	1.00 (0.22)	0.57 (0.53)	P > T, G > T
References to Images (M8)	0.33 (0.67)	4.992	NS	0.65 (1.09)	0.34 (0.81)	0.17 (0.55)	0.31 (0.54)	NS
Images with tool palettes (M9)	0.56 (0.68)	17.479	0.001	0.88 (1.48)	0.27 (0.73)	0.33 (0.38)	0.56 (0.62)	P > G, P > R
Images with parameters (M10)	0.40 (0.60)	27.450	0.001	**0.74 (1.34)**	0.21 (0.59)	0.09 (0.25)	0.40 (0.57)	**P > All**, T > G
Parameters in text (M11)	0.00 (0.24)	25.019	0.001	**1.31 (3.45)**	0.40 (0.84)	0.17 (0.45)	0.46 (0.77)	**P > All**
Images with annotations (M12)	0.00 (0.38)	2.714	NS	0.09 (2.25)	0.00 (0.17)	0.00 (0.16)	0.00 (0.28)	NS
Annotations (M13)	0.11 (0.25)	3.072	NS	0.11 (3.64)	0.00 (0.28)	0.00 (0.21)	0.00 (0.42)	NS
Tips and hints (M15)	0.05 (0.29)	29.316	0.001	0.23 (0.24)	0.16 (0.80)	0.27 (0.45)	0.00 (0.17)	P > T, G > T, R > T
Explanations (M16)	1.00 (5.00)	24.656	0.001	0.22 (0.49)	0.17 (0.54)	0.07 (0.33)	0.00 (0.17)	P > T, R > T
References to shortcuts (M17)	2.50 (5.00)	15.784	0.001	**7.50 (16.50)**	0.00 (3.00)	0.00 (4.25)	1.00 (4.00)	**P > All**
References to menus (M18)	4.50 (10.00)	21.193	0.001	**7.50 (9.00)**	2.00 (4.75)	1.00 (3.00)	2.00 (4.00)	**P > All**
References to tools (M19)	8.00 (9.00)	25.350	0.001	**18.00 (16.75)**	4.50 (11.75)	4.00 (10.75)	3.50 (7.00)	**P > All**

7. RESULTS: TUTORIAL QUALITY

We begin by examining the overall quality of the tutorials, and then turn to a discussion of differences in quality between tutorial sources. To compensate for the variability in the length of the tutorials and complexity of the tasks associated with them, we normalized the quantitative metrics by dividing by the number of steps. There were three exceptions (M17–M19); since experience with commands, menus, and shortcuts contributes to promoting expertise of the system, we analyzed their raw counts.

7.1 Quality across all Tutorials

Overall medians and inter-quartile ranges (IRQs) can be found in Table 3 (second column), with percentage values for categorical variables listed in Table 4 (second column). Below, we highlight some of the more notable results, dividing our discussion into the things tutorials are doing well, where the results are mixed, and where the tutorials are falling short; these results are summarized in Table 5.

7.1.1 Where Tutorials are Adhering to Guidelines

When examining values for the quality metrics in relation to previously established guidelines, we see that there are numerous things that the majority of tutorials appear to be doing right.

For example, 141 (91.6%) of the tutorials contained the original images (M23) that were being worked on, and 148 (96.7%) showed the final image (M20) that the tutorial produced. With both of these included in a tutorial, the user is more easily able to verify that they completed the tutorial correctly.

Of the 154 tutorials collected, 130 (84.4%) had at least one image per step (M1). This means that most of the time, users are looking at tutorials that have images that go with the text, helping them to follow along.

Tutorials also tended to refer back to previous steps when possible (M14; 60/67 or 89.6% of tutorials containing repeated steps).

Finally, 90.3% of the tutorials included links to all the source files needed to complete the tutorial (M22).

Table 4: Analysis of the Categorical Metrics by Website. Column 2 displays the % of “Yes” counts overall, while columns 5–8 report the observed and expected “Yes” counts for each source. Likelihood ratios are reported instead of Pearson’s chi-square statistics for cases with expected counts smaller than five. Observed counts that differed significantly from expected counts are shaded (at $p < 0.05$).

Metric	% “Yes”	X^2	p <	PhotoshopTutorials		Renderosity		Google		Tutorialized	
				“Yes” Count	Expected	“Yes” Count	Expected	“Yes” Count	Expected	“Yes” Count	Expected
Step Based (M3)	97.4	2.9	NS	20	19.5	20	19.5	13	13.6	97	97.4
Numbered Steps (M4)	60.4	24.2	0.001	**20**	**12.1**	**5**	**12.1**	7	8.5	61	60.4
Repetitive Steps Condensed (M14)	89.6	17.6	0.01	**16**	**7.8**	8	7.8	4	5.5	**32**	**39.0**
Final Image Present (M20)	96.1	9.1	0.05	19	19.2	19	19.2	**11**	**13.5**	**99**	**96.1**
Preview of Final Image (M21)	55.2	14.7	0.01	**19**	**11**	10	11	7	7.7	**49**	**55.2**
Source Files Present (M22)	90.3	8.9	0.05	20	18.1	17	18.1	**10**	**12.6**	92	90.3
Original Image Present (M23)	91.6	2.9	NS	19	18.3	18	18.3	11	12.8	93	91.6
Specified Version (M24)	16.9	14.6	0.01	**8**	**3.4**	3	3.4	**5**	**2.4**	**10**	**16.9**

7.1.2 Mixed Quality

Our analysis also reveals some strong variation in the extent to which tutorials were following established guidelines.

An example of this is the number of tips and hints per step (M15) (median=0.11, IQR=0.25). The presence of tips and hints is an indication that the authors are considering the problems that may be encountered by the audience. Many tutorials (40.7%), however, contained no attempts to address potential errors, while another 9.3% have an average of one attempt in every nine steps. This suggests that authors might be taking for granted their ability to convey instructions to the users. Examining the third quartile, however, reveals that some tutorials are frequently including this type of assistance, providing tips and hints for one out of every four steps (0.25).

The textual metrics concerning references to interface elements (M17–M19) all have relatively low median values (1, 2.5, and 4.5 respectively). This indicates that, on average, these tutorials may be limited in their ability to increase a user's exposure to the application's command set. That being said, much like the number of tips and hints per step, the third quartile (5, 6, and 12 respectively) reveals that many of the tutorials have potential to provide quite a bit of exposure to new commands.

7.1.3 Room for Improvement

Our analysis reveals that the web tutorials in our sample are falling short in a few areas. The most obvious example concerns the number of explanations per step (M16) (median=0.05, IQR=0.3). This is likely to impact long-term learning, as it can help users call upon past experience to more readily understand what needs to be done, and apply the current steps to similar situations they encounter in the future. While explaining every step might not be necessary, amongst the tutorials we examined, approximately only one in every twenty steps was explained.

We also found that only 72 (47.8%) of the tutorials contained annotations in any of their images (M13). Annotations can help describe things that are much harder to convey through text, such as where to apply brush strokes, or what areas of the image to select.

Most authors are not including what version of Photoshop the instructions work for (M24), with only 26 (16.9%) of the tutorials containing version information. Version-related questions were fairly common in the comments, either asking how to complete a step in a given version, or asking what version was being used.

Users are unlikely to invest the time into a tutorial if they cannot see the effect that they will be creating ahead of time, but authors often failed to include previews of the final image (M21), with only 85 (55.2%) of the tutorials having them. This is especially an issue in long tutorials.

One of the more surprising things that we saw in our analysis was that authors frequently did not number their steps (M4), with only 93 (60.4%) of the tutorials doing so. This makes it difficult to refer back to specific steps, either by the author within the tutorial, or by the user when seeking help in the comments.

7.2 Tutorial Source Quality

In this section we examine differences between the tutorial communities. Table 3 presents medians and IQRs for each group, and summarizes the significant pair-wise comparisons. We performed the non-parametric Kruskal-Wallis tests for effects of tutorial source on the quantitative metrics, and Bonferroni corrections on post-hoc pair-wise comparisons.

As Table 3 illustrates, only three of the main effects were not significant: the number of textual references to images per step (M8), annotations per step (M13), and images with annotations per step (M12). For the remaining significant main effects, pair-wise comparisons showed that tutorials from PhotoshopTutorials were better than or equivalent to the remaining tutorial sources. In 10 of the 15 quantitative metrics, PhotoshopTutorials were superior to those from Tutorialized (M2, M5, M7, M10, M11, M15–M19). PhotoshopTutorials also had seven quality metrics with measures that were significantly higher than both Renderosity and Google (M1, M9–M11, M17–M19).

Table 5: Summarizing Overall Quality of Tutorials

Quality	Metrics
High	Source files included, repetitive steps condensed, original image, final image, # of images, images with parameters, references to parameters
Mixed	Tips and hints, references to tools, references to shortcuts, references to menus
Room for Improvement	Explanations, annotations, version present, previews of final image, numbered steps, textual references to images

While PhotoshopTutorials were significantly better than the other sources for a number of metrics, the data did not suggest a clear ordering among the remaining three sources. This was especially true for Google and Renderosity, where the pair-wise comparisons revealed no significant differences between the two sources.

For the categorical data, we calculated Pearson's Chi Square values and likelihood ratios, and found significant differences between the sources in six of them (M4, M14, M20–M22, M24). Table 4 summarizes these results, and lists both, the actual and expected counts (we present only the "Yes" counts for sake of simplicity).

In examining the differences between the actual and expected counts, the most striking differences are often found with PhotoshopTutorials. In several cases (M4, M14, M21, M24; $p < 0.01$), the observed counts are close to double the expected. Renderosity, Google, and Tutorialized, on the other hand, had lower than expected counts for a number of metrics.

7.3 Discussion

7.3.1 Quality across all Tutorials

Our data indicates that while many of the tutorials in our sample are following principles of effective tutorial design, there were also some notable omissions by the tutorial authors, revealing opportunities for tutorial authoring systems. Among these omissions is a lack of inclusion of both, explanations for why a step should be undertaken, and tips for completing more difficult steps. Despite the recent surge in automatic tutorial authoring tools (e.g., [6, 10, 18, 19]), we have not come across any that have focused on these aspects, except to note that authors could add this information after the tutorials are generated (e.g., [6]).

Annotations are another missed opportunity for tutorial authors. Annotations allow authors to easily show screen-location information, such as the path to move your mouse when using the sharpen tool, or the exact spot a blur should be focused on. In many cases, annotations allow still images to convey as much information as a video clip would, without the overhead for the reader that is associated with scrubbing through a video's timeline.

Other problems that we found would have low overhead for authors (a low incidence of numbering steps, and inclusion of version information and previews of the final image) and are all things that either have been addressed in the automatic authoring tools, or could be addressed with simple extensions.

Outside of the realm of automatic tutorial creation, future authoring tools could include additional scaffolding to encourage users to include the types of information described above, that tutorial authors are frequently omitted, with some automatic formatting of tutorials post-authoring being possible. Examples of forms the scaffolding can take include separating and numbering steps, and including in each step a place for instructions and a place for explaining what the purpose of the step is.

7.3.2 Quality Differences between Communities

Our results suggest that you get what you pay for when it comes to tutorial authorship. PhotoshopTutorials, the site made up of professional tutorial authors, was the only one to have significantly higher quality than all of the other sites for a given metric, and this happened in five cases. It also had significantly higher values for than the other sources for a number of metrics, and there were no metric in which it performed worse than any of the other sites. These findings indicate that tutorial communities comprised of novice or intermediate authors might benefit from posting authoring guidelines or including tools to help authors create tutorials that adhere to these guidelines.

We note that the tutorials made by both PhotoshopTutorials, and Renderosity were longer, more complicated tutorials than those found in Tutorialized (or retrieved by Google Search using the CUTS technique). Those in PhotoshopTutorials, in particular, tended to focus on creating a full scene from many photographs. The complexity of these tutorials could be as a result of the authors being more invested in their creation. In the case of PhotoshopTutorials, the authors have several things at stake: they are paid for the tutorial, and, since these tutorials are being posted online with their names on them, these tutorials can represent their skill with Photoshop, and may serve as parts of their more formal portfolio. Renderosity's tutorials are also associated with members' online personas.

7.3.3 Showcasing Strategies

One interesting finding is that the showcasing methods on Tutorialized do not appear to be serving their intended purpose; the ratings do not vary enough to provide users with any power to perform between-tutorial comparisons. One possible explanation for this lack of variability is that it is difficult for users to determine what to base their ratings on, since what makes for a good tutorial is multidimensional and user-dependent. Without direction, users might be placing emphasis on different aspects of the tutorial, having an unpredictable effect on the resulting measurements. In the next section, we propose an alternative showcasing strategy that is based on a categorical rating scheme.

7.3.4 Generalizability

In answering our research questions, we faced a tradeoff between depth and breadth. We chose to focus on depth, examining a single tutorial type across a range of tutorial communities. Further data collection and analysis would be needed to determine the extent to which our findings generalize to tutorials in other domains, with such exploration able to build on our methodology. For example, we collected the majority of our metrics from research on tutorials in general, and thus these metrics can be used as a basis for studying tutorials in other domains, particularly for feature-rich applications that produce a visual output. Extending this work to video-based tutorials, however, would require first establishing a set of relevant metrics.

8. CATEGORICAL RATINGS

Ratings are a primary means for users to provide quantitative feedback on the quality of web-based tutorials based on their own experiences attempting to complete these tutorials. To help address some of the previously described limitations of a single rating scheme, in this section we explore the potential utility and

Table 6: Rating Categories and I-Statements

Category	I-Statement
Error Prediction	I felt the author made adequate attempts at helping users to avoid potential problems.
Coolness	I thought the effect that this tutorial created was awesome.
Learning	I learned at least one new technique through following this tutorial.
Ease of Following	I found the instructions were clear and easy to follow.
Enjoyment	I enjoyed completing this tutorial.
Writing Style	I liked the style of the text of this tutorial.
Image Helpfulness	The images of the tutorial were helpful in completing the tutorial.
Overall	I found that this tutorial was good overall.

feasibility of a categorical rating scheme, where users rate tutorials along a number of dimensions.

Categorical ratings have been explored in the context of recommender systems. For example, Sahoo *et al*. [23], and Adomacivius and Kwon [1] both used categorical ratings to generate single values representing a movie's overall appeal to a given user. In contrast, Lee and Teng [20] reject the idea of combining ratings of multiple scales into a single value, indicating that this results in a loss of information that users could find useful. These studies suggest that there is value in breaking up how users rate things into multiple categories.

To explore the feasibility and suitability of categorical ratings as applied to tutorials, we constructed a set of possible categories, shown in Table 6. We then conducted a study to elicit users' attitudes towards this categorical rating scheme, which we describe next.

8.1 Study Method

We recruited 12 people to participate in the study (five female, seven male, ages 18 to 28), through signs posted on our university campus. Each of the participants had prior experience with photo-manipulation software. Participants were provided with $15 for their time.

Our study compared two conditions: a single rating scheme and a categorical rating scheme based on seven different categories. These categories (error prediction, coolness, learning, ease of following, enjoyment, writing style, and image helpfulness) were initially informed by comments that users post to tutorials, and were subsequently refined through pilot testing. Each of the ratings was accompanied by an 'I'-statement (see Table 6 for details), with the ratings for each category ranging from strongly disagree (1) to strongly agree (5).

The procedure for the experiment was as follows. Each participant attempted four tutorials. After completing each tutorial, half of the participants rated the tutorial using the single rating, and the other half used the categorical rating scheme (task order was counterbalanced using a Latin square). We selected the four tutorials such that they were of varying quality (according to our metrics). To allow for 60-minutes sessions, participants were allowed to spend a maximum of 10 minutes on each tutorial.

To provide additional comparison points, after completing all of the tutorials, we asked the participants to look back over the tutorials and rate them a second time using the other rating style. We then asked participants to rank the seven categories in order of importance, where importance was defined as the degree to which the categories mattered to the participant in terms of tutorial selection. The session concluded with a short semi-structured interview.

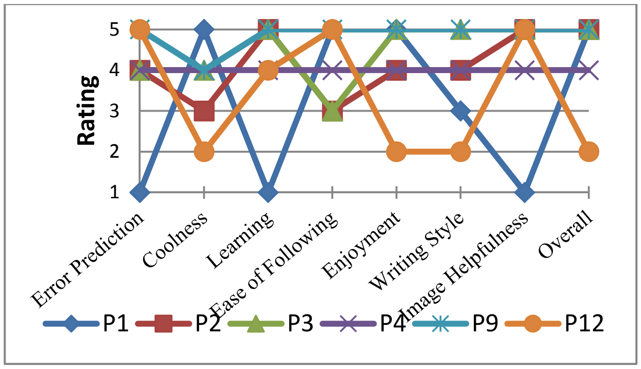

Figure 1: Ratings for task 1 for participants who rated the categories before providing the overall rating.

8.2 Findings

We focus our findings on participants' subjective impressions of the categorical rating scheme, the variability in their ratings across the categories and their thoughts on the relative importance of the individual categories.

When asked which of the two ratings schemes they preferred, all 12 participants indicated that they preferred rating on the set of categories to rating on the overall scale. The most commonly stated reason for this preference was the ability to provide more specific feedback. Participants also all felt that categorical rating data would be more useful for them in selecting tutorials, allowing them to focus on the things that they view as most important. For example, P6 said, *"[the categorical ratings] make it easier to evaluate the tutorial and know if indeed it is quite helpful"*. P5 agreed, indicating that he did not have confidence in the overall ratings:

> *"[The categorical ratings] are better, because they give a better kind of feedback […] but [the overall rating], you know, you could get lost trying to understand why the rating was this way."*

Participants' preferences were backed up by their rating data, which showed a great deal of variability within the ratings for each tutorial from individual users. Figure 1 illustrates participants' ratings for one of the tutorials (for the participants who rated the tutorial using the categorical scheme immediately after attempting the tutorial). From this figure, we see that only one of the participants (P4) selected a single score for all categories. In contrast, P1 used the entire range of the rating scale. Of the 48 sets of ratings that participants provided, only five sets had no variability (i.e., the participant provided the same rating for all categories for that particular tutorial). The mean number of points that a participant's seven category ratings varied from their corresponding overall rating was 5.04, with a standard deviation of 3.40.

Most users indicated that in choosing their overall rating, they considered their own performance at the task, citing the time it took to complete, and the difficulty they had with it. For example, P2 she said she rated it based on *"How fast I was"*, and *"if it didn't give me stress"*. Other participants spoke primarily to ease of following, but mentioned that they combined several different

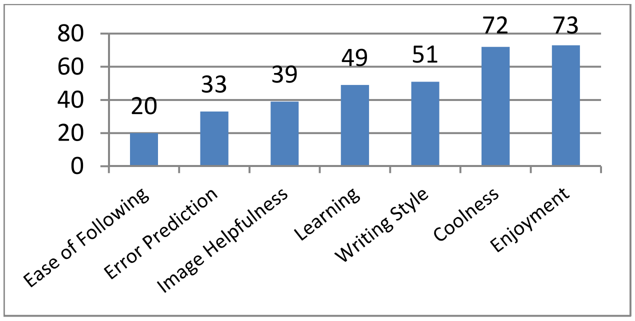

Figure 2: Sum of Participant Rankings of Categories.

concepts. P9 indicated that he primarily rated based on how much he enjoyed the images. *"I love colours, so when I was rating it, I was considering the colourful part of the tutorial, […] the images."* The above comments suggest that users assess tutorials using both varied and multiple criteria. They also suggest that the categories used in our study might require further refinement to cover aspects like efficiency and visual design of the tutorial itself (as opposed to the end result).

One concern with asking users to rate on multiple categories is that they may be unwilling to invest the extra time to do so. Of our twelve participants, only one indicated s/he would be unwilling to rate on more than one category, while four participants said they would be willing to rate as many as 10. The mean number that they said they were willing to rate was 5.9, with a standard deviation of 3.3. Participants indicated that they would be most keen to rate multiple categories if they had extreme reactions to the tutorial, such as finding it exceptionally helpful or difficult. Some participants also indicated that if there were too many categories, they would simply rate the ones they found most important. These results suggest that users do not feel that rating multiple categories would be prohibitive, and that when pressed for time, they might focus on the categories that they care most about.

Finally, participants' category rankings provide insight into their relative importance should tutorial websites wish to focus on only a subset of the categories examined in this study. Figure 2 displays the sums of participants' rankings (with lower sums implying greater importance). These rankings indicate that ease of following and error prediction were ranked as the most important overall by our participants, with coolness and enjoyment as the least.

8.3 Categorical Ratings: Discussion

The diversity of the ratings, as well as the participants' comments, support the notion that multiple rating categories are likely to provide a more comprehensive look at the quality of tutorials than the overall ratings do alone. While our study does not provide enough data to determine if the centralizing tendency that we saw with Tutorialized's ratings also exists with multiple categories, our results indicate that users do care about multiple tutorial dimensions and that they rarely feel that an individual tutorial performs equally on these dimensions. With a categorical rating scheme, tutorial websites could provide tutorial consumers with more discriminating data to guide their tutorial selection process. Our data also suggests that a categorical rating scheme would not be negatively perceived in terms of rating overhead.

9. SUMMARY AND FUTURE WORK

We presented a study of the current state of the quality of end-user authored online Photoshop tutorials, comparing communities of authors using quality metrics from the body of existing literature on tutorial design. We found that while many tutorials are of high quality, some authors likely require more scaffolding to create tutorials that follow the common practices of expert authors. Our results therefore motivate further research on tutorial authoring tools and highlight key areas to target. We also found that current methods of displaying information about tutorial quality suffer from a number of limitations and have proposed and explored a categorical rating scheme as a potentially richer source of quantitative user-supplied feedback.

In sampling a large number of Photoshop tutorials from the web and coding them according to established metrics, we have provided an in-depth characterization of current authoring practices for this target application and how these practices differ according to tutorial source. There are a number of important avenues of future work, one of which would be assessing the relative importance of these metrics from an individual user's perspective. Another promising direction would be to further explore the failing of the ratings in our sample to show any difference in level of quality, to determine whether our findings generalize to other tutorial-ranking websites, and to ranking websites in other domains. Finally, building on the methodology proposed in this paper, future work is needed to explore the generalizability of our findings to other types of tutorials, authoring communities, and tutorial formats, including the increasingly popular video format.

10. ACKNOWLEDGMENTS

We thank the National Sciences and Engineering Council (NSERC) and the Graphics, Animation and New Media (GRAND) Network Centre of Excellence for funding this work.

11. REFERENCES

[1] Adomavicius, G., and Kwon, Y. 2007. New recommendation techniques for multi-criteria rating systems. *IEEE Intelligent Systems 22*, 3 (May/June 2007), 48–55.

[2] Agrawala, M., Phan, D., Heiser, J., Haymaker, J., Klingner, J., Hanrahan, P., and Tversky, B. 2003. Designing effective step-by-step assembly instructions. *ACM Transactions on Graphics 22*, 3 (July 2003), 828–837.

[3] Booher, H. R. 1975. Relative comprehensibility of pictorial information and printed words in proceduralized instructions. *Human Factors 17*, 3 (June 1975), 266–277.

[4] Bunt, A., Dubois, P., Lafreniere, B., Terry, M. and Cormack, D. 2014. TaggedComments: Promoting and integrating user comments in online application tutorials. In *Proceedings of the ACM Conference on Human Factors in Computing Systems,* ACM, 4037–4046.

[5] Carroll, J. 1990. *The Nurnberg Funnel: Designing Minimalist Instruction for Practical Computer Skill*. The Massachusetts Institute of Technology Press, Cambridge, Massachusetts, USA.

[6] Chi, P.-Y., Ahn, S., Ren, A., Dontcheva, M., Li, W., and Hartmann, B. 2012. MixT: Automatic generation of step-by-step mixed media tutorials. In *Proceedings of the ACM Symposium on User Interface Software and Technology,* ACM, 93–102.

[7] Ekstrand, M., Li, W., Grossman, T., Matejka, J., and Fitzmaurice, G. 2011. Searching for software learning resources using application context. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, 195–204.

[8] Fernquist, J., Grossman, T., and Fitzmaurice, G. 2011. Sketch-Sketch Revolution: An engaging tutorial system for guided sketching and application learning. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, 373–382.

[9] Fourney, A., Mann, R., and Terry, M. 2011. Characterizing the usability of interactive applications through query log analysis. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 1817–1826.

[10] Grabler, F., Agrawala, M., Li, W., Dontcheva, M., and Igarashi, T. 2009. Generating photo manipulation tutorials by demonstration. *ACM Transactions on Graphics 28*, 3 (August 2009), 66:1–66:9.

[11] Grossman, T., and Fitzmaurice, G. 2010. ToolClips: An investigation of contextual video assistance for functionality understanding. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 1515–1524.

[12] Grossman, T., Matejka, J., and Fitzmaurice, G. 2010. Chronicle: Capture, exploration, and playback of document workflow histories. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, 143–152.

[13] Heiser, J., Phan, D., Agrawala, M., Tversky, B., and Hanrahan, P. 2004. Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ACM, 311–319.

[14] Kim, J., Nguyen, P., Weir, S., Guo, P.J., Miller, R.C., and Gajos, K.Z. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 4017–4026.

[15] Knabe, K. 1995. Apple guide: A case study in user-aided design of online help. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 286–287.

[16] Kong, N., Grossman, T., Hartmann, B., Fitzmaurice, G., and Agrawala, M. 2012. Delta: A tool for representing and comparing workflows. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 1027–1036.

[17] Lafreniere, B., Bunt, A., Lount, M., and Terry, M. 2013. Understanding the roles and uses of web tutorials. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, AAAI, 303–310.

[18] Lafreniere, B., Grossman, T., and Fitzmaurice, G. 2013. Community enhanced tutorials: Improving tutorials with multiple demonstrations. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, ACM, 1779–1788.

[19] Laput, G., Adar, E., Dontcheva, M., and Li, W. 2012. Tutorial-based interfaces for cloud-enabled applications. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, ACM, 113–122.

[20] Lee, H.-H., and Teng, W.-G. 2007. Incorporating multi-criteria ratings in recommendation systems. In *Proceedings of the International Conference on Information Reuse and Integration*, IEEE, 273–278.

[21] Novick, L. R., and Morse, D. L. 2000. Folding a fish, making a mushroom: The role of diagrams in executing assembly procedures. *Memory and Cognition 28*, 7 (October 2000), 1242–1256.

[22] Palmiter, S., and Elkerton, J. 1993. Animated demonstrations for learning procedural computer-based tasks. *Human-Computer Interaction 8*, 3 (September 1993), 193–216.

[23] Sahoo, N., Krishnan, R., Duncan, G., and Callan, J. 2012. The halo effect in multicomponent ratings and its implications for recommender systems: The case of Yahoo! movies. *Information Systems Research 23*, 1 (March 2012), 231–246.

[24] Stern, K. R. 1984. An evaluation of written, graphics, and voice messages in proceduralized instructions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, HFES, 314–318.