Computers & Education 58 (2012) 551–559

Contents lists available at SciVerse ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Studying teacher selection of resources in an ultra-large scale interaive system: Does metadata guide the way?

Samuel Abramovich*, Christian Schunn

Learning Research and Development Center, University of Pittsburgh, Suite 830, 3939 O'Hara St., Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 19 January 2011
Received in revised form
16 August 2011
Accepted 2 September 2011

Keywords:
Computer-mediated communication
Cooperative/collaborative learning
Human-computer interface
Improving classroom teaching

ABSTRACT

Ultra-large-scale interactive systems on the Internet have begun to change how teachers prepare for instruction, particularly in regards to resource selection. Consequently, it is important to look at how teachers are currently selecting resources beyond content or keyword search. We conducted a two-part observational study of an existing popular system called TeachersPayTeachers hypothesizing that 'evaluative metadata' (i.e. comments, ratings, and popularity measures) would drive selection of resources. The first part examined patterns in tens of thousands of sales overall, and the second part focused on patterns of sales in one focal topic that could be expert coded. We find that there are significant gaps in available metadata, that some aspects of metadata are closely associated with sales, and that metadata are weak correlates of expert-determined quality. We conclude by making suggestions for additional research and suggesting how ultra-large scale-interactive systems such as TeachersPayTeachers could be used to improve teacher education.

1. Introduction

The Internet and Internet-based technologies allow for the creation of what has been dubbed ultra-large-scale interactive systems; systems that leverage millions of users and huge databases to create unique and complex tools (Gabriel, Northrop, Schmidt, & Sullivan, 2006). Such systems, such as YouTube, Ebay, and Wikipedia, have begun to impact many facets of modern society and, because of the newness of these technologies, science has just begun to unravel how people use them.

Ultra-large-scale interactive systems have also begun to change how teachers prepare for instruction (Franklin, 2007; Jaen, Bohigas, & Novell, 2007; Russell, Bebell, O'Dwyer, & O'Connor, 2003). Currently, a teacher with access to the Internet now has at their disposal an almost unlimited number of resources provided by large interactive online systems. With this access, the previous problem of how to give teachers access to the resources (Greenhow, Robelia, & Hughes, 2009; Watson, 2006) has been supplanted with the new challenge of how to support teachers' effort in finding the appropriate resources (Hill & Hannafin, 2001; Jaen et al., 2007). This challenge is particularly important given that the creation and analysis of teacher resources provides a learning opportunity for teachers (Ball & Cohen, 1996; Leat & Higgins, 2002; Sparks & Loucks-Horsley, 1989) that can result in improved student learning.

Fortunately, ultra-large-scale interactive systems have various features designed to help users navigate the large amount of resources made available. The most common method is the inclusion of search engines (Fallows, Rainie, & Mudd, 2004), such as is found in Amazon. For example, users can search Amazon for a book based on content or metadata like genre or date of publication. However, search engines using only content and genre/date metadata are possibly limited in their ability to return optimal results. Specifically, a lack of completeness, accuracy, or uniqueness of a resource's content and metadata could generate search results that are skewed or incorrect.

Another common method for assisting user search is to allow for and take advantage of user ratings or comments on a specific resource (Schafer, Konstan, & Riedi, 1999); these ratings and comments are what we call "evaluative" metadata. Users looking for a book at Amazon can narrow their search results by looking at what ratings or comments that other users, often described as the crowd, have left about the resources. These ratings and comments allow for lists of resources to be ordered by ratings, essentially creating recommendations from the

* Corresponding author. Tel.: +1 412 624 3909.
E-mail addresses: sja24@pitt.edu (S. Abramovich), schunn@pitt.edu (C. Schunn).

0360-1315/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2011.09.001

input of other users. Evaluative metadata is commonly found in systems that wish to leverage their users' ability to evaluate in order to provide additional filtering of their vast resources. Other examples include Youtube, which provides ratings and comments of submitted movies, and Ebay, which provides ratings and comments of buyers and sellers. Of course, ratings can be biased and comments can be noisy, limiting the benefit of evaluative metadata.

If we are to better understand how ultra-large-scale systems impact teachers then it is important to look at how teachers are currently selecting resources beyond content or keyword search. How are teachers using evaluative metadata when selecting resources in ultra-large-scale interactive systems? Based on the preponderance of crowd-sourced "evaluative" metadata in highly successful commercial websites, we selected what we observed to be the most prevalent types: comments, ratings, and popularity. Based on their association with commercial success, we hypothesized that these three specific "evaluative" metadata would drive teacher selection of resources in large-scale online systems of teacher materials. To test our theory, we conducted an observational study of an existing popular ultra-large-scale online system called TeachersPayTeachers.

1.1. Theoretical Background

A recurring theme in research on how teachers use technology and the Internet involves ways in which metadata could contribute to more optimal resource selection. Recker et al. (2005) conducted educator workshops and found gaps in metadata that otherwise would have helped educators discover appropriate resources. Wang and Hsu (2006) designed a demonstration system for how different types of searches of metadata can result in good resource selection. Others have explored the design of systems in which accurate metadata could also help teachers find resources that minimize the need for adaption for instruction (Recker, Dorward, & Nelson, 2004; Suthers, 2001). Many computer scientists have focused on the creation, testing, and implementation of metadata standards for online learning resources (Anido et al., 2002; Bohl, Schellhase, Sengler, & Winand, 2002; Gonzalez-Barbone & Anido-Rifon, 2008; Plodzien, Stemposz, & Stasiecka, 2006).

Additionally, research on measuring usability in digital libraries has created heuristic models that can be applied to a variety of repositories of online resources (Jeng, 2005). Often the models for high quality digital libraries include objective measures of metadata quality (Goncalves, Moreira, Fox, & Watson, 2007) or suggest methods for improving metadata accuracy (Nesbit, Belfer, & Vargo, 2002).

Seemingly absent is research on how metadata impacts teacher selection of resources in the type of large-scale interactive environments becoming more common on the Internet. This absence is indicative of the social-technical gap that often occurs in this type of research (Ackerman, 2000). The social approach examines the social needs of a particular group, in this case teachers, given existing technologies and suggests how new technology could meet those needs. The technology approach usually designs new technologies to benefit a social group, such as designing innovative technology to benefit teachers. The missing approach involves examining how new technology is currently being used by these social groups, such as how teachers are currently using innovative technology.

To help close this gap, a good strategy would involve examining teacher use of metadata in ultra-large-scale interactive systems. Large-scale interactive environments present unique opportunities for system designers and teachers to access large quantities of materials. Furthermore, metadata can provide a window into how teachers are finding resources in these systems.

We limited our investigation by focusing on the use of evaluative metadata. There are studies that have looked at how best to incorporate evaluation of resources into searchable metadata (Nesbit & Li, 2004; Nesbit, Li, & Leacock, 2006) and examined how teacher communities can collaborate to provide more accurate evaluative metadata (Recker, Walker, & Lawless, 2003). Given that Internet search engines are becoming more powerful and that non-evaluative metadata is static, we believe that finding content-relevant teaching resources based on non-evaluative metadata will eventually become a relatively straightforward process. We suggest that evaluative metadata, in terms of its potential ability to distinguish quality levels among relevant resources, is where the power of ultra-large-scale interactive systems lies. However, this expectation does not imply that all methods of collecting and using evaluative metadata will be successful—there likely will be some socio-technical challenges to effectively manage collection and use.

Conducting any laboratory analysis of ultra-large scale interactive systems is unrealistic simply because of the number of subjects involved. Finding thousands of users to participate in an experiment is impractical. The only practical option is to analyze a system that currently operates with both a very large user-base and a very large amount of resources. These requirements necessitate finding a system that is reliable and sustainable (Schlager, Fusco, & Schank, 2002) with a critical mass of users and a robust incentive system for attracting new users, adding new resources, and evaluating existing resources (Ackerman, 2000). Analyzing a currently operating system also has the added benefit of placing our findings within the reality of current teacher behavior.

1.2. Data Source

The online system we examined is a website called TeachersPayTeachers (TPT) that allows teachers to sell resources they have designed and created directly to other teachers. The operators of TPT graciously provided us with a complete copy of the main database with user name and credit card information removed. Included in the database were resource names, descriptions, overall rating of the resource, number of times a resources had been rated, number of times a resources had been sold, price of the resource, comments left by users for each resource, a resource ID number, and the ID number of the author. Also included was metadata generated by the author of the resource and displayed on each resource's webpage in TPT. The metadata consisted of grade level appropriateness, subject area, and file type. Standard web metrics such as web page hits were not available.

The following description of TPT is a result of direct interaction with the system along with discussion with TPT's owner. All of the information provided has been confirmed for accuracy with the owner.

Operating since February of 2006 to March of 2010 at the time of the database dump, TPT had over 200,000 registered users, 62,781 resources, and over $900,000 in sales. TPT allows teachers of numerous subjects, grade levels, and experience to put lesson plans, unit plans, assessments, or other teacher resources up for sale. Built into the system is the opportunity for buyers of materials to generate metadata in the form of several ratings and comments. Sellers categorize their resource according to subjects, grade level appropriateness, and type of resource such as assessment, worksheet, presentation, and review materials. Buyers can search the available resources by the non-evaluative metadata such as subject, grade level, type, and keyword. Searches can also be based on evaluative metadata such as what

are the bestselling resources and which resources have the highest rating. Buyers can further use content descriptions and evaluative metadata to then inform their purchases from the searched items.

Writing reviews of resources that cost money in TPT is only allowed for registered users who have purchased and downloaded the resource. The rubric is a rating scale with 8 rankings in the form of grade letters (F, D, D+, C, C+, B, B+, A) across six categories: Overall Quality, Accuracy, Practicality, Thoroughness, Creativity, and Clarity. Ratings for each category are transformed by TPT to a five-point scale (0–4 presented to the tenth decimal point), averaged, and displayed prominently for each resource. In addition, alongside the author's name is the average rating of all of their submitted resources. Users entering a rating also have the option of writing a comment about the resource that will be displayed along with the average rating of the resource.

A typical use case scenario would begin with a teacher logging into the system and entering some keywords along with a specific grade level and subject area. The results of the search are resources listed according to the relevancy of the search. The teacher can then scroll through the list looking at a short summary, overall rating from other users, number of times the resource had been rated, subject areas, grade levels, type of resource such as assessment or activity, instruction time required, author, and a price. The list can be resorted according to rating, alphabetical according to resource name, price, sales, and date of creation. Once the teacher selects a resource that seems promising they can view the resource's page (see Fig. 1) where a more detailed description is available along with a preview, individual ratings, individual comments, a question and answer forum for each resource, and a short biography of the resource creator. The teacher can then choose to add the resource to their shopping cart or go back to the search results to look for a more promising resource. The teacher can also choose to browse the collection of resources directly looking at a specific grade level, subject area, price category, or bestselling materials. Once a resource has been purchased, teachers are encouraged to use the purchased resource in instruction before returning to rate it (see Fig. 2). Ratings are not required but encouraged (Fig. 3).

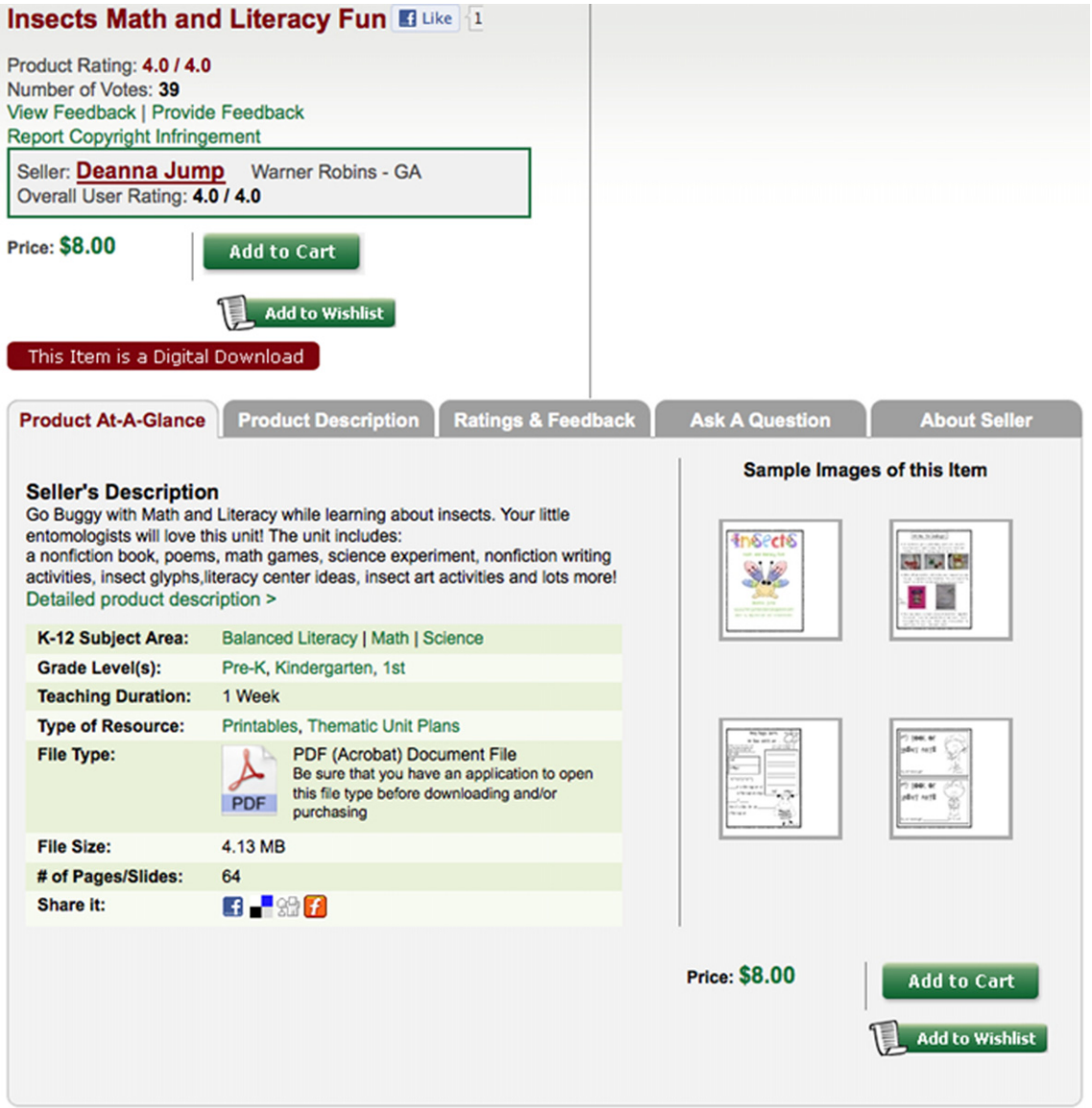

Fig. 1. Resource Page.

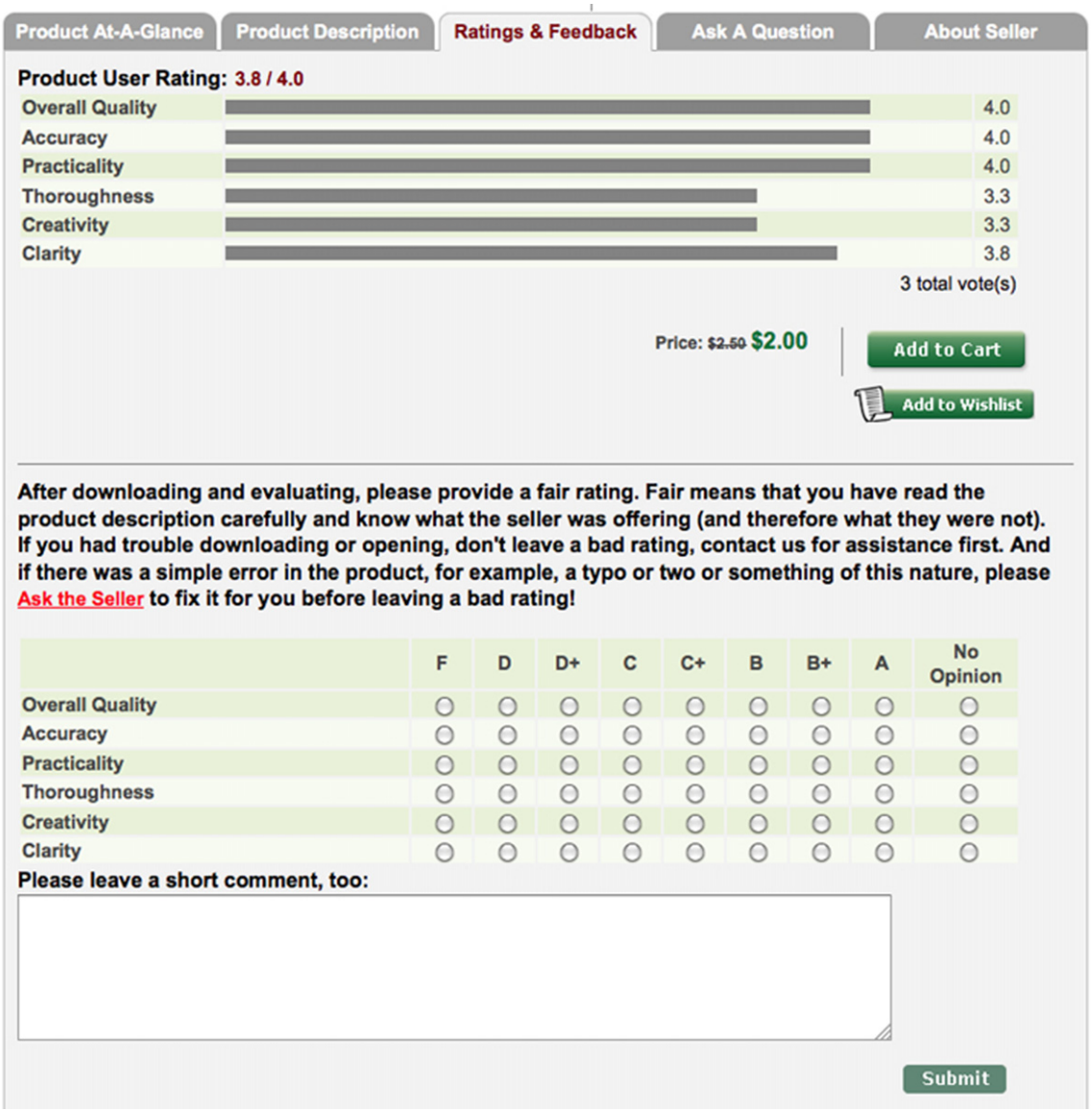

Fig. 2. Rating Rubric for Resources.

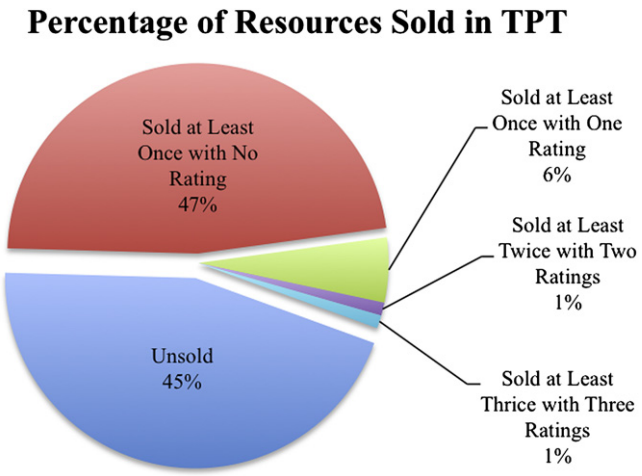

Fig. 3. Percentage of Resources Sold in TPT with Number of Ratings.

TPT meets the previously outlined requirements for sustainability and incentivized participation. The large number of participants insures a critical mass of resource submissions and purchases. The fact that resources are sold incentivizes sellers to provide resources and the overall low cost of the resources, typically 1 to 5 dollars per resource, incentivizes buyers to purchase resources.

As an ultra-large-scale interactive system, the issue of search and selection is critical. What evaluative metadata are teachers using to select resources?

2. Study 1: Evaluative metadata consistency

2.1. Methods

The measure by which resource metadata is useful to teachers is to the degree that teachers can find, identify, select, and acquire the resource that they are seeking (Bruce & Hillmann, 2004). To abstractly determine metadata quality, we could have examined completeness, accuracy, provenance, consistency, conformance, or accessibility (Bruce & Hillmann, 2004). Each of these measures has drawbacks for the purpose of determining a practical use to teachers. For example, it is hard to attach an analysis to a rating that could be made for subjective reasons.

TPT introduces another quantitative measure of curriculum value, namely sales. Teachers spend an average of $623 dollars a year of their own money on classroom preparation (National Teaching Realities Survey, 2010), a non-trivial expenditure for a teacher in the current global economy. As a result, the number of sales of a resource indicates a certain valuation by the purchasing teachers. Evaluative metadata that correlates positively with sales might indicate teacher criteria for selection.

Because evaluative metadata is optional in TPT, it is not necessarily the case that users would rate products or leave comments. The absence of such metadata would limit its use. Further, large skews in the metadata, such as all positive reviews, would limit the value as well. So to determine what evaluative metadata guides teachers to select resources in an ultra-large-scale interactive system we first looked for some overall patterns of the evaluative metadata. Our research questions were:

1. What percentages of items that are sold receive evaluative metadata (i.e. have a rating)?
2. What is the distribution of evaluative metadata (i.e. ratings), positive and negative, amongst items that are sold?

To examine the content of the thousands of comments in the database, we used Pennebaker, Booth, and Francis's Linguistic Inquiry and Word Count (LIWC) (2007). LIWC is a computer program that can identify positive and negative affect words with high reliability and has been successfully used in a variety of research contexts (e.g., Paletz & Schunn, 2009).

Having established the basic descriptive data, we then turned to correlational analyses of evaluative metadata and sales:

3. Does the evaluative metadata relate to how the resources are valued? Is the evaluative metadata and average sales related?

2.2. Results & discussion

Mining the TPT database for answers to these questions revealed some surprising findings. The vast majority of sold items never receive any additional evaluative metadata. Of the 34,602 items that were sold at least once, approximately 86% of the sold items never receiving any type of rating, 10% rated once, 2% rated twice, and 2% rated three times or more. Similarly, 88% of all items sold at least once receive no comments, 7% received one comment, 3% received two comments, and 2% received three comments or more. This level of activity is analogous to the 90-9-1 rule of participation (Nielsen, 2006).

Turning to the content of the ratings, of those items that receive at least one rating, the vast majority of the ratings are very positive: 76% received a mean rating of three or higher and 40% have a mean rating of four, the highest possible rating. A similar pattern was observed with resource comments left by users. According to our LIWC analysis, positive comments outweighed negative comments 13 to 1. Thus the evaluative metadata generated in the form of ratings or comments is usually very positive and generally not very discriminating. The large amount of positive ratings and comments could be the result of teachers not perceiving quality differences between different resources and finding them all sufficient for instructional purposes. Another possibility could be that teachers are only choosing to provide ratings and comments in the positive cases. Additional research is necessary to better understand the nature of the positive ratings.

To examine the relationship among variables, we present analyses using Pearson correlations and linear regression—the R^2 captures the percent of variance accounted for by the analyses. The large skew in all the variables raises a question about whether raw or transformed data, or whether Pearson or Spearman correlations should be used. There are a large absolute number of data points in the tails. Further, those tails are pragmatically important to viability of TPT and describe the most common monetary transactions. Therefore, we prefer to report the Pearson correlations that proportionally weight the contributions of these data points in the tails. However, analyses using log-transformed data or Spearman correlations produce similar conclusions.

Despite the relatively small percentage of items that are rated, the very large pool of sold items ($N = 34{,}602$) meant that we could still determine with high precision how closely sales are related to evaluative metadata. The correlations are presented in Fig. 4.

Three of the four key relationships were highly statistically significant and two of the four were at least moderate in strength. Positive emotion words were significantly associated with sales, but at a very weak level; further the relationship with negative emotion words was not statistically significant ($r = -0.02$, $p > 0.19$). Ratings correlate with sales at a weak level, which could indicate that sales are weakly related to material quality, as according to TPT users, or that displayed ratings of quality is weakly related to sales.

Another salient finding is that sales are more strongly related to number of ratings and comments than to average rating. Combining number of ratings and comments into a linear regression reveals that they account for just under half of the variance in sales ($R^2 = 0.47$, $F\ (2{,}34{,}602) = 15{,}415$, $p < 0.0001$), with independent statistically significant contributions from both factors. Here there is likely

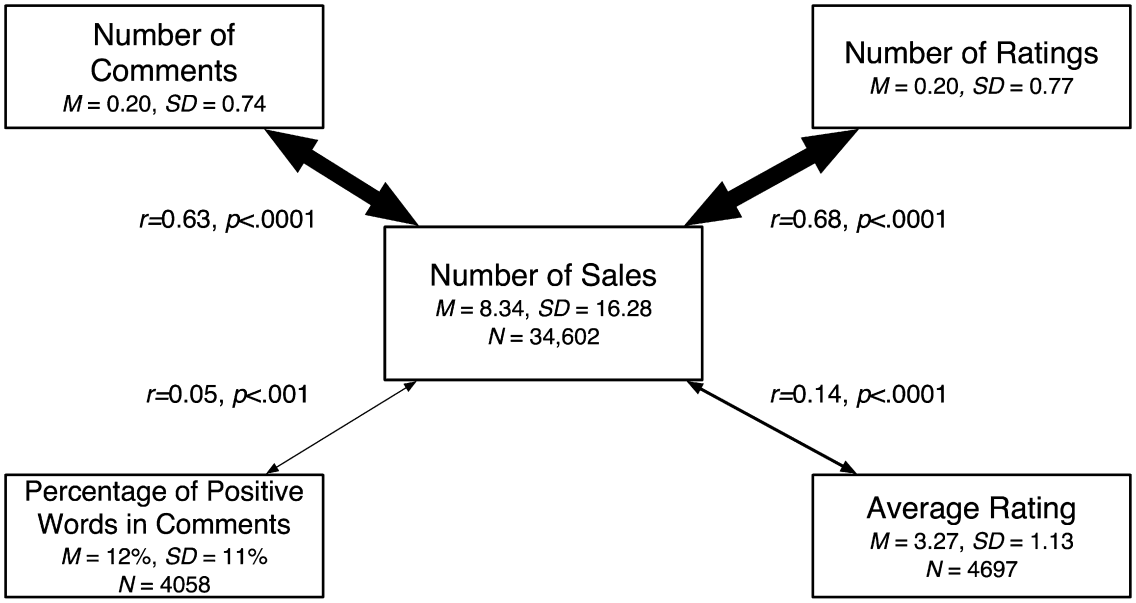

Fig. 4. The strength of correlations between Evaluative Metadata and Sales, with Evaluative Metadata broken into comments vs. ratings by their number vs. content. Thicker arrow bars represent a stronger correlation with sales.

a bidirectional influence: items that have higher sales also have more opportunities for generating evaluative metadata while evaluative metadata may influence others to buy.

The small correlation between sales and mean rating does not have that possible mediation through number of opportunities to provide ratings so the evidence of a direct but weaker relationship is more evident. However, it is important to acknowledge possible third variable confounds, such as some topics being more popular with higher quality ratings and higher sales. Including topics, grade, type of resources in the multiple regression did not reduce the relationships between mean quality ratings and sales, and thus a topic-based third variable confound is unlikely to be the driving factor.

But the initial aggregate analyses for number of ratings and number of comments do not specify a causal directionality because we did not have time-stamped information about when the evaluative metadata was generated. We cannot definitively determine what were the number of sales for a resource prior to generation of the evaluative metadata and what were the number of sales after the evaluative metadata was visible to potential buyers. To partially address this issue, we ran the same statistical analyses for items that sold at least fifteen times. The premise for this analysis is that buyers will eventually be influenced by a critical number of ratings and comments no matter what other reasons for purchasing a resource (Standifird, 2001). Consequently we hypothesize that by the time a resource has been purchased a minimum of fifteen times then the critical number of ratings and comment necessary to influence sales will have been reached. The minimum of fifteen sales is a simple threshold chosen for being well above the average number of sales per resource while still providing a large enough number of resources to merit statistical analysis.

The same statistical patterns that existed for all items that sold at least once also existed for those items that sold 15 times or more as well, including the variance of sales predicted by number of ratings and comments ($R^2 = 0.48$, $F(2,5080) = 2332$, $p < 0.01$). If the effect had been purely from sales to opportunity to rate and comment, then we would have expected the correlation between sales and number of ratings and comments to be at least partially attenuated in this higher sales subset.

3. Study 2: Evaluative metadata validity

Although the initial analyses suggest a weak relationship between sales and user quality ratings and comments, we do not yet know whether the evaluative metadata could indicate a quality resource because we do not yet know the validity of user ratings and comments. Further, it is interesting to examine whether number of comments and ratings is perhaps also indirectly correlated with resource quality. With such a skew in user ratings, perhaps the presence of evaluative metadata could be a useful clue to teachers regarding material quality. Finally, to evaluate the overall success of the search tools in providing access to higher quality materials, it is interesting to see how sales are associated with material quality. Therefore, our next step was to determine the relationship of TPT's evaluative metadata (sales, mean rating, comment content, number of ratings, and number of comments) to actual material quality by comparing them to ratings generated by curriculum experts.

3.1. Methods

TPT includes content from a very wide range of domains and therefore it was necessary to focus on a particular content domain in order to narrow the scope of our study to meet our available capitals. We chose to analyze math resources because of our access to math resource experts in addition to the widely recognized importance of providing math teachers with good curriculum materials (Stein, Smith, Henningsen, & Silver, 2000). The percentage of resources in TPT varied somewhat by subject area, with math resources accounting for 11% of the

Table 1
Number of math resources selected for analysis by experts by removing resources no longer for sale, removing cases too rare to produce informative results, and balancing cell size across cells that could be examined.

Number of Items per Cell		Number of Ratings		
		Hi = 3+	Med = 2	Low = 1
Average Rating	Hi = 3–4	18 (of 19 possible)	21 (of 26 possible)	20 (of 137 possible)
	Med = 2–2.9	0 (of 1)	0 (of 2)	20 (of 21 possible)
	Low = <2	0 (of 0)	0 (of 2)	24 (of 28 possible)

total. The patterns that we established in Study 1 for all TPT resources also existed for the subset of math resources, further legitimizing our selection of the domain for further study.

In the TPT database, there were 4060 math resources that had at least one sale in its database. Because we wanted to compare user ratings with experts, we eliminated those resources that did not have a rating. It is possible that some purchases of resources are made simply for curiosity or an impulse purchase, especially when offered for free or very cheaply, so we also restricted the pool of math resources to those that cost $2 or more. After further eliminating resources that were no longer for sale (and thus we could not acquire them), we randomly selected a stratified sample of resources that represented the different combination of number of ratings and average rating for the resources of TPT from a possible pool of 236 math resources (see Table 1). Because of how ratings were distributed, not all combinations of rating number and average rating occurred often enough to be studied. Resources that were rated frequently and given a poor or average rating were extremely rare. Consequently our sampling was done such that the correlations of expert ratings with each of quality and number of ratings or average rating could be examined while holding the other dimension constant. We eliminated certain cells from our analysis since they would not have provided enough resources to allow for statistical findings. This filter resulted in our elimination of the rare cases of any resource that had 2 or more ratings in addition to an average rating of less than 3. For the resources in the remaining cells, we selected all available resources (that were still for sale). The one exception to this complete inclusion was for resources that were rated highly only once, which had a pool of 137 possible resources. Of those 137, we selected 20 at random in order to prevent them from having an overwhelming affect on our analysis based on sheer sample size and to make the rating task for experts manageable. This sampling choice was also consistent with our overall research approach of examining the naturally occurring behavior within the system. Overall, the resources had a mean price of $3.08 with a standard deviation in price of $1.29.

We selected five math resource experts based on the criteria that they had previously taught math in a K-12 school setting and were either pursuing their doctorate or had already received it in a field directly related to mathematics education. Three of the experts are currently studying advanced math education topics and one rater, with a doctoral degree in education, is an assistant professor in a math department.

The experts were given the same rubric and description of the resource available in TPT. The resources were randomly presented in their original format in an online structure similar to how the resources are made available in TPT.

For 12 of the resources, some change to the format was required. These resources were originally formatted for a Smartboard and had to be converted to a more compatible web format. In these cases, the experts, all familiar with how a Smartboard works, were told which resources were converted for their review and were asked to judge those resources more favorably on the basis that some formatting had been altered likely to the resources' detriment.

Judging the quality of instructional resources is a difficult task, even for experts. Ideally it requires knowing the intended learning goals and the way in which the materials are to be used (Kesidou & Roseman, 2002). Experts were told to use the information provided by the seller regarding intended use and intended goals. However much information was provided. Resources could be classified by their author for any grade level and up to three different categories. Although the experts reported the task difficult, reliability of the expert judgments at the level of average quality rating was moderate but acceptable given the difficulty of the task and the unknown quality of the rubric used by TPT, Cronbach's Alpha = 0.64.

Because there were a more manageable set of comments, the comment content for all 103 resources were classified by hand as either positive, negative, or simply resource related (e.g. "I used this resource in my classroom."). By way of validation, LIWC ratings were significantly correlated with human ratings: positive rated comments with percentage of positive emotion words ($r = 0.61$, $p < 0.01$), and negative rated comments with percentage of negative emotion words ($r = 0.28$, $p < 0.01$). The lower correlation of negative ratings likely stems from their rare occurrence (see below).

3.2. Results

Unsurprising to us, expert ratings of the resources were much lower than the ratings given by users of TPT. The average expert rating was 1.9, with no expert giving a mean rating for a resource higher than 2.9, in contrast to a mean user rating of 3.3 for these items. Experts in general tended to be harsher in their evaluation than novices. We fully expected this result given that experts have the ability to be more critical because of their familiarity with the many potential flaws found in typical resources for supporting student learning.

It is interesting that expert ratings were significantly correlated with sales ($r = 0.19$, $n = 103$, $p = 0.04$). Thus, there is some evidence at the TPT structure does enable some discrimination according to expert-determined material quality, although the relationship is weak. Further, this finding suggests that users could use prior sales information as a valid but weak clue regarding material quality.

Turning to the more putative system measures of material quality, the correlation between the expert ratings and the mean rating from TPT was not statistically significant ($r = 0.15$, $p = 0.13$). Thus, the ratings found in the system are not in themselves useful predictors of quality according to experts. Because we selected items with varying quality ratings, the explanation cannot lie in the skew in the ratings. However, the very small number of ratings used to produce these means may be an underlying cause. It is likely that there is a lot of variability across teachers just as there is across experts and only one rating is not sufficient data to produce a reliable quality estimate.

An examination of the content of comments in TPT also revealed no significant correlations. The number of overall comments ($r = 0.01$, $p = 0.45$), the number of positive comments ($r = 0.01$, $p = 0.54$), the number of negative comments ($r = 0.01$, $p = 0.26$), or the number of resource comments ($r = 0.01$, $p = 0.91$) were not correlated with expert ratings.

We could discern no other statistically significant correlation between the experts' rating of resources and the other metrics from TPT other than sales, including the correlation between experts' rating and the number of ratings ($r = 0.14$, $p = 0.14$).

The correlation between expert ratings and number of ratings was the closest to statistically significant, but when ran as a multiple linear regression with sales and number of ratings as the independent variable and expert ratings as the dependent variable, the independent contribution of number of ratings was not statistically significant ($p = 0.17$).

Our conclusion is that the positive ratings and comments described in Study 1 are not the equivalent of expert analysis. Additionally, the other available metadata for each resource is also not the equivalent of expert analysis. The positive correlation between sales and expert ratings does indicate to a degree that teachers are purchasing quality resources but the criteria for selecting the resources is probably not directly based on the evaluative metadata available in TPT.

4. General discussion

So what evaluative metadata drives teachers to select resources in ultra-large-scale interactive systems? Sales did correlate weakly with TPT ratings and, to an even lesser degree, expert ratings. These statistically significant relationships could indicate that the bestselling items in TPT are due in part to the evaluative metadata listing their rated quality. However, the correlations between user ratings and actual quality are low enough that it would be hard to conclude that an expert-determined level of quality was anything more than a minor influence on sales. A safer inference is that mean rating metadata does have a small influence on sales but not to a degree that a buyer can be assured of quality based on the rating alone or that a seller can be assured that a highly selling item is because of higher ratings than a poorly selling item. These findings are important in that they serve as a benchmark for comparison to other teacher resource systems that rely on peer-based ratings.

A stronger correlation occurred between sales and the number of ratings and comments of a TPT resource. One explanation of this relationship is that the relative scarcity of evaluative metadata accentuates those items that do receive evaluation. Because TPT has so many resources, using the system without whittling down choices might be overwhelming (Schwartz et al., 2002). Users might seek some initial way to differentiate the vast number of items in the database in order to choose from a smaller pool. Items that have evaluative metadata are rarer and, by setting a minimum rated threshold, allow a user to have a more comfortable item pool to choose from, even if those ratings are negative (Smith, Menon, & Sivakumar, 2005).

Turning to the issue of the accuracy of the evaluative metadata, it is not surprising to find that experts rated the resources in TPT much lower than TPT users. Curriculum experts should have a much higher expectation of quality for teacher resources or then they wouldn't be experts. In addition, teachers may have rated the materials on a different basis than the ones sellers intended or experts used, resulting in a higher rating. The reliability of the rating rubric must also be examined to investigate consistency of the ratings across users.

Another reason why number of ratings might have a large effect on sales is that some of the TPT default views list the most reviewed items in descending order. With companies such as Google and Microsoft producing more optimized search engines it is possible that users have increasing amounts of faith in how search engines list results.

4.1. Future implications

Despite the findings presented, additional research is necessary to better understand how teachers interpret and use evaluative metadata in resource exchange systems. We only conclude that the evaluative metadata in TPT is not the equivalent of expert reviews. We hope to further unpack how teachers use metadata when selecting resources with additional research such as surveys and case studies.

Our findings do allow us to theorize about possible motivations that drive participation in TPT. The most obvious potential motivators are profit for sellers and the ease of finding a specific resource for buyers. However, ultra-large-scale interactive systems must also rely on a community of users who identify with the goals of the system and provide the data necessary for operation. These communities can be based on social-bonds or identity-bonds. Social-bond based communities are where the participants have a social connection with each other that exist independent of the online community. An example would be a community of teachers at a singular school. Identity-bond based communities occur where the participants desire to be in a community is based on a perceived identification with the community independent of any social-bonds. An example would be a music teacher who seeks an online community of other music teachers despite being part of a school's teacher community (Ren, Kraut, & Kiesler, 2007).

TPT could also rely on an identity-based community to insure its success, namely the contribution of resources and associated metadata. The resource classification and ratings system are designed to be similar to what teachers use for their students, utilizing commonly accepted vocabulary and performance indicators. Essentially TPT, capitalizing on that teachers have a strong identity with their professional community, encourages an identity-based community (Goodson & Cole, 1994). While we were not focused on understanding how motivation impacts teacher resource selection, this topic is an important subject to explore if we are to form better conclusions about teacher behavior online. For instance, altering the incentives for evaluating resources to better leverage the benefits of an identity-based community could encourage more generation of evaluative metadata (Cheng & Vassileva, 2006) and increase TPT's ability to more accurately evaluate resources.

An ultra-large-scale interactive system could fit a model of a knowledge building community (Scardamalia & Bereiter, 1994) or of a resource-based learning environment (Hill & Hannafin, 2001). By being focused on the creation and review of teaching materials such as problems, lessons, or assessment the system would fulfill some of the elements of good teacher professional development include being content specific (Borko & Putnam, 1995), use of actual curricular materials (Ball & Cohen, 1999; Hawley & Valli, 1999), providing opportunity for reflection (Hawley & Valli, 1999), and being ongoing (Hawley & Valli, 1999). Findings from research on ultra-large-scale interactive systems could benefit teachers who do not use online resources by uncovering improvements that could be applied to all teacher

professional development. Another benefit could be an improved machine learning approach to identifying resource quality (Bethard, Wetzer, Butcher, Martin, & Sumner, 2009) or more efficiency in the usability design of the systems (Sumner, Khoo, Recker, & Marlino, 2003).

As we develop a better understanding of how teachers currently use these systems we can close the social-technical gap of our understanding and begin to mold better learning opportunities for our teachers. Hopefully, in the near future, we can directly measure the effect between ultra-large-scale interactive systems on student learning.

References

Ackerman, M. S. (2000). The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human-Computer Interaction, 15*(2), 179–203.

Anido, L. E., Fernandez, M. J., Caeiro, M., Santos, J. M., RodrIguez, J. S., & Llamas, M. (2002). Educational metadata and brokerage for learning resources. *Computers & Education, 38*(4), 351–374.

Ball, D., & Cohen, D. (1996). Reform by the book: what is: or might be: the role of curriculum materials in teacher learning and instructional reform? *Educational Researcher, 25*(9), 6–14.

Ball, D., & Cohen, D. (1999). Developing practice, developing practitioners: toward a practice-based theory of professional education. In L. Darling-Hammond, & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). San Francisco: Jossey-Bass Inc.

Bethard, S., Wetzer, P., Butcher, K., Martin, J. H., & Sumner, T. (2009). Automatically characterizing resource quality for educational digital libraries. In *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries* (pp. 221–230). Austin, TX: ACM.

Bohl, O., Schellhase, J., Sengler, R., & Winand, U. (2002). The Sharable Content Object Reference Model (SCORM) - a critical review. In *Proceedings of the international conference on computers in education* (pp. 950). IEEE Computer Society.

Borko, H., & Putnam, R. (1995). Expanding a teacher's knowledge base: a cognitive psychological perspective on professional development. In T. Guskey, & M. Huberman (Eds.), *Professional development in education: New paradigms and practices* (pp. 35–65). New York, NY: Teachers College Press.

Bruce, T. R., & Hillmann, D. (2004). The continuum of metadata quality: defining, expressing, exploiting. In D. Hillmann, & E. Westbrooks (Eds.), *Metadata in practice* (pp. 238–256). Chicago, IL: ALA Editions.

Cheng, R., & Vassileva, J. (2006). Design and evaluation of an adaptive incentive mechanism for sustained educational online communities. *User Modeling and User-Adapted Interaction, 16*(3), 321–348.

Fallows, D., Rainie, H., & Mudd, G. (2004). The popularity and importance of search engines. In *Pew Internet & American Life Project*.

Franklin, C. (2007). Factors that influence elementary teachers use of computers. *Journal of Technology and Teacher Education, 15*(2), 267.

Gabriel, R. P., Northrop, L., Schmidt, D. C., & Sullivan, K. (2006). Ultra-large-scale systems. In *Companion to the 21st ACM SIGPLAN symposium on object-oriented programming systems, languages, and applications* (pp. 632–634). Portland, Oregon, USA: ACM.

Goncalves, M. A., Moreira, B. L., Fox, E. A., & Watson, L. T. (2007). What is a good digital library?" - a quality model for digital libraries. *Information Processing & Management, 43*(5), 1416–1437.

Gonzalez-Barbone, V., & Anido-Rifon, L. (2008). Creating the first SCORM object. *Computers & Education, 51*(4), 1634–1647.

Goodson, I., & Cole, A. (1994). Exploring the teacher. *Teacher Education Quarterly, 21*(1), 85–105.

Greenhow, C., Robelia, B., & Hughes, J. (2009). Learning, teaching, and scholarship in a digital age - web 2.0 and classroom research: what path should we take now? *Educational Researcher, 38*(4), 14.

Hawley, W., & Valli, L. (1999). The essentials of effective professional development: a new consensus. In L. Darling-Hammond, & G. Sykes (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 127–150). San Francisco: Jossey-Bass Inc.

Hill, J., & Hannafin, M. (2001). Teaching and learning in digital environments: the resurgence of resource-based learning. *Educational Technology Research and Development, 49*(3), 37–52.

Jaen, X., Bohigas, X., & Novell, M. (2007). The need for virtual information managers in education. *Computers & Education, 49*(2), 254–268.

Jeng, J. (2005). What is usability in the context of the digital library and how can it be measured? *Information Technology and Libraries, 24*(2), 46–56.

Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? Findings from Project 2061's curriculum review. *Journal of Research in Science Teaching, 39*(6), 522–549.

Leat, D., & Higgins, S. (2002). The role of powerful pedagogical strategies in curriculum development. *Curriculum Journal, 13*(1), 71–85.

National Teaching Realities Survey. (2010). Culver City, CA: Kelton Research. http://multivu.prnewswire.com/mnr/officemax/43900/.

Nesbit, J., Belfer, K., & Vargo, J. (2002). A Convergent participation model for evaluation of learning objects. *Canadian Journal of Learning and Technology, 28*(3), 105–120.

Nesbit, J., & Li, J. (2004). Web-based tools for learning object evaluation. In *International conference on education and information systems: Technologies and Applications*, Orlando, FL.

Nesbit, J., Li, J., & Leacock, T. (2006). Web-based tools for collaborative evaluation of learning resources. *Journal on Systemics, Cybernetics and Informatics, 3*(5), 102–112.

Nielsen, J. (2006). Participation inequality: encouraging more users to contribute. In *Jakob Nielsen's alertbox (2011)*.

Paletz, S. B. F., & Schunn, C. D. (2009). A new metric for assessing group level participation in fluid teams. In *Atlanta conference on science and innovation policy* (pp. 1–6).

Pennebaker, J., Francis, M., & Booth, R. (2007). *Linguistic Inquiry and Word Count (LIWC2007)*. Austin, TX: University of Texas.

Plodzien, J., Stemposz, E., & Stasiecka, A. (2006). An approach to the quality and reusability of metadata specifications for e-learning objects. *Online Information Review, 30*(3), 238–251.

Recker, M., Dorward, J., Dawson, D., Liu, Y., Mao, X., & Palmer, B. (2005). *You can lead a horse to water: teacher development and use of digital library resources*. In Digital Libraries, 2005. JCDL '05. Proceedings of the 5th ACM/IEEE-CS Joint Conference on (pp. 1–8).

Recker, M., Dorward, J., & Nelson, L. (2004). Discovery and use of online learning resources: case study findings. *Educational Technology & Society, 7*(2), 93–104.

Recker, M., Walker, A., & Lawless, K. (2003). What do you recommend? Implementation and analyses of collaborative information filtering of web resources for education. *Instructional Science, 31*(4), 299–316.

Ren, Y., Kraut, R., & Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. *Organization Studies, 28*(3), 377–408.

Russell, M., Bebell, D., O'Dwyer, L., & O'Connor, K. (2003). Examining teacher technology use: implications for preservice and inservice teacher preperation. *Journal of Teacher Education, 54*(4), 297–310.

Scardamalia, M., & Bereiter, C. (1994). Computer support for knowledge-building communities. *Journal of the Learning Sciences, 3*(3), 265–283.

Schafer, J. B., Konstan, J., & Riedi, J. (1999). Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on electronic commerce* (pp. 158–166). Denver, Colorado, United States: ACM.

Schlager, M., Fusco, J., & Schank, P. (2002). Evolution of an on-line education community of practice. In K. A. Renninger, & W. Shumar (Eds.), *Building virtual communities: learning and change in cyberspace* (pp. 129–158). Cambridge, U.K: Cambridge University Press.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: happiness is a matter of choice. *J Pers Soc Psychol, 83*(5), 1178–1197.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing, 19*(3), 15–37.

Sparks, D., & Loucks-Horsley, S. (1989). Five models of staff development for teachers. *Journal of Staff Development, 10*(4), 40–57.

Standifird, S. S. (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management, 27*(3), 279.

Sumner, T., Khoo, M., Recker, M., & Marlino, M. (2003). Understanding educator perceptions of "quality" in digital libraries. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on digital libraries* (pp. 269–279). Houston, TX: IEEE Computer Society.

Suthers, D. D. (2001). Evaluating the learning object metadata for K-12 educational resources. In *IEEE international conference on advanced learning technologies* (pp. 371–374), Madison, WI.

Wang, H.-C., & Hsu, C.-W. (2006). Teaching-material design center: an ontology-based system for customizing reusable e-materials. *Computers & Education, 46*(4), 458–470.

Watson, G. (2006). Technology professional development: long-term effects on teacher self-efficacy. *Journal of Technology and Teacher Education, 14*(1), 151–166.