How Much Information? Effects of Transparency on Trust in an Algorithmic Interface

René F. Kizilcec
Department of Communication, Stanford University
kizilcec@stanford.edu

ABSTRACT

The rising prevalence of algorithmic interfaces, such as curated feeds in online news, raises new questions for designers, scholars, and critics of media. This work focuses on how transparent design of algorithmic interfaces can promote awareness and foster trust. A two-stage process of how transparency affects trust was hypothesized drawing on theories of information processing and procedural justice. In an online field experiment, three levels of system transparency were tested in the high-stakes context of peer assessment. Individuals whose expectations were violated (by receiving a lower grade than expected) trusted the system less, unless the grading algorithm was made more transparent through explanation. However, providing too much information eroded this trust. Attitudes of individuals whose expectations were met did not vary with transparency. Results are discussed in terms of a dual process model of attitude change and the depth of justification of perceived inconsistency. Designing for trust requires balanced interface transparency—not too little and not too much.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): User Interfaces; K.3.1. Computers and Education: Computer Uses in Education.

Author Keywords

Interface Design; Algorithm Awareness; Attitude Change; Transparency; Trust; Peer Assessment.

INTRODUCTION

Advances in machine learning and artificial intelligence aim to meet the growing challenge of managing an abundance of information for human consumption. Algorithmic interfaces that curate online news stories, create custom radio stations, and personalize search results have become commonplace and seemingly indispensable. Yet many people are unaware of these systems' hidden intelligence despite their potential impact on society [9]. This raises a new set of questions for designers, scholars, and critics of media. The consequences of increased algorithm awareness through more transparent interface design are not well understood, especially in real world situations where the stakes are high. Transparency may promote or erode users' trust in a system by changing beliefs about its trustworthiness.

Trust is a key concern in the design of technology, as it affects the initial adoption and continued use of technologies [4, 24]. In light of people's tendency to treat new technologies as social actors [23], the present definition of trust draws on prior work in offline interpersonal contexts. Trust is understood as "an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited" (p.740) [5]. One way to assure individuals that they will not be exploited is through transparency in design, which may foster a better understanding of the system and the extent to which it is fair and accurate.

Early research on transparency in complex systems focused on explanations to expose the data or reasoning underlying a system's output [11]. One of the first artificial intelligence interfaces, MYCIN, provided explanations to help users understand its reasoning and instill confidence [2]. Providing explanations can increase performance on information retrieval tasks [13] and improve attitudes toward automated collaborative filtering [10]. An experiment in an e-commerce context found that complementing product recommendations with different kinds of explanations positively influenced consumer beliefs [31]. In particular, 'why', 'how', and 'trade-off' explanations raised perceptions of competence, benevolence, and integrity, respectively. Another experiment found 'why' explanations to increase recommendation acceptance but not trust in the system [6]. The evidence suggests that added explanations can promote positive attitudes toward a system, but not necessarily trust. In a qualitative study of factors influencing trust in a complex adaptive agent, system transparency emerged as a core theme in user interviews [8]. Increased transparency is also associated with fewer misconceptions [16] and higher confidence in system recommendations [29]. However, an experimental test of increased transparency in an e-commerce system found no gains in trust or perceived competence [21]. Finally, a study of the Facebook News Feed found that increased algorithm awareness may not raise satisfaction, though it can promote engagement with the service [7].

The available evidence on how transparency affects trust is mixed—some studies found positive effects, while others found no effect. The literature offers few rigorous experimental tests, and the ones reviewed above took place in controlled lab settings or on hypothetical e-commerce websites. The

CHI 2016, May 7–12, 2016, San Jose, CA, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3362-7/16/05...$15.00
DOI http://dx.doi.org/10.1145/2858036.2858402

present study addresses a number of these shortcomings by testing the effects of transparency in a natural and high-stakes environment. Additionally, the current experiment compares between three levels of transparency (low, medium, and high) and evaluates the moderating role of expectation violation, the extent to which the system output matches user expectations.

A MOTIVATING ANECDOTE

A true story inspired this research and informed the study design and hypotheses. In a large, in-person HCI class, some students noticed that they received lower homework grades than their peers who were in a discussion section with a different teaching assistant (TA). What happened was that each TA had graded all homework questions for a subset of students, resulting in inconsistent grading between TAs. Students who got harsher graders were naturally more upset. To resolve the issue, the instructor informed students that their grades would be statistically adjusted for this bias to make grades fair. This apparent solution comforted the students at first. However, when the instructor announced the original and adjusted grades, the students were once again upset about it. What went wrong? Three important constructs embedded in this narrative are trust in the system, transparency (revealing the grading procedure), and the violation of positive expectations (receiving unfair grades). In particular, there were three consecutive levels of transparency: no explanation, a purely procedural explanation, and additionally providing data. Around the time that this grading issue occurred, there was a relevant development in peer assessment practices in large online courses.

TRUST AND TRANSPARENCY IN PEER ASSESSMENT

Online peer assessment provides a suitable context to study the effects of interface transparency on trust. It is a natural environment with high stakes that parallels the context of the motivating anecdote, while the digital format enables random assignment to different experimental conditions. Peer assessment is a proven method for scaling the grading of a large number of assessments, such as is required in massive open online courses (MOOCs) [14, 19]. In peer assessment, every person evaluates several submissions by peers and has their own submission evaluated by several peers. Peer grading is often supplemented by self grading to encourage the development of self-evaluative skills and is generally thought to "augment student learning" [25]. Surveys found positive student attitudes toward peer grading, but also some concern over the fairness and reliability of peer grades [27, 26, 12, 32].

Traditionally, the final assessment grade is simply the mean or median of the peer grades, which turns out to be similar to formal instructor grades [25, 14]. Kulkarni and colleagues [14] tested peer assessment in a MOOC and found that 66% of median peer grades were within 10% of instructor grades. Piech and colleagues [19] were able to improve accuracy by 30% using algorithms that adjust grades for grader bias and reliability. Their "tuned models" of peer assessment were a substantial improvement, but it was unclear how to communicate this to online learners. The details of the algorithm would be overwhelming and most learners were still under the impression that peer grades were simple averages. A social media analogue of this issue is Facebook's News Feed ranking algorithm that is designed to provide a better user experience than a chronological view. Yet many users are unaware of this algorithm [7] or develop personal beliefs about it [22]. In both cases, people's mental model of how the system works, informed by analogues of the physical world [17], is not how the system functions. Finding out how the system actually functions could induce positive or negative attitudes toward it.

THEORY AND HYPOTHESES

Information processing plays a fundamental role in how transparent interface design influences a person's trust in the system. Dual process models of communication [3], and specifically of attitude change [18], posit that individuals process information either consciously or unconsciously. A violation of personal expectations, such as receiving a lower than expected grade, is expected to directly influence arousal valence (e.g., lose trust in peer grading) [3].

H1. Trust is lower if expectations are violated.

Moreover, expectation violation prompts individuals to process available information more consciously to find a possible justification for the inconsistency between expected and actual system output. This increased attention to the information provided facilitates changes in individuals' attitudes [18]. Specifically, when receiving a lower than expected grade, individuals pay more attention to available information.

H2. Changes in interface transparency affect trust depending on whether expectations are violated.

Procedural justice theory [15] posits that individuals can be satisfied with a negative outcome as long as the underlying procedure is considered to be just. Consistent with this theory, providing some information (i.e., explaining the grading procedure) fostered trust in the motivating anecdote. However, providing more information (i.e., adjusted and unadjusted grades) eroded students' trust. The additional information may have confused students and shifted their focus away from procedural justice and back to the unsatisfactory outcome. Accordingly, transparency about tuned peer grading [19] is predicted to have similar effects if expectations are violated.

H3. If expectations are violated, procedural transparency increases trust, but additional information about outcomes erodes this trust.

METHODS

Participants and Design

Participants were enrolled in a MOOC offered on the Coursera platform. The study only involved learners who participated in peer assessment by submitting an essay for peer grading. Out of 120 learners who took part in the study, 17 had either failed to self-assess their essay or submitted it too late for peer grading. All analyses were conducted on the remaining 103 learners: 33% women and average age = 37.15 (SD = 10.85), based on 79 participants' self-report. Each person was randomly assigned to a transparency condition: 39 low, 34 medium, 30 high. Once a learner and her peers had graded her essay, she would receive her combined and adjusted peer grade accompanied by different amounts of information about the grading process depending on the transparency condition.

Transparency Manipulation

In the *low transparency* condition (the system default), only one sentence was shown: "Your computed grade is X, which is the grade you received from your peers." In the *medium transparency* condition, more information about the computation of the final grade was provided: "Your computed grade is X, which is based on the grades you received from your peers and adjusted for their bias and accuracy in grading. The accuracy and bias are estimated using a statistical procedure that employs an expectation maximization algorithm with a prior for class grades. This adjusts your grade for easy/harsh graders and grader proficiency." In the *high transparency* condition, in addition to the explanation of the grading process, participants saw the raw individual peer grades they received and how these were adjusted to arrive at their final grade.

Measures

Immediately following the transparency manipulation, on the same web page that the grade information was presented, participants answered questions to measure their trust in the peer assessment system. Four items assessed facets of trust (i.e., attitude of confident expectation that one's vulnerabilities will not be exploited): 'To what extent do you understand how your grade is computed in peer grading?'; 'How fair or unfair was the peer grading process?'; 'How accurate or inaccurate was the peer grading process?'; and 'How much did you trust or distrust your peers to grade you fairly?'. Participants responded on construct-specific and fully labeled response scales with 5 points for the unipolar item about understanding ('No understanding at all' to 'Excellent understanding'), and 7 points for all other items (e.g., 'Definitely fair' to 'Definitely unfair'). As expected, ratings of system understanding, fairness, accuracy, and trust in peers' fair grading were highly correlated (Cronbach's $\alpha = 0.83$). They were combined by simple averaging into an index that measured each participant's trust in the system ($M = 3.55$, $SD = 1.13$, range from 0 to 6). In addition, each participant's self-assessment grade (*self grade*) and adjusted peer grade before late submission penalties (*peer grade*) were available. The difference between the two grades served as a measure of expectation violation, as either a binary (Figure 1) or continuous variable (Figure 2). In the binary case, expectation violation was defined as a self grade that was over 2 points above the peer grade.

RESULTS

Figure 1 shows the average trust index for participants who either received a grade that matched their expectations or one that violated expectations. A 2 (expectations violated vs. not violated) by 3 (transparency: low, medium, high) ANOVA was conducted to test the first two hypotheses. Consistent with *H1*, trust was lower when the received grade was worse than expected ($F_{1,97} = 13.4$, $p < 0.001$, $\eta_p^2 = 0.12$). Moreover, as hypothesized in *H2*, this gap in trust varied with the level of transparency ($F_{2,97} = 3.48$, $p = 0.035$, $\eta_p^2 = 0.07$). There was no main effect of transparency ($F_{2,97} = 0.87$, $p = 0.42$).

In the low transparency condition, trust was lower when expectations were violated ($M_0 = 3.02$, $SD_0 = 1.34$, $M_1 = 4.18$, $SD_1 = 0.85$, $t_{37} = 3.15$, $p = 0.003$, $d = 1.01$). However, in the medium transparency condition, trust in both groups was similar ($M_0 = 3.72$, $SD_0 = 0.99$, $M_1 = 3.70$, $SD_1 = 0.87$, $t_{32} = 0.06$, $p = 0.95$). In the high transparency condition, trust was once again lower when expectations were violated ($M_0 = 2.77$, $SD_0 = 1.28$, $M_1 = 3.89$, $SD_1 = 0.77$, $t_{28} = 2.96$, $p = 0.006$, $d = 1.08$). This pattern is consistent with *H3*.

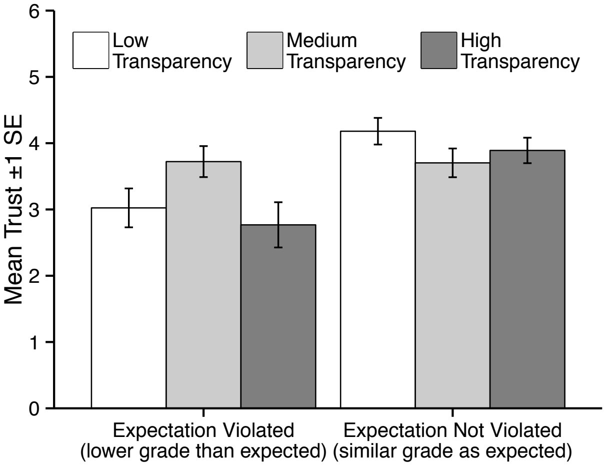

Figure 1. Trust examined as a function of expectation violation and randomly assigned transparency condition. Standard error bars are shown.

Instead of a binary cutoff, expectation violation can be measured continuously as the difference between the expected (self) and received (peer) grade. Figure 2 illustrates trust ratings against the degree of expectation violation in each transparency condition. In the low and high transparency conditions, expectation violation was negatively correlated with trust (low: $r = -0.59$, $t_{37} = 4.47$, $p < 0.001$; high: $r = -0.55$, $t_{28} = 3.45$, $p = 0.002$). However, consistent with *H3*, trust was uncorrelated with expectation violation in the medium transparency condition ($t_{32} = 0.26$, $p = 0.79$).

DISCUSSION

This study tested the effect of system transparency on user trust in the context of peer assessment in an online course. Trust is a critical issue in this setting, which involves high stakes, as course certification hinges on grades from peer assessment. The results provide strong evidence in support of the three hypotheses put forward: Expectation violation reduced trust overall (*H1*), but interface transparency moderated this effect (*H2*), such that providing some transparency with procedural information fostered trust, while additional information about outcomes nullified this effect (*H3*).

Consistent with a dual process model of attitude change, expectation violation was a critical moderator of the effect of transparency on trust. If users' expectations were met, interface transparency did not affect trust, as individuals were less likely to examine information thoroughly and more likely to rely on general impressions or their own mood [18]. The effects of transparency were only detected among individuals whose expectations were violated negatively and who would therefore be motivated to evaluate relevant information to understand the inconsistency and potentially change their

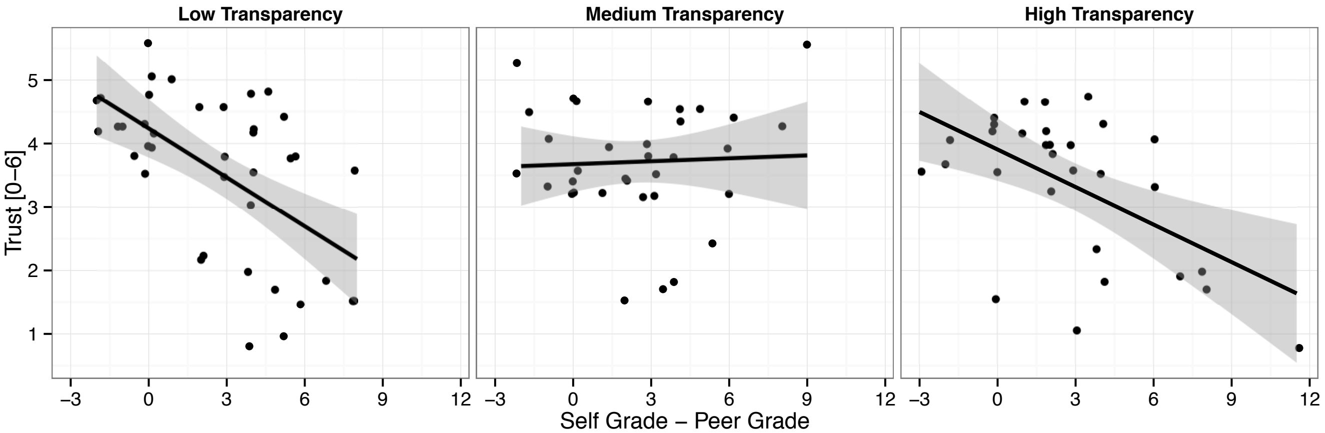

Figure 2. Trust examined as a function of the difference in expected (self) grade and received (peer) grade in each randomly assigned transparency condition. Points are jittered for presentation; OLS regression lines with 95% confidence bounds.

attitude.[1] This accounts for why interface transparency might influence some people's attitudes but not others. It may also explain mixed outcomes in prior empirical work on interface transparency, where expectation violation is rarely taken into account.

Transparency was manipulated in this study by providing different types of explanations. Consistent with procedural justice theory [15], a procedural explanation re-instilled trust in those whose expectations were violated. Notably, the explanation explicitly described the grade-adjustment algorithm as a fair method to reduce bias in grades—this may be necessary to elicit a positive response. Supplementing the procedural explanation with outcome-specific information to further increase transparency undermined the positive impact of procedural transparency. One reason for this result is that additional information was confusing and reduced understanding instead of opening the 'black box' (c.f. [20]). Ratings of system comprehension were in fact higher in the medium than high transparency condition. An alternative explanation is that additional information shifted the focus away from procedural justice and back to grading outcomes, re-emphasizing the perceived inconsistency und unfairness of those outcomes (c.f. [1]).[2] Effective applications of transparency in interface design can be informed by a deeper understanding of the mechanisms by which explanations shape user attitudes.

The current work has implications for theory on interface transparency and algorithm awareness. It demonstrates the critical role of user expectations in relation to system output and it provides initial evidence for a bell-shaped relation between transparency and trust. The practical implications of this work most immediately concern the design of online peer assessment systems, which should provide procedural transparency to avoid losing some learners' trust. More broadly, the results encourage engineers and designers to consider adaptive interface transparency in response to expectation violation—providing procedural information to confused users. While procedural transparency is most effective when expectations were violated, it may not matter to individuals whose expectations are met by the system. Still, providing the option to find out more about the system could build trust, help manage expectations, and preempt experiences of inconsistency. Limitations of the current work include the small sample size, the focus on self-report outcomes, and the absence of qualitative interviews to gain deeper insights into the user experience. Future work should replicate the results in different contexts, assess longitudinal behavioral outcomes, and investigate the proposed mechanisms.

The ongoing debate in the HCI community around system transparency [9], or seamless versus 'seamful' design, documents the importance and complexity of the issue. Interface design should foremost be tailored to its application context. In education, academic assessment is traditionally a highly opaque system. New forms of data-enriched assessment provide novel challenges and opportunities for transparent design in digital learning environments [30]. As with any new technology, its adoption and potential benefits depend on individual attitudes and public opinion. Intelligent technologies have been around for some time, but their intelligence is becoming increasingly difficult to hide—the intelligence behind the anti-lock breaking system is easier to hide than that behind a self-driving car, for instance. In an algorithmic interface, the right amount of system transparency—not too little and not too much—can foster positive attitudes and encourage people to reap the benefits of intelligent technology.

ACKNOWLEDGMENTS

I thank Clifford Nass for his inspiration and guidance, Chinmay Kulkarni for his help with the implementation, Daniel McFarland for providing access to his online course, and Emily Schneider and the CHI reviewers for their helpful feedback.

[1] Not enough data was available to study positive expectation violation, as few self grades underestimated the peer grades.

[2] To protect self-integrity, individuals may have attributed the perceived inconsistency between grades to a lack of comprehension, reflected in low ratings of system understanding [28].

REFERENCES

1. Frank Bannister and Regina Connolly. 2011. The Trouble with Transparency: A Critical Review of Openness in e-Government. *Policy & Internet* 3, 1 (2011), 1–30.
2. Bruce G Buchanan, Edward Hance Shortliffe, and others. 1984. *Rule-based expert systems*. Vol. 3. Addison-Wesley Reading, MA.
3. Judee K Burgoon and Jerold L Hale. 1988. Nonverbal expectancy violations: Model elaboration and application to immediacy behaviors. *Communications Monographs* 55, 1 (1988), 58–79.
4. Karen S Cook, Chris Snijders, Vincent Buskens, and Coye Cheshire. 2009. *eTrust: Forming relationships in the online world*. Russell Sage Foundation.
5. Cynthia L Corritore, Beverly Kracher, and Susan Wiedenbeck. 2003. On-line trust: concepts, evolving themes, a model. *International Journal of Human-Computer Studies* 58, 6 (2003), 737–758.
6. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (2008), 455–496.
7. Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in the news feed. In *Proceedings of the 33rd Annual SIGCHI Conference on Human Factors in Computing Systems*. 153–162.
8. Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. 2008. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*. ACM, 227–236.
9. Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. A path to understanding the effects of algorithm awareness. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 631–642.
10. Jonathan L Herlocker, Joseph A Konstan, and John Riedl. 2000. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM, 241–250.
11. Hilary Johnson and Peter Johnson. 1993. Explanation facilities and interactive systems. In *Proceedings of the 1st international conference on Intelligent user interfaces*. ACM, 159–166.
12. Julia H Kaufman and Christian D Schunn. 2011. Students' perceptions about peer assessment for writing: their origin and impact on revision work. *Instructional Science* 39, 3 (2011), 387–406.
13. Jürgen Koenemann and Nicholas J Belkin. 1996. A case for interaction: a study of interactive information retrieval behavior and effectiveness. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 205–212.
14. Chinmay Kulkarni, Koh Pang Wei, Huy Le, Daniel Chia, Kathryn Papadopoulos, Justin Cheng, Daphne Koller, and Scott R Klemmer. 2013. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 6 (2013), 33.
15. E Allan Lind and Tom R Tyler. 1988. *The Social Psychology of Procedural Justice*. Springer Science & Business Media.
16. Jack Muramatsu and Wanda Pratt. 2001. Transparent Queries: investigation users' mental models of search engines. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 217–224.
17. Stephen J Payne. 2003. Users' mental models: the very ideas. *HCI models, theories, and frameworks: Toward a multidisciplinary science* (2003), 135–156.
18. Richard E Petty and John T Cacioppo. 1986. *The elaboration likelihood model of persuasion*. Springer.
19. Chris Piech, Jon Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. 2013. Tuned Models of Peer Assessment in MOOCs. In *Proceedings of the 6th International Conference on Educational Data Mining*.
20. Wolter Pieters. 2011. Explanation and trust: what to tell the user in security and AI? *Ethics and information technology* 13, 1 (2011), 53–64.
21. Pearl Pu and Li Chen. 2007. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20, 6 (2007), 542–556.
22. Emilee Rader and Rebecca Gray. 2015. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 173–182.
23. Byron Reeves and Clifford Nass. 1996. *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.
24. Jens Riegelsberger, M Angela Sasse, and John D McCarthy. 2005. The mechanics of trust: A framework for research and design. *International Journal of Human-Computer Studies* 62, 3 (2005), 381–422.
25. Philip M Sadler and Eddie Good. 2006. The impact of self-and peer-grading on student learning. *Educational assessment* 11, 1 (2006), 1–31.
26. Kay Sambell, Liz McDowell, and Sally Brown. 1997. "But is it fair?": an exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation* 23, 4 (1997), 349–371.

27. Mien Segers and Filip Dochy. 2001. New assessment forms in problem-based learning: the value-added of the students' perspective. *Studies in higher education* 26, 3 (2001), 327–343.

28. David K Sherman and Geoffrey L Cohen. 2002. Accepting threatening information: Self–Affirmation and the reduction of defensive biases. *Current Directions in Psychological Science* 11, 4 (2002), 119–123.

29. Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*. ACM, 830–831.

30. Candace Thille, Emily Schneider, René F Kizilcec, Christopher Piech, Sherif A Halawa, and Daniel K Greene. 2014. The future of data-enriched assessment. *Research & Practice in Assessment* 9, 2 (2014), 5–16.

31. Weiquan Wang and Izak Benbasat. 2007. Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems* 23, 4 (2007), 217–246.

32. Meichun Lydia Wen and Chin-Chung Tsai. 2006. University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education* 51, 1 (2006), 27–44.