What you think and what I think: Studying intersubjectivity in knowledge artifacts evaluation

By: Dmytro Babik, Rahul Singh, Xia Zhao, Eric W. Ford

Babik, Dmytro, Rahul Singh, Xia Zhao and Eric W. Ford. "What You Think and What I Think? Studying Intersubjectivity in Knowledge Artifacts Evaluation," *Information Systems Frontiers*, 2015, 1-26.

The final publication is available at Springer via http://dx.doi.org/10.1007/s10796-015-9586-x

Abstract:

Miscalibration, the failure to accurately evaluate one's own work relative to others' evaluation, is a common concern in social systems of knowledge creation where participants act as both creators and evaluators. Theories of social norming hold that individual's self-evaluation miscalibration diminishes over multiple iterations of creator-evaluator interactions and shared understanding emerges. This paper explores intersubjectivity and the longitudinal dynamics of miscalibration between creators' and evaluators' assessments in IT-enabled social knowledge creation and refinement systems. Using Latent Growth Modeling, we investigated dynamics of creator's assessments of their own knowledge artifacts compared to peer evaluators' to determine whether miscalibration attenuates over multiple interactions. Contrary to theory, we found that creator's self-assessment miscalibration does not attenuate over repeated interactions. Moreover, depending on the degree of difference, we found self-assessment miscalibration to amplify over time with knowledge artifact creators' diverging farther from their peers' collective opinion. Deeper analysis found no significant evidence of the influence of bias and controversy on miscalibration. Therefore, relying on social norming to correct miscalibration in knowledge creation environments (e.g., social media interactions) may not function as expected.

Keywords: Intersubjectivity | Miscalibration | Longitudinal analysis | Knowledge artifacts | Peer-evaluation | Latent classes

Article:

Introduction

Rapid advances in Web 2.0-based social media technologies have made creating and distributing digital knowledge artifacts (KAs) easier, more accessible, and less expensive. As a result, online collaboration to produce, refine and manage new knowledge is increasing in popularity (Sage and Rouse 1999). For example, digitally published articles outlining current knowledge about a topic, are often co-created, co-evaluated and maintained by self-organized knowledge communities in IT-mediated social media (Dede 2008). KAs are creations that social groups use to represent and share knowledge. In organizations, for example, KAs are integrated as part of organizational knowledge management systems (KMS) to retain organizational memories. In social media, wikis, blogs and other public shared knowledge bases, such as eHow.com and About.com, enable quick access of users from any walks of life to rich and yet easily comprehendible information. Our understanding of KAs is based on the notion of cognitive artifacts – things that help us understand and perform tasks (Heersmink 2013; Norman 1992). A cognitive artifact is an artificial device or object designed to display or operate on information to serve a representational function (Norman 1992). Sharable and transferable representations of knowledge such as cognitive artifacts have multiple distributed cognition benefits and provide structured shareable referents to coordinate thought or a process in a variety of domains (Kamsu-Foguem et al. 2014; Kirsh 2010; Sutton 2001; Tung et al. 2014. KAs facilitate information sharing in knowledge communities where KA creators, evaluators and users collaboratively learn and evolve the collective understanding of a topic (Salazar-Torres et al. 2008). As new, more dynamic and participatory IT-enabled modes of information exchange emerge, understanding how KAs are created and evaluated in social systems, and how differences between creators' and evaluators' opinions regarding KA evolve over time has very important implications for the future information systems research.

Anyone with Internet access can produce a digital KA or modify an existing one and share it with the world instantaneously. However, users' or reviewers' perceptions and opinions about the quality, utility and verity of KAs, captured in evaluations, may vary significantly (Black and Wiliam 1998). In knowledge communities, KAs emerge through social interactions between creators and evaluators. These interactions and evaluation are subjective in nature. All knowledge codification and transfer incorporates subjective elements and greatly depends upon individual interpretations (Sutton 2001). As a result, the facts, opinions, values and beliefs around knowledge captured by KAs may or may not be shared by all community members. Over time, subjective preferences impact how KAs' quality and utility are perceived by both creators and evaluators. Unexplained variances in user's perceptions raise questions about the reliability and validity of the evaluation of socially produced KAs.

Much research has focused on socially constructed knowledge in technology-mediated, web-based communities (Brown et al. 1989; Cusinato et al. 2009; Dede 2008; Hu et al. 2007; Lauw et al. 2008; Miranda and Saunders 2003). Despite wide-spread societal acceptance of, and reliance on digital KAs, very little research examines the intersubjective nature of KA creation and evaluation (Walsham 2006) and organizational social learning (Wang 2011) in IT-enabled peer-based environments. Without research to understand the inter-twined and intersubjective processes involved in creating and evaluating KAs, effective designs of information systems that facilitate creation of reliable and valid KAs remain underinformed.

The purpose of this paper is to present a systematic investigation of the inter-twined and intersubjective social creation and evaluation processes that online knowledge communities use to create KAs. In particular, miscalibration is explored, which, for the purpose of our study, refers to the difference in perceptions of a KA's quality between the creator and peer evaluators (Kruger and Dunning 2002; Sadler and Good 2006; Sargeant et al. 2008. The impact of repeating iterations of multi-peer evaluation and feedback as a mechanism to improve creators' self-assessment competency through ongoing social interactions is examined. Lauw et al. (2008) suggested subjectivity of evaluations is reflected in evaluators' bias (individual evaluators' deviation from opinions of other evaluators) and controversy of KAs (the level of disagreement of evaluators' opinions about a particular object). Therefore, we also examine the interactions between miscalibration and these important concepts of the evaluation process.

In this paper, we address the research question: How does miscalibration of self-perceptions with the perceptions of others in a knowledge community changes over multiple iterations in the presence of bias and controversy?

We study the longitudinal dynamics of miscalibration between creators' and peer evaluators' perceptions regarding the KA, under different levels of bias and controversy, as members engage in multiple iterations of creating and evaluating KA. Using a sample of 435 students at a large public university, we explore changes in students' peer assessment, self-assessment, miscalibration metrics, controversy and bias over time. We tested our hypotheses using standard inferential statistics (t-tests). The major methodological advance that we propose is to apply Latent Growth Modeling (LGM) to study the longitudinal dynamics of miscalibration between creator's and peers' evaluation of KA. The nature of intersubjectivity, as many other social phenomena and theories explored by the IS research, is not static but rather intrinsically longitudinal. Yet, longitudinal analyses, and LGM in particular, have been sparsely used by IS researchers Zheng et al. (2014). Our study provides a systematic method to evaluate and understand the longitudinal development of expertise and KAs in IT-enabled knowledge communities and contributes to the currently sparse research of temporal evolution of such communities.

Understanding creator-evaluator interactions around KAs is important for three audiences. In the business community, technology-aided knowledge management is a critical factor for sustained competitive advantage. Being able to statistically assess employees' shared belief and knowledge systems will be an important advance. For those engaging in open online social KA creation (e.g., eHow.com, wikiHow.com, about.com, Pinterest, Yelp, etc.), the ability to capture actionable reliability, validity and utility metrics for KAs and creators will improve the efficacy of these knowledge communities. In education, the shift toward Massive Open Online Courses (MOOCs) and other pedagogical approaches that rely heavily on peer learning, requires educators to be able to identify and influence intersubjective agreement or disagreement around KA evaluations.

The paper proceeds as follows. In Section 2, we give overview of the relevant prior research, develop our hypotheses and present the research model. In Section 3, we describe our research methodology. Section 6 reports the results of hypotheses testing. Finally, we discuss our findings and present the conclusions, limitations and areas of future research in Section 7.

Literature review

Knowledge communities and intersubjectivity

Knowledge communities are social systems of knowledge creation and exchange between members situated in the context of the particular domain of interest (Amin and Roberts 2008; Brown et al. 1989; Edwards 2001; Sutton 2001). In such social system, creators and users of equal status, or peers, interact through communicative actions of providing peer evaluation and feedback intended to achieve rational cooperative or conflicting objectives and advance perspectives, while collectively searching for shared meaning, clarification and agreement (Habermas 1981; Hermida 2011). Common meaning, intersubjectively derived through exchange of multiple perspectives in the form of qualitative and quantitative evaluations and feedback results in better comprehension of KAs representing problems and solutions by actors (Miranda and Saunders 2003; Walsham 2006). Intersubjectivity can be viewed as the shared understanding that emerges from socio-technical interactions of individual in the pursuit of the common goal (Bostrom and Heinen 1977; Markus and Robey 1988; Miranda and Saunders 2003).

At the same time, individual subjectivity poses a threat to the objective evaluation of KA produced in knowledge communities. Knowledge encoded in a KA means different things to different people depending on their backgrounds, positions and social context. When dealing with an artifact developed as a result of a simple task, with a single straightforward outcome exists and non-conflicting, objective goodness criteria (Zigurs and Buckland 1998), attainment of KA's goodness can be assessed by comparing the outcome to this set of objective criteria. Most KA, however, result from complex open-ended tasks or problems with attributes such as multiple acceptable outcomes, multiple solution schemes, conflicting interdependence, and outcome uncertainty (Campbell 1988). In different literature streams such problems are known as "ill-structured" (Simon 1969, 1973), "wicked" (Rittel and Webber 1973), or "design" problems (Conklin 2001). Consequently, they typically involve creativity, subjectivity and ambiguity about KA's goodness. Since such problems are multifaceted, and facets are difficult to measure objectively, evaluations suffer from bounded rationality (Dorst 2003; Kreps 1997; Simon 1959). Evaluators' judgments regarding the KA depend on their subjective understanding of the content and the context. Expertise is often focused and limited, whereas the complexity of tasks and corresponding KAs is virtually unlimited. Moreover, every subjective evaluation is affected by evaluators' knowledge of the KA and the domain, as well as their individual perspective towards them including any individual biases regarding the relevant subject matter, content and context (Matusov 1996; Walsham 2006). One of the common ways to alleviate subjectivity of evaluations of KAs built around complex tasks is subject to multiple peer reviews and evaluations (Hardaway and Scamell 2012). A variety of models have been proposed in different settings, which are largely based on the belief that a collective evaluation, or the wisdom of crowd, is more objective than a single evaluator's subjective, or biased assessment (Surowiecki 2004). Thus, goodness, or attainment, of a KA becomes intersubjective. It describes understanding that emerges from the shared experiences (Schutz 1967), and is determined by subjective states shared by multiple individuals (Scheff 2006) and, thus, by the intersubjective interpretation by multiple creators and evaluators (Dorst 2003; Miranda and Saunders 2003; Walsham 2006). Intersubjectivity emphasizes that shared cognition and consensus is essential in the shaping of ideas and relations.

Peer evaluation

KA evaluation based on multiple peer evaluators' perspectives and opinions presents its own challenges. While in some instances it may result in intersubjective consensus regarding KA goodness and produce constructive and complete recommendations for its improvement, in other it may lead to contradicting conclusions or prescribe conflicting directions, leaving the creator perplexed regarding desired properties of KA. Yet, since multi-peer evaluation is believed to be a more valid and reliable alternative to KA quality assessment than a single-evaluator's opinion, it is worthwhile to investigate its outcomes and how it may be employed to improve KA creation and refinement process in knowledge communities. Since the 1990s, peer evaluation and its impact on learning process and outcome has been extensively studied by social science and education research (Topping 2005). Defined by Topping (1998) as "an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar status", peer evaluation usually carries some combination of formative and summative assessment. Summative assessment seeks to monitor performance by providing a quantitative summary evaluation of the attainment of a particular task objective (Shepard 2007). It is typically used for external accountability and is expressed in the form of a score or grade. Formative assessment involves qualitative feedback with the aim to promote improvement by suggesting adjustments and modifications to the artifact (Crooks 2001; Huhta 2008).

Despite the large volume of research, conditions for efficacious peer evaluation remain inconclusive. When actors interact with their peers in learning or problem-solving situations, intersubjective disequilibration occurs, inconsistent knowledge is exposed, opposing perceptions and ideas are explored, and inadequate logical reasoning and strategies are challenged (Piaget and Gabain 1926; Slavin 1992; Yu et al. 2005. As a result, peer evaluation helps build cognitive ability because it facilitates peer interaction, exchange and absorption of critical concepts (King 1989). While some studies highlighted the importance of peer evaluation in social and self-regulated learning systems to achieve competency gains, other noted that peer group 'value diversity' negatively impacts collaborative learning's efficacy by inhibiting the formation of a shared perspective and understanding (Van Gennip et al. 2009, Van Gennip et al. 2010).

Peer evaluation improves competencies of actors who assess and are being assessed by exposing them to the practice of evaluating others' KAs and receiving feedback on their own KAs (Brutus and Donia 2010). Peer feedback enhances individuals' meta-cognitive learning and critical thinking skills (Wang and Wu 2008) and enables learning at high cognitive levels (Bouzidi and Jaillet 2009). Helping others to improve their creations by giving feedback, as well as improving one's own creations based on feedback received from peers is a competency that is acquired through practice (Sluijsmans et al. 2004). Providing, receiving and incorporating constructive summative and formative feedback to improve KA attainment, as well as creators' and evaluators' competencies, are important aspects of peer interactions that extend beyond the classroom into technology-enabled knowledge communities. However, translating social interactions of KA peer evaluations into effective KA creation and refinement requires a great degree of self-awareness and maturity of peer reviewer and feedback recipient.

Self-evaluation: Challenges in calibration

Just as peer evaluation, self-evaluation is a complex social activity. In particular, assessing one's own competencies relative to those of other knowledge community members through the process of self-reflection is a critical thinking skill Lin et al. (2001)In the context of KA creation and refinement, self-reflection, self-evaluation, and self-regulation play dual roles: they stimulate creator's motivation and creativity to produce and refine new artifacts and they guide the creator to be responsive to external feedback and evaluations. These cognitive activities are intrinsic to professional behavior and creative pursuits. Accurate self-evaluation results in greater satisfaction with the accomplished results and stimulates aspiration to reach new goals (Bandura 1977).

Accurate self-evaluation is difficult to achieve. Previous theoretical and empirical studies showed mixed results when comparing self-evaluation to peer evaluation. While some evidence suggested that self-evaluation may be as accurate (or even more accurate) as external assessment Dochy et al. (1999); Lindblom-ylänne et al. 2006), other studies found that self-evaluation can be significantly miscalibrated, i.e., creators tend to inflate their own self-assessments compared to evaluations of their work by peers (Falchikov 1986; Falchikov and Boud 1989; Ryvkin et al. 2012; Sargeant et al. 2008). Results on whether specific criteria and good guidance improve self-evaluation accuracy are also mixed (Buchy and Quinlan 2000; Lindblom-ylänne et al. 2006; Orsmond et al. 2000; Sluijsmans et al. 2002; Taras 2002).

The behavioral economics literature, specifically the research on the "unskilled-and-unaware" problem by Kruger and Dunning (1999), indicate that subjects with lower task competency (the "unskilled") tend to overestimate their performance, thus showing overconfidence. On the other hand, individuals with higher competency levels (the "skilled") typically underestimate their performance compared to peers' efforts, showing underconfidence (Kruger and Dunning 1999; Ryvkin et al. 2012). Further, this miscalibration is not normally distributed because there are more "unskilled" members that overestimate their own performance than "skilled" who underestimate theirs. According to Kruger and Dunning (1999), the unskilled lack the metacognitive ability to realize their incompetence. In effect, they are afflicted by a "double curse" of low skill and low ability to recognize competence when presented. Moreover, when assessing other artifacts' relative values, unskilled peers may introduce into the intersubjective group dynamics biases, which may compromise the reliability of evaluations. In the absence of objective criteria for evaluating complex-problem KAs, where reliability serves as a source of validity of KA assessment (Uebersax 1988), this presents a serious issue for peer evaluation in knowledge management systems. The theory of social modeling suggests that self-assessment of KA would tend to converge toward peer assessment as social interactions between members of the knowledge community ensue (Bandura 1962). A number of studies addressed the issue of reducing miscalibration (for example, creator's overconfidence) through experience of multiple iterations and feedback (Ryvkin et al. 2012). While some studies demonstrated this effect (e.g., Koriat et al. 1980; Lichtenstein and Fischhoff 1980; McKenzie 1997; Sieck and Arkes 2005; Sieck et al. 2007; Stone and Opel 2000), other found that miscalibration is robust with respect to feedback (e.g., Pulford and Colman 1997; Sharp et al. 1988. The present study contributes to this discourse by investigating the discrepancy between creator's self-perception of the artifact goodness and peer evaluators' perceptions in the specific context of a technology-enabled peer-based knowledge creation and evaluation environment.

Controversy and bias

Given the inevitable dependency of perceived KA attainment on the variation in subjective peer evaluations, two other important aspect of intersubjectivity that must be considered are controversy of a KA and evaluator bias (Gillespie and Cornish 2010; Lauw et al. 2008). Even if most evaluators in the community are subjectively fair, some of them may have idiosyncratic preferences or opinions that will distort consensus (Douceur 2009; Shah et al. 2013. Moreover, since not all KAs are evaluated by all community members (due to physical constraints), aggregating KA evaluations also requires consideration of systematic biases of individual evaluators, such as their “confidence” or “reputation”.

Different KAs may vary in the variety and divergence of opinions they generate. Therefore, some KAs may be more “controversial” than the other. Controversy does not necessarily imply the lack of “goodness”; it is a concept orthogonal to quality; i.e., a controversial KA may have high or low goodness attainment depending on who evaluates it (Lauw et al. 2008). In the domain of online peer-based evaluation systems, a number of models have been recently proposed to account for impacts of individual evaluators' idiosyncrasies, their impact on intersubjective assessment of evaluated objects, implications for individual reputation and overall consensus (Cusinato et al. 2009; Dai et al. 2012; Lauw et al. 2006; Lauw et al. 2008; Mizzaro 2003; Roos et al. 2012).

Cognitive biases of individual actors influence their perceptions about their own and their peers' work (in the form of KAs) and, therefore, affect how intersubjective agreement or disagreement about KAs' goodness is reached. As a result, whenever individual perceptions of evaluators about a particular KA align, less dissonance in evaluations is observed, whereas when perceptions conflict, we observe more variation in evaluations and, hence, treat a KA as more controversial. Lauw et al. (2008), in particular, introduced and discussed operationalization of the notions of controversy and bias in evaluation systems, which are briefly summarized here. Controversy and bias are inter-related measures of the departures among evaluations of an object, such as a KA. Controversy captures a degree of divergence of evaluations among reviewers. In other words, controversy reflects the lack of agreement among peers about a specific KA’s goodness attainment. Evaluator bias refers to a degree of deviation of the particular evaluator’s assessments from other evaluators. Stated simpler, bias reflect how much the actor’s opinion, in general, is different from the opinions of others.

Applying the concepts of controversy and bias to modeling intersubjectivity of evaluations in knowledge communities serves three purposes:

(a) to differentiate actor’s competencies into creation and evaluation competencies by separating the impact of individual evaluation competency on overall evaluation (in bias) and creator’s propensity to produce KA’s with higher or lower degree of evaluation consensus (in controversy);
(b) to assess reliability of subjective evaluations of artifacts by segregating KAs with higher consensus (lower controversy) and lower consensus (higher controversy); and
(c) to examine the relationship between actor’s creation and evaluation competencies and self-evaluation by relating bias and controversy to miscalibration.

Therefore, in our approach, we differentiate between bias and controversy as cognitive traits (which are outside the scope of this paper) and bias and controversy as operationalization of aspects of intersubjective evaluation.

Conceptual model and hypotheses

The reviewed literature shows that it remains ambiguous whether peer evaluation interactions among members of a knowledge community always lead to better intersubjective shared evaluation of KAs' goodness (Topping 1998). Moreover, this shared understanding can be viewed in multiple ways, reflecting its complex and subjective nature. Thus, to conceptualize and operationalize the research question motivating this study, we present a temporal model of intersubjectivity of peer evaluations of KAs.

For the purposes of this paper, we conceptualize KA evaluation in a knowledge community as a dynamic system of several concepts that describe the interrelationships among individual valuations given to KAs by their creators and multiple peer evaluators. We define attainment as the reflection of the goodness of KA in the view of evaluators. We define miscalibration as the difference between the creator’s self-perception of the artifact attainment and the aggregate peer evaluation perceptions about it. We adopt and apply the definitions of bias by Lauw et al. (2008) to describe deviations of individual peer evaluation from other peer evaluations. Similarly, we adopt their definition of controversy to describe the overall aggregate divergence of peer evaluations of an individual KA. We further conceptualize KA evaluation subjectivity of an actor as a combination of miscalibration with respect to her own KA and evaluator bias with respect to KAs created by other actors in the knowledge community. Accordingly, intersubjective assessment is conceptualized as a combination of attainment of a KA and its controversy. Together, attainment, controversy, miscalibration, and bias reflect a snapshot, at a particular time, of intersubjective understanding of the KA goodness in the knowledge community stakeholders in the particular KA (Fig.1).

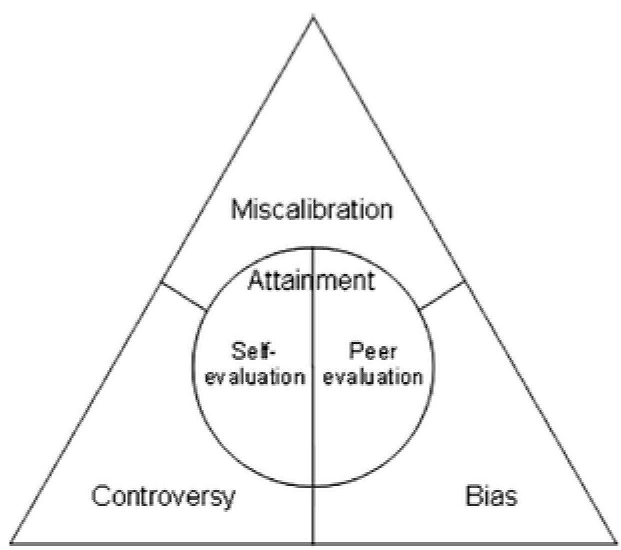

Figure 1. A Single-instance intersubjective KA evaluation

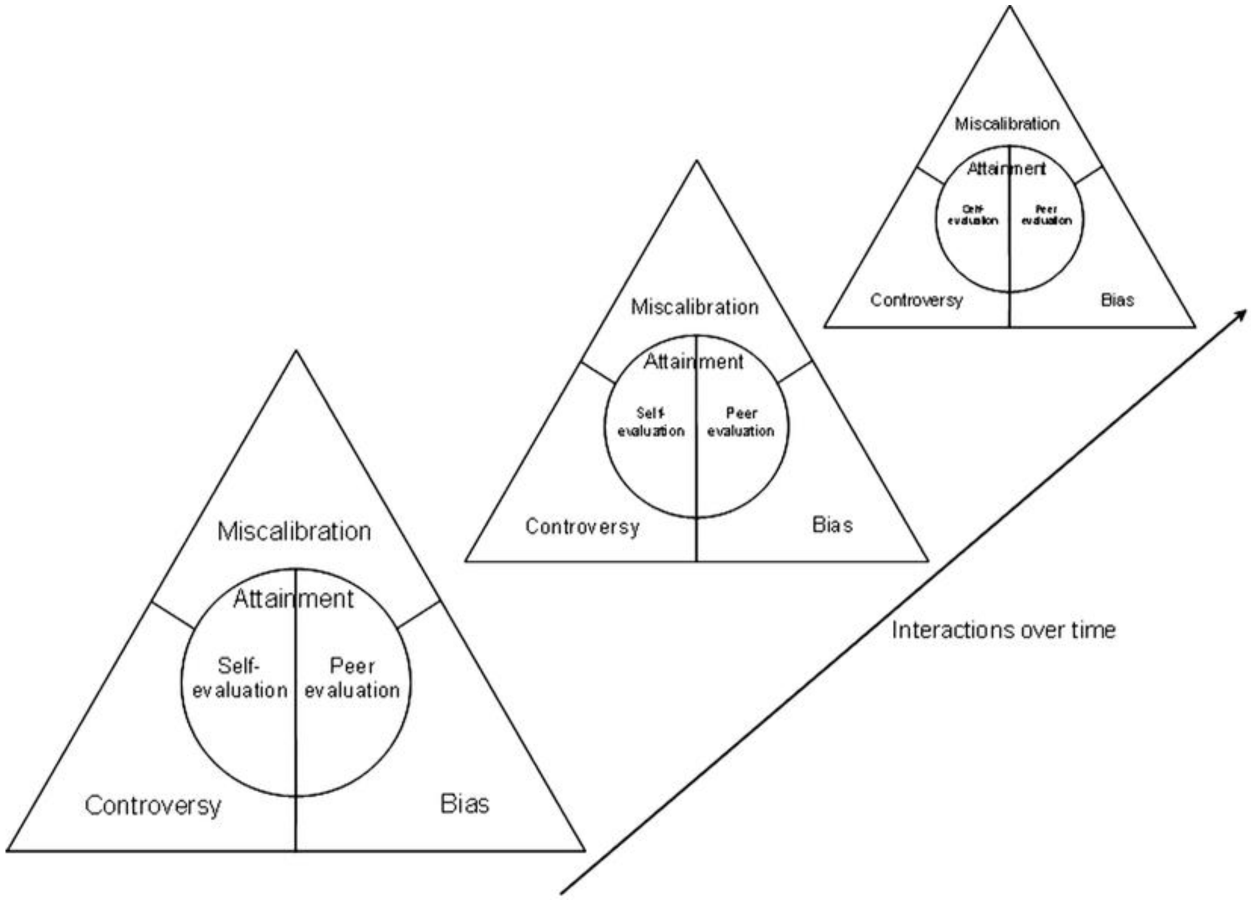

Figure 2. Conceptual model of the change of intersubjective understanding of attainment over time

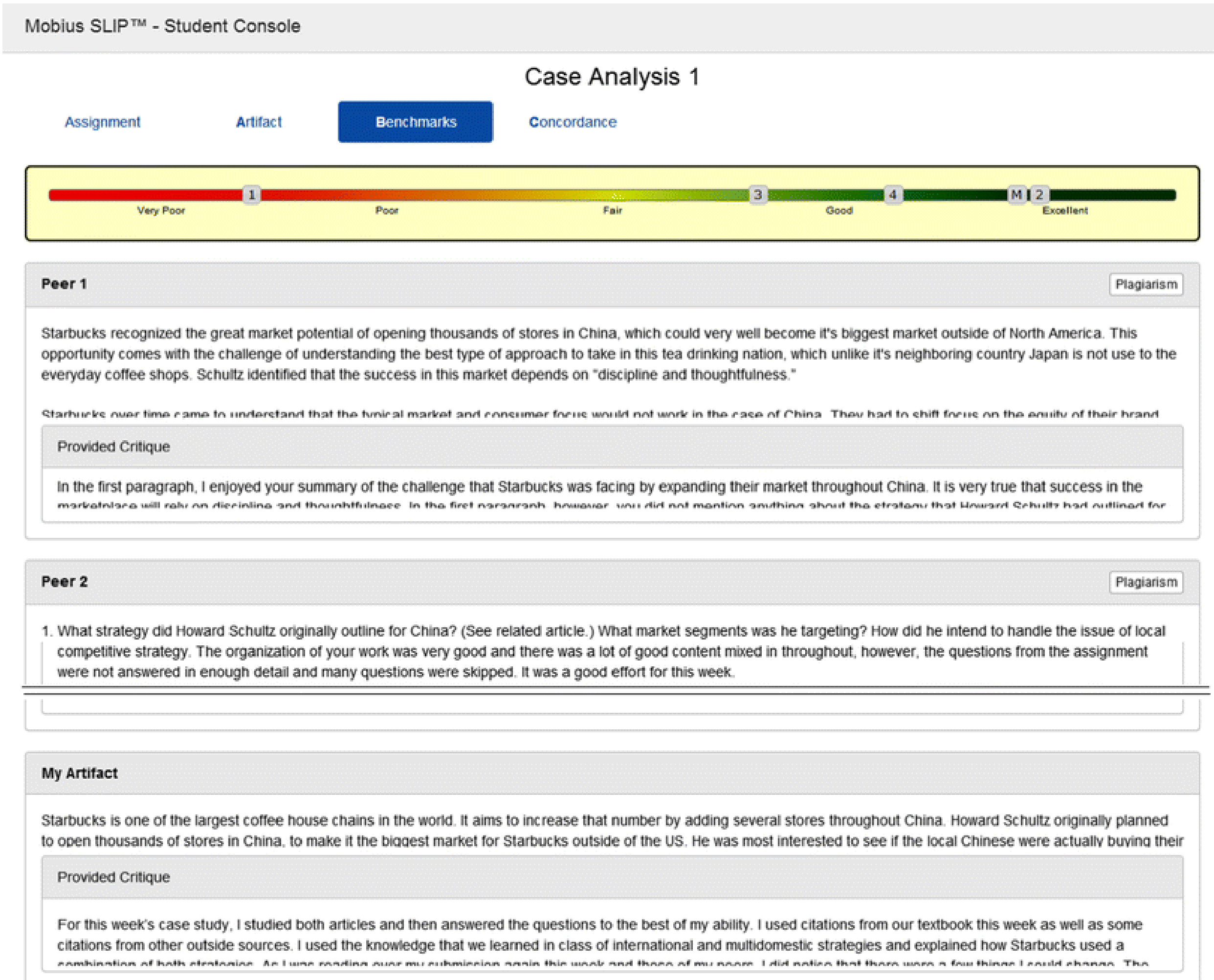

Figure 3. The SLIP benchmarking interface used to collect assessments of case analyses

Actor		Peer Evaluation (Inverted Ranks Given)				
		I	II	III	IV	V
Artifacts (Inverted Ranks Received)	I		1	1	1	1
	II	1		2	2	2
	III	2	2		3	3
	IV	3	3	3		4
	V	4	4	4	4	

Figure 4. Example scenario of mutual peer evaluation ranking

Actor		Peer Evaluation (Inverted Ranks Given)				
		I	II	III	IV	V
Artifacts (Inverted Ranks Received)	I	*4*	1	1	1	1
	II	1	*2*	2	2	2
	III	2	2	*3*	3	3
	IV	3	3	3	*3*	4
	V	4	4	4	4	*4*

Figure 5. Mutual peer evaluation and self-assessment ranking

Actor		Peer Evaluation (Inverted Ranks Given)				
		I	II	III	IV	V
Artifacts (Inverted Ranks Received)	I		1	1	4	4
	II	4		3	2	1
	III	1	2		1	2
	IV	2	3	2		3
	V	3	4	4	3	

Figure 6. Controversy in mutual peer evaluation

Actor		Peer Evaluation (Inverted Ranks Given)				
		I	II	III	IV	V
Artifacts (Inverted Ranks Received)	I		1	1	1	4
	II	1		2	2	3
	III	2	2		3	2
	IV	3	3	3		1
	V	4	4	4	4	

Figure 7. Bias in mutual peer evaluation

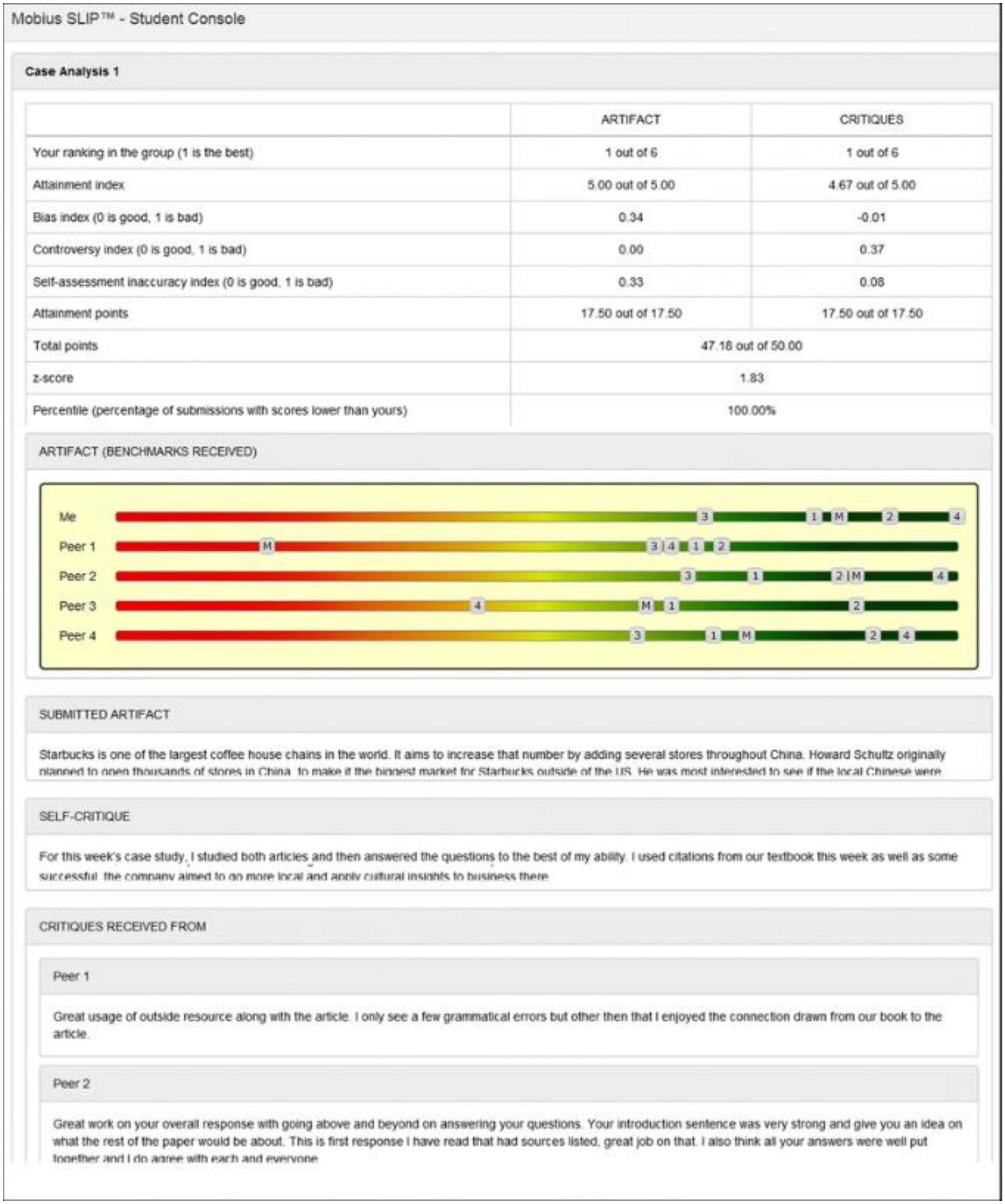

Figure 8. The student interface showing self- and peer assessment of case analyses

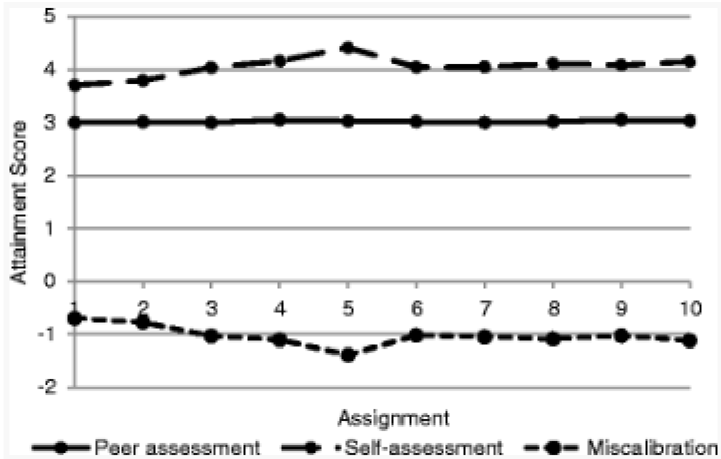

Figure 9. Changes in peer and self-assessment and miscalibration

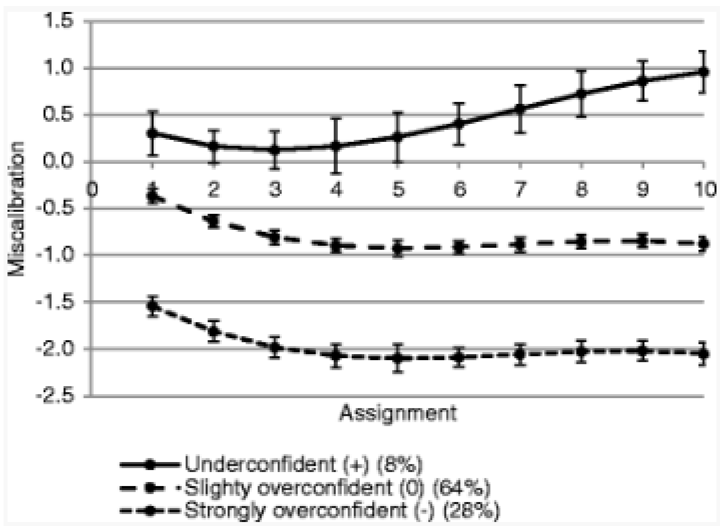

Figure 10. Latent classes of the miscalibration (with indicated standard errors)

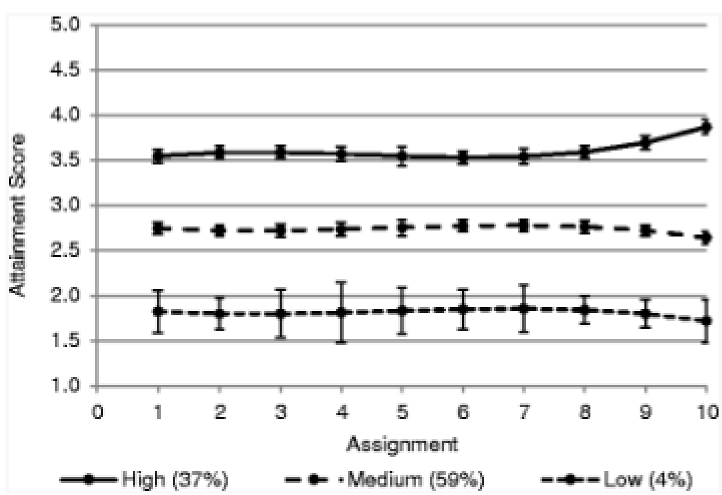

Figure 11. Latent classes of peer-evaluated attainment (with indicated standard errors)

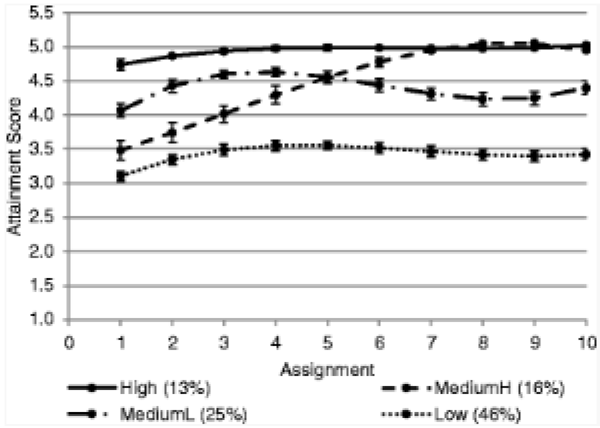

Figure 12. Latent classes of self-assessed attainment (with indicated standard errors)

Since we are interested in exploring dynamics of this system over time, the following conceptual model describes the longitudinal view of intersubjectivity in a knowledge community (Fig. 2).

Existence of systematic miscalibration

Based on the theory that suggests that external assessment and self-evaluations do not necessarily align, we first of all are interested in confirming that difference exists in evaluations of KAs. Therefore, we hypothesize that:

- H1a: KA attainment based on creator's self-evaluation is systematically different from attainment based on aggregate peer evaluations.
- H1b: Attainment based on creator's self-evaluations is systematically higher than attainment based on aggregate peer evaluations (i.e., creators are systematically overconfident).

Systematic longitudinal change of miscalibration

Social psychology, including literature on information processing and collective intelligence (Conklin 2001; Simon 1979), collective action (Hargrave and Van de Ven 2006), social cognitive theory (Bandura 1962; Bandura and Walters 1963; Gagne 1985), social construction of meaning (Miranda and Saunders 2003), knowledge as social practice (Brown and Duguid 2001), and intersubjectivity (Gillespie and Cornish 2010; Matusov 1996) suggests that over time, knowledge community members – creators and evaluators – learn from each other, develop shared understanding and individual competencies.

Through the experience of observing and evaluating the work of others, as well as being evaluated and receiving peer feedback, they realize implicit expectations with regard to attainment, as well as reconsider and adjust their actions to avoid being penalized for producing more controversial artifacts and reviews and to conform to social norms (Kreps 1997). Repeated experience in evaluating others and oneself and being evaluated results in subjects' better understanding of what is expected and, hence, in improved quality of artifacts and reviews, and more accurate evaluations (Brutus et al. 2013. Divergent opinions find settlement, shared understanding emerges. As subjects become more confident in evaluating others' KAs, they also become more critical of their own creations, and, therefore, self-evaluations become more reflective of the KA attainment perceived by peers. Based on this theory, we expect that multiple sequential mutual assessments create intersubjective dynamics that lead to the convergence between peer and self-evacuations in a group. Thus, over time they lead to the reduction of miscalibration in KA evaluations.

On the other hand, while it is possible, that mutual peer feedback leads to better common understanding, and, therefore, to the reduction of miscalibration (that is, both overconfidence and underconfidence should decrease), the double-curse argument of the Unskilled-and-unaware theory suggests that miscalibration may also systematically increase. Thus, past research suggests that although over time (i.e., multiple iterations) miscalibration may changes systematically, it provided no conclusive suggestions about the direction of these systematic changes in miscalibration. Therefore, we hypothesize that:

- H2: Over multiple successive mutual creator-evaluator interactions, miscalibration changes systematically (i.e., non-randomly).

Moreover, we anticipate that the intersubjective dynamics among subjects with homogeneous skills should be different from those in a population with substantially diverse skill levels.

Impact of controversy

Theory also suggests that subjectivity and intersubjectivity of perceived goodness of an evaluated KA and, consequently, its attainment evaluations, are influenced by the degree of evaluators' disagreement and manifest themselves through evaluator biases and KA controversy (Lauw et al. 2008). Intuitively, since higher controversy means a higher degree of divergence of assessments among evaluators and therefore implies a significant portion of evaluations in the mix that may not align with creator's own perception of her KA goodness captures, and since controversy may or may not be evident to the creator, we would expect creator's self-evaluation to diverge from the aggregate peer evaluation. Therefore, we hypothesize that:

- H3: Actors, who as creators produce KAs with higher controversy levels, show higher miscalibration than peers, who produce KAs with lower controversy levels.

Impact of bias

Finally, due to variation in backgrounds, competence, skills and perspectives among evaluators, and because expertise of any given reviewer is limited and focused, a reviewer's evaluations may systematically deviate from those of other peers. That is, the evaluator's judgments towards KAs in a specific topic domain may be biased compared to the peers' aggregate assessment. Since bias reflects higher degrees of divergence of an individual's evaluations from evaluations by other peers, we would expect that a creator, who tends to diverge from others on evaluations of others' KAs, will also diverge from others on evaluation of her own work. Therefore, we hypothesize that:

- H4: Actors, who as evaluators show higher bias level, also show higher level of miscalibration than peers with lower evaluator bias levels.

Research methodology

Participants

Participants were 435 undergraduate students at a large public university taking a sophomore level business course on the Principles of Management. The course was taught in spring 2013 in two face-to-face sections. The entire student body taking the course participated in this study. Students engaged in ten take-home assignments designed to improve analytical writing, critical thinking and information evaluation skills. The dropout rate of the students during the course was negligible (0.6 %), i.e., practically all students participated in the experiment throughout the entire course. The overall response rate was about 90 % for the case analyses submissions and 88 % for evaluations. 3617 usable records were obtained across ten assignments. For assignment 5, due to a university holiday, multiple students did not turn in their submissions and the peer review group size significantly deviated from the intended. Therefore, the data from this assignment was excluded from our analysis. Students completed these assignments as part of their course work and were not offered any monetary or social incentives. The score resulting from peer evaluation was included as a component of their grade. Age, gender or other demographic data were not collected.

System design, protocol and procedure

To operationalize and conduct our study, we used the online Social Learning Interaction Platform (SLIP) system developed by researchers at a large public university and designed to facilitate and monitor students' interactions in multiple online evaluations of each other's submissions in randomly assigned and double-blind peer groups (Ford and Babik 2013). In the SLIP system, interactions between subjects are executed according to the following protocol. Before each assignment, subjects are randomly allocated into peer review groups of, typically, five or six individuals. This group size is chosen for the following reasons: on the one hand, four or five peer evaluations give better grounds for statistical reliability of evaluations than two or three; on the other hand, evaluating (and especially ranking) more than four or five submissions causes substantially higher cognitive load, and, hence, is less accurate (Alwin and Krosnick 1985; Miller 1956). The group size of four to six peers is also advocated in other online peer review systems Cho et al. (2008); Joordens et al. 2009 (while the specific choice of the peer group size may affect the results of evaluation, we leave this issue outside the scope of this study, and for its purpose keep the group size constant). To eliminate biases due to non-anonymity, such as "friendship" or "retaliation" grading, all reviews and evaluations are double-blind (i.e., the identities of the reviewers are not revealed to the recipients of reviews and vice-versa at any time). Students in each peer group work individually and independently on a single common task requiring two electronic submissions in the SLIP – case analysis write-up (the "Artifact"), and peer and self-assessment (the "Benchmarks and Critiques"). Thus, in each group, all students act as creators as well as mutual evaluators of each other's KA.

For the Artifact, each student writes a case analysis report (an essay up to 800 words in length) and submits it in the SLIP. The topic of each assignment is the same for all students. While topics of the cases vary across assignments, the substantive task and the evaluation rubrics of the assignments remain the same over all ten iterations. Thus, we can claim that the difficulty of the assignment remained constant. After the Artifact submissions are collected, they are distributed anonymously among the peers in the group for summative and formative evaluation. Every student in the group has to review, evaluate and critique every other peer's Artifact, as well as her own Artifact.

Measures

Generally, summative evaluation can be conducted using either of two types of scales – ranking or rating. Rating refers to the comparison qualities of different objects using a common absolute, or cardinal, scale. Ranking, sometimes called forced-distribution rating, means comparing different objects directly one to another on a relative, or ordinal, scale (Schleicher et al. 2008). Both ranking and rating have their strengths and weaknesses, and have been used in research of social phenomena, including studies of peer evaluations (Krosnick 1999; Krosnick et al. 2003).

The SLIP system forces students to make judgmental choices about the merits of Artifacts relative to each other and at the same time allows capturing peer and self-evaluation data as both rating and ranking simultaneously. Specifically, for peer and self-evaluation, each student benchmarked their peers' Artifacts, along with their own Artifact using the SLIP Slider GUI control (Fig. 3). The SLIP Slider displays a continuum from "Very Poor" to "Excellent", on which numbered handles corresponding to each peers' Artifacts can be positioned based on the reviewer's judgment. A rubric to assess "goodness" of analyses was provided to all participants. The M-handle represents the student's self-evaluation. Importantly, students cannot overlap any two handles to indicate an identical "goodness" level. That is, judgments about "goodness" of any two Artifacts have to be at least marginally distinct. Rating is recorded as an integer between 1 and 100 reflecting a position of the essay on the continuum from "Very poor" to "Excellent" independently of the quality of other Artifacts. Ranking was recorded as an integer between 1 (the highest rank) to group size minus 1 (N-1) (the lowest) reflecting a relative position of the essay among other essays in the group.

Several past studies demonstrated that for many tasks ranking-based evaluations are cognitively easier, and therefore, contain significantly less noise, than rating-based evaluations (Barnett 2003; Carterette et al. 2008; Stewart et al. 2005). Ranking has been advocated as a more robust approach to evaluation of complex KA, for example, in peer reviewing and evaluation of conference papers (Douceur 2009) and submissions in online courses (Raman and Joachims 2014; Shah et al. 2013). The results of our analysis are consistent with these findings. For the sake of space, in this paper we present operationalization and empirical analysis of ranking-based assessment. Peer and self-evaluation ranking data are converted into the variables of attainment, bias, controversy and miscalibration (self-assessment inaccuracy). Here we provide only general conceptual summary of these measures; Appendixes 1 and 2 present a detailed algebraic description of calculations based on the algorithm by Ford and Babik (2013).

Attainment score is computed by inverting the ranking; i.e., the rank of 1 is converted to the maximum score, and the rank of (N-1) is converted to the minimum score of 1. Aggregate peer-evaluated attainment is computed as the average of attainment scores produced by peer ranking; self-ranking is excluded from the computation of aggregate peer-evaluated attainment to avoid attainment inflation. Let us illustrate this with as simple example (Fig. 4). Consider the following matrix of the mutual assessment attainment scores in a group of five actors acting as creators of Artifacts and peer evaluators of each other's Artifacts. Each column represents ranks given by each actor to peers' Artifacts, each row represents ranks received by each Artifact. The empty diagonal elements indicate the exclusion of self-assessment from attainment calculations. The higher numbers signify the higher ranking. As can be seen from Fig. 4, actor I receives the aggregate attainment score of 1, and the actor IV receives the aggregate attainment score of 4.

Miscalibration of each Artifact is computed as the difference between the average peer-evaluation attainment score and the self-evaluation attainment score. For illustration, consider the same scenario as in Fig. 4 but now with self-assessment ranks given on the matrix diagonal (Fig. 5). Obviously, actor V shows very low miscalibration (her peer-evaluated attainment score is 4, and her self-assessment attainment score is also 4, hence, miscalibration is zero); whereas, actor I shows very high overconfidence (her peer-evaluated attainment score is 1, and her self-assessment attainment score is also 4, thus, miscalibration is negative 3).

In this study, controversy of a particular Artifact is computed as deviation from mean (DFM); i.e., as the average absolute value of deviations between the attainment score given to the Artifact by each evaluator and the average attainment score given by the rest of evaluators (excluding creator's self-evaluation). Figure 4 illustrates a scenario where each Artifact has zero controversy, i.e., all Artifacts were assigned the same ordinal positions (ranks) by all evaluators. Consider now the following scenario on Fig. 6. Artifacts III, IV, and V show very little variation in received ranks; the average deviation from mean of four respective attainment scores of each of these Artifacts is not very large, and, hence, these Artifacts can be considered non-controversial. In contrast, peer evaluations of Artifact I are polarized (two peers gave it the highest rank and two other – the lowest), the average deviation from mean of four respective attainment scores of this Artifact is large and, therefore, it shows higher level of controversy. Similarly, peer evaluations of Artifact II are scattered through the entire ranking scale, hence, the variation of peer evaluations is large, and, therefore, this Artifact is also more controversial than Artifacts III, IV and V.

Bias of a particular evaluator in this study is also computed as deviation from mean (DFM), i.e., as the average absolute value of deviations between the attainment score given to every Artifact by the evaluator and the average attainment score given to these Artifacts by the rest of evaluators (excluding creator's self-evaluation). Figure 4 illustrates a scenario where each evaluator shows zero bias (with respect to other evaluators), i.e., all evaluators assigned all Artifacts the same ordinal positions (ranks). Consider now the following scenario on Fig. 7. Actors I, II, III, and IV assigned all Artifacts the same ordinal positions (adjusted for exclusion of their self-assessment). Thus, they are in an implicit agreement about attainment of all Artifacts. Actor V, however, assigned ranks to all Artifacts in the reverse order; thus, the average deviation of evaluations of actor V from the average of four evaluations by actors of each respective Artifacts is large, and, hence, actor V can be considered a highly biased evaluator (irrespective of the source of her bias).

For this study, the version of Ford and Babik (2013)'s algorithm that computes controversy and bias as deviations from mean was used as a starting point of our exploration. Alternative approaches to estimation, such as deviations from co-evaluators have also been advocated (Lauw et al. 2006, 2008). The algorithm permits these alternative estimation approaches, and we explored them in our other studies. The algorithm also makes adjustments for the number of submitted Artifacts and evaluations (benchmarks); for the peer group size to make these scores comparable across different groups; and for the bias and controversy nonlinearity due to the use of ranking and the exclusion of self-assessment from the attainment computation (see Appendixes 1 and 2).

After the completion of the assignment, attainment, controversy, bias and self-assessment accuracy scores, as well as other performance indicators, were presented to students in the format shown in Fig. 8.

Table 1 and Table 2 below provide descriptions, descriptive statistics and a correlation matrix of the data collected.

Table 1
Descriptive statistics ($N = 3617$, group size = 5)

Variable	Mean	St Dev	Min	Max
Attainment score (peer ranking), c	3.025	1.077	1.00	5.00
Self-ranking, α	1.954	1.057	1.00	5.00
Attainment score (self-ranking), α_s	4.046	1.057	1.00	5.00
Miscalibration, Δ	−1.021	1.333	−4.00	3.40
Controversy, γ	0.487	0.340	0.00	1.00
Bias, δ	0.494	0.314	0.00	1.00

Table 2
Correlations

Variable	c	α	α_s	Δ	γ
Attainment score (peer ranking), c	1.00				
Self-ranking, α	−0.22	1.00			
Attainment score (self-ranking), α_s	0.22	−1.00	1.00		
Miscalibration, Δ	0.63	0.62	−0.62	1.00	
Controversy, γ	0.09	−0.02	0.02	0.05	1.00
Bias, δ	−0.12	0.07	−0.07	−0.04	0.40

Note that controversy and bias have very little correlation with peer and self-assessment attainment scores.

Since this study is only concerned with the aspects of peer evaluation intersubjectivity, such as miscalibration vis-à-vis peer evaluations, controversy and bias, and not with the "true" performance of creators, the data on the external/expert evaluation of submissions was not collected. The level of competency, i.e., attainment of creators, was only assessed through peer evaluation that represents participants' perception of the Artifacts' "goodness". The following section describes the analysis and results of our hypotheses testing.

Analyses and results

The results of empirical testing of our hypotheses are summarized in Table 3. For consistency and completeness of analysis, we tested our hypotheses using both measurement scales – ranking and rating. In this paper, for the sake of space, we present the results based on ranking data only. While both modes consistently evidence the same effects, we find the rating data to contain more idiosyncratic noise than ranking data and the results derived from it much less conclusive. In the next four subsections, we present the results of hypotheses testing using simple inferential statistics. Since this method produced inconclusive results, we further show the results of hypothesis testing using LGM (in Subsections 4.5 and 4.6).

Table 3
Summary of hypotheses testing

Hypotheses		Conclusion
H1a	KA attainment based on aggregate peer evaluations is systematically different from attainment based on creator's self-evaluation.	Supported
H1b	Attainment based on creator's self-evaluations is systematically higher than attainment based on aggregate-peer-evaluations-based is systematically lower than (i.e., creators are systematically overconfident).	Partially supported
H2	Over multiple successive mutual creator-evaluator interactions, creator miscalibration changes systematically (i.e., non-randomly).	Supported
H3	Creators who produce KAs with higher controversy levels, have higher miscalibration than peers, who produce artifacts with lower controversy levels.	Not supported
H4	Evaluators, who show higher bias levels, have higher level of miscalibration than peers showing lower bias levels.	Not supported

Existence of systematic miscalibration

To test for the existence of miscalibration in our sample, we compared the attainment scores resulting from peer and self-evaluation using the one-sample t-test (Table 4, Table 5, Fig. 9). The presence of a significant miscalibration, i.e., the difference between these two measures, supports our core hypothesis that creators' and peer evaluators' perceptions about the attainment of a KA generally diverge. In assignment 1, the ranking-based peer evaluation produced in a lower attainment score ($M = 3.004$, $SD = 1.021$) than self-assessment ($M = 3.703$, $SD = 1.089$). This difference was significant (in two-tail t-test, $t(399) = -10.706$, $p = 0.000$). Note that while attainment based on peer ranking is essentially constant slightly above 3 across all assignments (because forced ordinal distribution (i.e., ranking) with the median equal 3was used for peer evaluation), attainment based on self-ranking shows signs of steady increase from early assignments to the later assignments.

Table 4
Attainment and miscalibration across assignments

Assignment	Attainment based on peer ranking	Attainment based on self-ranking	Miscalibration	t-statistic	*p*-value
1	3.004	3.703	−0.698	−10.706	0.000
2	3.014	3.790	−0.776	−12.432	0.000
3	3.004	4.038	−1.034	−0.895	0.000
4	3.058	4.162	−1.103	−0.967	0.000
5*	3.029	4.414	−1.384	−1.234	0.000
6	3.023	4.044	−1.020	−0.893	0.000
7	3.001	4.048	−1.047	−0.905	0.000
8	3.024	4.111	−1.087	−0.944	0.000
9	3.056	4.084	−1.028	−0.890	0.000
10	3.037	4.153	−1.116	−15.536	0.000

Assignment with the anomaly in peer review group size; removed from further analysis

Table 5
Descriptive statistics of the pools with positive and negative miscalibration

Pool	No of obs.	Average	St Dev	Min	Max
Overconfident (negative)	2649	−1.627	0.951	−4	−0.070
Underconfident (positive)	968	0.637	0.641	0	3.400

Notably self-evaluation attainment score on average exceeds the peer-evaluation attainment score. Moreover, the significant overconfidence (the negative difference between peer- and self-evaluation ranking-based attainment scores) is observed in all consecutive assignments. This result shows that miscalibration does occur and persists across the series of assignments, i.e., student systematically perceive their attainment to be higher than that of the artifacts created by their peers, and on average this overconfidence increases.

Systematic longitudinal change in miscalibration

Contrary to our expectations guided by social learning theory that in the intersubjective evaluations of complex-tasks KAs, where problem solving and learning are tightly intertwined and where perceptions of actors about the goodness of solutions may differ initially but should converge over multiple iterations of creator-evaluator interactions, we do not observe this pattern in our sample (Table 4, Fig. 9). As students continuously receive feedback from their multiple peers, they are expected to become more self-aware towards later assignments, and, therefore, their self-assessment should be closer to peer evaluation, or, in other words, miscalibration should diminish. Instead, the overall miscalibration increased towards the last assignment. In our study, miscalibration increased from assignment 1 until assignment 5, then suddenly dropped from 5 to 6 (which we attribute the data collection anomaly described above), and then remained practically flat (Fig. 9). Thus, instead of expected convergence, on aggregate, we observe overall widening miscalibration. This counter-theoretic finding leads us to further investigation of longitudinal dynamics of miscalibration (Subsection 4.5).

Impact of controversy

Next, we compared miscalibration for the pools of subjects with high and low controversy in their KAs. In the overall sample, for the ranking-based attainment scores, the students whose Artifacts were not very controversial (controversy below 0.33, lowest one-third) showed a greater miscalibration ($M = -1.199$, $SD = 1.511$) than the students whose Artifacts were very controversial (controversy above 0.67, highest one-third) ($M = -1.010$, $SD = 1.180$). Overall, this difference was significant, ($t(2237) = -3.397$, $p = 0.000$). In assignment 1, the difference between the two pools was significant ($t(186) = -2.224$, $p = 0.027$). This result, however, was not stable: in assignment 2, no evidence of significant difference was found; in assignment 3, the difference is significant at 5 % but not significant at 1 %; in the consecutive assignments, a marginally significant difference was found only in assignment 6 ($p = 0.057$).

We also investigated whether the overconfident subjects (who have a negative peer-self difference (65 % of the sample) differ in controversy from the underconfident subjects (those with a positive peer-self difference). In assignment 1, based on ranking, the pool with negative peer-self differences showed the average controversy ($M = 0.553$, $SD = 0.312$) that was insignificantly different from the pool with positive peer-self differences ($M = 0.526$, $SD = 0.307$) ($t(377) = -0.833$, $p = 0.406$). In other assignments, the results was the same, except assignment 4 ($p = 0.01$), 5 ($p = 0.000$), 7 ($p = 0.01$), 8 ($p = 0.000$), 10 ($p = 0.002$).

Impact of bias

Further, we compared miscalibration for the pools of subjects with high and low evaluator bias. In the overall sample, for the ranking-based attainment scores, the students whose evaluator judgments were unbiased vis-à-vis other reviewers (bias below 0.33) showed a smaller miscalibration ($M = -0.962$, $SD = 1.334$) than the students whose judgments were more biased (bias above 0.67) ($M = -1.080$, $SD = 1.333$). This difference was significant, ($t(2314) = 2.130$, $p = 0.033$). In the individual assignments, however, the difference between the two pools was significant only in assignments 8 ($t(224) = 2.097$, $p = 0.037$) and 9 ($t(219) = 1.978$, $p = 0.049$).

Further, we examined whether the overconfident subjects differ in bias from the underconfident subjects. In assignment 1, based on ranking, the pool with negative peer-self differences showed the average controversy ($M = 0.508$, $SD = 0.300$) that was not significantly different from the pool with positive peer-self differences ($M = 0.541$, $SD = 0.295$) ($t\ (285) = -1.080$, $p = 0.281$). In other assignments, the results was the same, except assignment 5 ($p = 0.000$), 8 ($p = 0.002$), marginally in 9 ($p = 0.084$), and 10 ($p = 0.024$).

In summary, we see that simple inferential statistics and t-tests do not afford conclusive results in our hypotheses testing. Therefore, we continue our investigation with LGM, the method that allows us to look inside the population and differentiate actors with dissimilar traits and behaviors.

Latent longitudinal trajectory dynamics in miscalibration, peer evaluation, and self-assessment

Given the fact that at the aggregate sample level we did not observe theoretically plausible decrease in miscalibration that would indicate social learning and due to inconclusive results with regard to controversy and bias, we conjectured that our sample includes subjects drawn from sub-populations with specific and dissimilar behaviors. These different behaviors cannot be revealed with conventional statistical techniques, such as t-test. LGM is a nascent technique that allows deeper investigation of whether there are latent classes of subjects that demonstrate different behaviors of miscalibration, as well as peer and self-assessment with LGM following the group-based trajectory modeling proposed by Jones, Nagin, and Roeder (2001). LGM models relationships within a single variable as well as between variables by discovering longitudinal patterns characteristic of latent sub-populations Curran et al. (2004); Duncan 1999; Zheng et al. 2014). The existence of such latent classes can provide a basis for explaining our observations. Following the discovery of such classes with LGM, to further understand the relationships between different aspects of intersubjective evaluations, we cross-tabulated frequencies of subjects in different latent classes and controversy and bias categories and applied chi-squared test to check whether actors in different latent classes of miscalibration demonstrate different propensities in attainment, self-assessment, controversy and bias.

LGM represents systematic change (e.g., growth) of repeated measures of a dependent variable as a function of time and other variables and allows investigating inter-individual variability in this change. In LGM, a construct that affects longitudinal dynamics is modeled as a latent random variable with individual unobservable realizations in a sample. Therefore, in our study, with this method, we hypothesize the existence of unobserved latent classes in the subject population that have distinct temporal trajectories of peer and self-perceptions regarding attainment. The SAS TRAJ procedure was used to fit a series of mixture models to the data (Jones et al. 2001). The Bayesian information criterion (BIC) was used to identify the number of classes in the model (Schwarz 1978). Specifically, $2\Delta BIC$, twice the difference between the BIC for the full model (larger number of classes) and that for the reduced model (smaller number of classes), is interpreted as the degree of evidence for the full model. This interpretation is justified because $2\Delta BIC$ is approximately equal to $2\ln B10$, where B10 is the Bayes factor (Kass and Raftery 1995). A value of $2\ln B10$ greater than 10 is interpreted as very strong evidence against the reduced model, which can be replaced by a more complex model, suggesting the presence of an additional latent class (Kass and Wasserman 1995). Table 6 below shows the refinement process through which we select the most reasonable number of latent classes.

Table 6
Tabulated BIC and 2ΔBIC for peer-evaluation attainment, self-assessment attainment and miscalibration from the latent growth analysis

No of classes	Miscalibration		Peer evaluation		Self-assessment	
	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC
1	−5556.03		−5042.26		−4718.46	
2	−5409.39	293.28	−4925.64	233.24	−4297.53	841.86
3	−5385.52	**47.74**	−4910.39	**30.50**	−4254.79	85.48
4	−5390.16	−9.28	−4916.73	−12.68	−4219.47	**70.64**
5	−5393.93	−7.54	−4923.88	−14.30	−4215.05	8.84

The bolded values of 2ΔBIC indicate the largest significant number of latent classes.

For miscalibration, the best fitting model shows three latent classes and a significant cubic trend. Around 64 % of students showed a slight overconfidence (a small negative difference that increased first and then remained stable) (the “0” class); 28 % of students showed substantial overconfidence (the “-” class with a larger negative difference and the pattern similar to that of the “0” class). About 8 % showed slight and growing underconfidence (the “+”class; i.e. these subjects' self-perception is below peer perception and this gap widens as the progress over assignments (Fig. 10).

To understand the nature of the three latent classes discovered in miscalibration, we also looked tested for possible latent classes in attainment based on peer evaluation and self-assessment. For peer evaluation, the best fitting model had three latent classes: around 59 % of students receiving stable median ranking over multiple assignments (Medium), 37 % starting high and continuing to improve (High), and 4 % starting low and declining over time (Low), as shown in Fig. 11.

For self-assessment, the best fitting model had four latent classes and a significant cubic trend: around 46 % of students self-assessing at about median score with the tendency to higher self-assessment over time (Low); 25 % starting above median but below high, increasing toward the middle of the course and then dropping (MediumL); 16 % starting just above median and monotonically increasing towards the end (MediamH); 13 % starting high, and remaining flat at near ceiling (High), as shown in Fig. 12.

To investigate whether a significant relationship exists between the latent classes of miscalibration and peer-evaluated attainment, we cross-tabulated the miscalibration classes against the peer-evaluated attainment classes and conducted the chi-square tests (Table 7).

Table 7
Cross-tabulation of miscalibration classes against peer-evaluated attainment classes

		Latent Miscalibration Class			Total
		Strongly overconfident (−)	Slightly overconfident (0)	Underconfident (+)	
Latent	Low	3 %	1 %	0 %	4 %
Attainment	Medium	21 %	35 %	2 %	58 %
Class	High	2 %	31 %	4 %	38 %
Total		26 %	67 %	7 %	100 %

The differences in the distribution of miscalibration across the latent peer-perceived attainment classes in this sample are statistically significant (Pearson chi-square (df = 4, N = 3617) = 566.08, p-value =0.000. Interestingly, the dominant majority of underconfident creators (i.e. with positive miscalibration) come from the high peer-perceived attainment class, whereas the creators who produced low-attainment essays (as perceived by the peers) mostly show large overconfidence (negative miscalibration). This finding is consistent with the theory of unskilled-and-unaware problem that suggests that the unskilled tend to overestimate the quality of their work, whereas the skilled tend to underestimate the quality of their work.

At the same time, the majority of creators who produced medium attainment Artifacts (the largest class) tend to show near-zero (or slightly negative) miscalibration, suggesting that they are the most accurate self-evaluators (35 % of the overall sample, 60 % of the medium attainment class). Yet, 37 % of the creators with medium attainment Artifacts (21 % of the overall sample) showed large overconfidence. These findings indicate that among creators with a medium skill level, while the majority are accurate evaluators, the inaccurate evaluators (the unskilled-and-unaware) constitute a significant minority.

In summary, by applying LGM we discovered that miscalibration behavior is not uniform. Contrary to the social learning theory, despite multiple iterative creator-evaluator feedback interactions, actors' self-assessment did not converge to peer evaluation, indicating that their self-awareness takes its own path. This finding is, nevertheless, consistent with the predictions of the unskilled-and-unaware problem theory that the “unskilled” are “doomed” to remain unaware. Moreover, looking inside miscalibration, we find that that it is driven by even more complex interactions in peer-evaluated and self-evaluated attainment.

Effects of controversy and bias on miscalibration in the latent growth model

LGM revealed no significant latent trajectories in either controversy or bias. In addition, cross-tabulation and chi-square testing showed that various levels of controversy or bias are associated with particular latent classes of miscalibration. Specifically, in each of the three miscalibration classes, subjects with high, low and near-zero controversy are also distributed practically uniformly, further supporting the lack of stable specific patterns of controversy in our sample (Table 8).

Table 8
Cross-tabulation of miscalibration classes against controversy pools

		Latent Miscalibration Class			Total
		Strongly overconfident (–)	Slightly overconfident (0)	Underconfident (+)	
Controversy	Low	9 %	22 %	2 %	33 %
	No	7 %	19 %	2 %	28 %
	High	10 %	26 %	3 %	38 %
Total		26 %	67 %	7 %	100 %

Similarly, in each of the three miscalibration classes, subjects with high, low and near-zero bias are distributed practically uniformly, further supporting the lack of stable specific patterns of bias in our sample (Table 9).

Table 9
Cross-tabulation of miscalibration classes against bias pools

		Latent Miscalibration Class			Total
		Strongly overconfident (–)	Slightly overconfident (0)	Underconfident (+)	
Bias	Low	7 %	23 %	2 %	32 %
	No	9 %	24 %	2 %	35 %
	High	10 %	21 %	2 %	33 %
Total		26 %	67 %	7 %	100 %

Thus, our hypotheses that controversy and bias are intrinsically related to miscalibration are not supported.

Discussion and conclusion

Discussion

Our analyses revealed some interesting and counter-theoretic results that have implications for researchers, educators and managers. Using simple statistics and t-tests, we found that in concordance with theory, miscalibration between creator’s self-perception and peer evaluators' perceptions regarding KA goodness does exist. Moreover, in the overall sample, we observed significant changes in miscalibration over multiple iterations or creator-evaluators interactions. Social learning and norming theories suggests that perceptions of knowledge community members regarding qualities of the KA will settle, calibrate and converge towards each other over time (Gersick 1991). However, our results showed that contrary to our theory-based expectations, miscalibration does not reduce over multiple iterations as subjects receive peer feedback on their performance. Instead, this miscalibration, prevailingly overconfidence, in the overall sample increased over time.

Intrigued by our counter-theoretic findings, we undertook deeper investigation of evaluation behavior beyond traditional t-tests and analysis of variances, and employed LGM to scrutinize the homogeneity of the data and to understand the temporal changes (Bentein et al. 2005; Leite and Stapleton 2011). This scrupulous analysis of data over multiple iterations of mutual peer and self-evaluations between subjects in random anonymous peer groups revealed latent classes of community members involved in KA creation and evaluation that can be characterized by specific patterns of miscalibration, peer-evaluated and self-evaluated attainment. That is, the results showed that the creation attainment and self-evaluation behaviors are not homogenous across subjects. Specifically, we observed significantly distinct patterns in peer evaluation, self-evaluation and miscalibration among several latent classes.

The miscalibration data showed that the predominant majority of subjects tend to overestimated their performance in comparison with peer-evaluated attainment. Moreover, approximately one in every three overconfident subjects showed significantly increasing overconfidence over multiple iterations. A relative minority of subjects showed underconfidence by self-assessing their KAs lower comparing with peer-evaluated attainment. Furthermore, contrary to theory, underconfidence also increased over multiple iterations. Thus, contrary to theoretic predictions, our findings showed that miscalibration does not attenuate over multiple iterations but instead increases regardless of whether subjects are initially overconfident or underconfident, and regardless of the initial competency level measured by peer-evaluated attainment. Further, we discovered that overconfident subjects tend to become more overconfident, while underconfident subjects become more underconfident. This adversely affects development of shared understanding in the knowledge community. Moreover, miscalibration increases contrary to the expectation that repeated mutual evaluation interactions between actors working on the same problem and feedback provided by peer evaluators to creators reduce miscalibration through social learning.

In peer evaluation, subjects whose Artifacts were evaluated higher by peers in early iterations, tend to show even higher peer-evaluated attainment in the later iterations, thus demonstrating learning as perceived by peer evaluators. In contrast, subjects whose Artifacts received lower peer evaluations in earlier iterations, also receive distinctly lower evaluations in subsequent iterations, thus demonstrating regress in their performance. Latent-class analysis of self-assessment also revealed several interesting behavioral patterns; in particular, just under half of all subjects tend to rank themselves around or slightly above median quite persistently over multiple iterations.

Our results also indicated that different subjects miscalibrate in different ways. In concordance with the unskilled-and-unaware theory, subjects whose work is evaluated highly by their peers tend to showed underconfidence by self-assessing their own work lower; in contrast, subjects whose work is peer-evaluated low demonstrated overconfidence in their work by self-assessing their own work higher than peers. Most strongly overconfident students come from the latent class of medium attainment performers – roughly, two out of three overconfident students are students given medium scores by their peers.

While there are theoretic reasons to assert that controversy and bias may be important intersubjectivity factors associated with miscalibration, we could not find any significant effect of controversy and bias on miscalibration with the models that we tested. In our study, we used the average deviation from mean approach to capture the phenomena of bias and controversy. Alternative approaches could be applied, such as the average deviation from co-evaluators, that have been argued to have advantages over the deviation from mean approach (Lauw et al. 2006, 2008). The lack of the evidence of the association of miscalibration with bias and controversy may be either due to the model (mis)specification or the choice of measures. To keep the present study parsimonious, we deliberately applied the approach described by Lauw et al. (2006) as naïve, that uses the deviation from mean approach and ignores the mutual dependency of bias and controversy in capturing these phenomena. This may have resulted in a weaker signal of bias or controversy in our data, but helps establish the base line for our future studies of the phenomenon. It informs and inspires an interesting avenue for our future research to employ more advanced models for studying bias and controversy, such as the reinforcement-based model Lauw et al. (2008) or non-linear models Roos et al. (2012), to investigate the impact of bias and controversy on miscalibration. The lack of evidence that latent classes of controversy and bias exist is indicative of the conflict between theories of intersubjectivity and the unskilled-and-unaware problem, which explains that in peer-based knowledge creation and refinement systems where objective external criteria of goodness are not applicable, we cannot evaluate one subject's competency unless we establish competencies of other subjects. This is a circular problem that is difficult to solve endogenously.

We also found that the use of forced-distribution ordinal scale (ranking) to measure attainment is more robust with respect to the overall level of competency of the actors in a given population and less noisy with respect to individual idiosyncratic biases of evaluators than continuous cardinal scale (rating). In our analyses, the use of ranking reveals sharper differences in attainment and miscalibration across latent classes than rating.

In summary, our results show, in agreement with theory and findings of earlier studies that a significant miscalibration (i.e., a gap between peer and self-evaluation) does exist and that this gap changes over multiple mutual review iterations in re-mixed creator-evaluator groups. Contrary to the theory and our expectations, miscalibration does not decrease over time but, instead, it increases. On average, miscalibration increases regardless of the goodness of Artifacts and competency of creators, as well as its initial magnitude and direction. This implies that miscalibration is endogenous in nature (Meyer 1995). By observing the change of miscalibration over time, we removed the effect of initial conditions (and thus, fixing some of the weaknesses of the quasi-experimental design). This allows us to focus on changes endogenous or systemic to the creator-evaluator social system. We observe that miscalibration is amplified over time as the signal of self-perception. This is also consistent with the unskilled-and-unaware problem view (the unskilled underestimate how bad they are, while the skilled underestimate how good they are) and contrasts with the social system view (in the social system where the abilities of actors vary, high-ability actors and low-ability actors act differently, but over time converge in their abilities and behaviors).

Contributions

This study adds to the existing literature in IT-enabled knowledge creation and learning social systems by further exploring the phenomenon of miscalibration and relating it to other factors of creation and evaluation performance and intersubjectivity, such as peer-evaluated and self-assessed attainment, controversy and bias. Specifically, we investigated the longitudinal dynamics of miscalibration over multiple anonymous creator-evaluator interactions.

We make theoretical contribution by developing hypotheses regarding the miscalibration of self-assessment with respect to peer evaluation in creator-evaluator interactions, its change over multiple iterations, and the effects of the KA controversy and evaluator bias on this gap. Our methodological contribution is in developing a method for measuring the KA controversy and evaluator bias in mutual creator-evaluator social interactions based on forced-distribution ordinal scale (ranking). We make empirical contribution by testing these hypotheses with a large sample of student case analyses subjected to multiple peer evaluation and the creators' self- assessment. We designed an information system to test the hypotheses about socio-technical interactions occurring in epistemic, practice and other open knowledge creation communities. Importantly, to test our hypotheses, we applied LGM method rarely used in the information systems research. Our study is among the very few to study intersubjectivity in knowledge creation and evaluation on a longitudinal basis. Our thorough literature review did not reveal any studies that identified distinct classes of the knowledge community actors that exhibited distinct behaviors regarding peer and self-evaluations of KAs. The applications of the LGM methodology produced results consistent with the unskilled-and-unaware problem theory.

Limitations and directions for future research

We would like to acknowledge the following limitations of our study, which also grant promising future research opportunities. First, we used a sample of undergraduate students in one university; therefore, any generalizations to other populations should be made with caution. We intend to replicate our results with multiple samples of undergraduate and graduate students in various disciplines from different universities, as well as in non-academic settings (we have access to these data through the SLIP). Second, we observed an idiosyncratic anomaly in the sample in one of the iterations. While we removed this instance from the analysis and assumed that this shock had no significant effects on the findings of the study, we did not test whether this is true; for the future studies we will use samples with no such anomalies. Third, some important components of intersubjective interaction between creators and evaluators, such as the intra-group inter-observer reliability, were not considered in this study. The SLIP computes this metric, and we intend to incorporate it in our empirical model in the future studies. Fourth, in this study, we calculated controversy and bias using the naïve approach that Lauw et al. (2008) criticized for ignoring the fact that bias and controversy mutually affect each other. We agree with this criticism and in the future studies intend to modify our system design to account for this reciprocal effect. Simultaneous estimation of controversy and bias as suggested by Lauw et al. (2008) and application the cultural consensus theory (Anders and Batchelder 2012; Batchelder and Anders 2012) are possible solutions to the endogeneity problem in studies evaluation of complex-task KA where objective criteria of goodness are not applicable.

Implications

Our findings have important implications for a number of domains. In knowledge production, refinement and management, especially in the business community, identifying and differentiating the value of contributions made by actors with varying abilities and skills are highly important. We suggest a set of metrics and a methodology for identifying competent contributions and evaluations and tracing them over time. Specifically, these metrics and methodology can be used in design of peer-based knowledge creation and refinement information systems. These metrics and methodology can also be applied in social-media-based open knowledge and recommender systems that people rely on to guide their daily decisions. For learning, alleviating miscalibration has effect on self-awareness and self-efficacy. Educational systems for peer review and peer learning should be designed to help actors with varying abilities recognize their strength and weaknesses. For decision making and decision support systems, recognizing controversial and biased contributors and identifying the nature and sources of their controversy and bias helps prevent undesired behaviors such as “group-think”.

Appendix 1: Measurement Operationalization (Single-Group Ranking Model)

Suppose, each peer group consists of N subjects that are indexed $i = \{1, 2, \ldots, N\}$. Each subject $i$rank-orders other (N-1) subjects' Artifacts so that the “best” is ranked 1 and the “worst” is ranked (N-1), that is, here only subject i’s rank-ordering of his peers' Artifacts is considered. The matrix of mutual evaluations of Artifacts produced by the group is

$$\boldsymbol{A}_{N\times N} = [a_{ij}]_{N\times N} = \begin{bmatrix} N & a_{12} & \cdots & a_{1N} \\ a_{21} & N & & a_{2N} \\ \vdots & & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & N \end{bmatrix}$$

where ranks given are in rows, ranks received are in columns, a_{ij} denotes a rank given by a subject i to a subject j for the Artifact (or, symmetrically, received by a subject j from a subject i). Note that $\boldsymbol{a}_i = [a_{i1} \; a_{i2} \ldots a_{ij} \ldots a_{iN}]$ is a row vector of ranks given by a subject i to all his peers such that

$$\begin{cases} a_{ij} = N \; if \; i = j \\ a_{ij} \in \{1, 2, \ldots, N-1\} \; if \; i \neq j \\ a_{i1} \neq a_{i2} \neq \cdots \neq a_{ij} \neq \cdots \neq a_{iN} \\ a_{ij} = N \; if \; E_j = 0 \end{cases}$$

where E_j is the indicator function such that

$$E_j = \begin{cases} 1 \; if \; the \; Artifact \; was \; submitted \; by \; a \; subject \; j \\ 0 \; if \; the \; Artifact \; was \; not \; submitted \; by \; a \; subject \; j \end{cases}$$

Note that the row vector $\boldsymbol{e} = [E_1 \; E_2 \ldots E_N]$ is the vector of the indicator function's values. These conditions constitute the data integrity constraints. The first condition means that a subject *i*'s assessment of his own Submission is not included in this matrix. The second condition means that each peer's Artifact, without exceptions, needs to be rank-ordered. The third condition means that rank-ordering is enforced, that is, no two peers' Artifacts may have the same rank. Note that relaxing this assumption allows each Artifact to be rated rather than ranked. The forth condition means that a missing Artifact is given *N* points.

Suppose $C > 1$ is the maximum possible attainment score for an Artifact; i.e., the attainment score of *C* is given to an Artifact that received the rank of 1, and the attainment score of 1 is given to an Artifact that received the rank of (*N-1*). A failure to submit an Artifact results in the attainment score of 0. Then the following transforms the rank a_{ij} given by a subject *i* to subject *j* into the attainment score c_{ji} received by subject *j* from subject *i*:

$$c_{ji} = \begin{cases} 1 + \left(\sum_{h=1}^{N} E_h - 1 - a_{ij}\right) \frac{D-1}{\sum_{h=1}^{N} E_h - 2} \; if \; a_{ij} \neq N \\ 0 \; if \; a_{ij} = N, \end{cases}$$

Or

$$c_{ji} = \begin{cases} a_{ij} \frac{(1-C)}{\sum_{h=1}^{N} E_h - 2} + \frac{C\left(\sum_{h=1}^{N} E_h - 1\right) - 1}{\sum_{h=1}^{N} E_h - 2} \; if \; a_{ij} \neq N \\ 0 \; if \; a_{ij} = N \end{cases}$$

For example, suppose $N = 6$ subjects and $C = 5$ points. Then, the transformation rule is:

Rank a_{ij}	Attainment score c_{ij}
1	5
2	4
3	3
4	2
5	1
Not submitted (6)	0

The matrix of the individual Artifact attainment scores for the entire group is (scores received are in rows, scores given are in columns)

$$C_{N\times N} = A' \frac{(1-C)}{1e' - 2} + 1'1 \frac{C\left(1e' - 1\right) - 1}{1e' - 2}$$

where scores received are in rows, scores given are in columns,

$\mathbf{1}_{1\times N} = [1\ 1 \ldots 1]$ is the row vector of ones,

$c_{ji} = 0$ for all $a_{ji} = N$, and

$$1e' = \sum_{j=1}^{N} E_j \leq N.$$

A subject *i*'s attainment score for the Artifact is the average attainment score received from all his peers, who submitted their evaluations, ideally (*N-1*). Hence, the column vector of attainment scores for Artifacts is

$$\bar{c} = \frac{C\,1'}{1f' - 1}$$

where the row vector $\boldsymbol{f} = [F_1\ F_2 \ldots F_N]$ is the vector of values of the indicator function F_j such that

$$F_j = \begin{cases} 1\ if\ the\ Artifact\ Evaluation\ was\ submitted\ by\ subject\ j \\ 0\ if\ the\ Artifact\ Evaluation\ was\ not\ submitted\ by\ subject\ j, \end{cases}$$

$$1f' = \sum_{j=1}^{N} F_j \leq N$$

Hence, the attainment score for the Artifact of subject i is

$$\bar{c_i} = \frac{\boldsymbol{c_i}\,1^{'}}{1\boldsymbol{f}^{'} - 1}$$

where $\boldsymbol{c}_{i1\times N}$ is the row vector of Artifact attainment scores received by a subject i.

Subjects are also required to self-evaluate, that is, rank-order their own Artifacts among those of other peers. However, their self-ranking is not included in the calculation of attainment of their own Artifact. The difference between attainment measures derived from self-assessment and peer assessment results is computed as follows. Suppose α_i is a rank given by a subject i to his own Artifact, such that α_i coincides with one of the values $\boldsymbol{a}_i$. Then, self-evaluation score for Artifact attainment is defined as

$$\varepsilon_i = \begin{cases} \left(1 + (N - \alpha_i)\frac{C-1}{N-1}\right)\ if\ F_i = 1 \\ \varnothing\ if\ F_i = 0, \end{cases}$$

and miscalibration, i.e., the difference between the attainment score produced by self-assessment and the attainment score produced by peer assessment) for the Artifact of a student i is defined as

$$\Delta_i = \begin{cases} \frac{\bar{c}_i - \varepsilon_i}{C-1}\ if\ F_i = 1 \\ \varnothing\ if\ F_i = 0. \end{cases}$$

The assessor error (*ER*) measures a given subject's divergence from assessments of each peer's Artifact by the rest of the peer group. Subject i's *assessor error* is defined as

$$\delta_i = \begin{cases} \sum_{j=1}^{N} |\bar{c}_j - c_{ji}|\ \forall\, i \neq j\ if\ F_i = 1 \\ \varnothing\ if\ F_i = 0 \end{cases}$$

where |. | denotes the absolute value operator, and ø denotes an undefined (missing) value (assessor error cannot be calculated if assessment was not submitted). The column vector of assessor error measures is

$$\boldsymbol{\delta} = \left|(\bar{\boldsymbol{c}}1)^{'} - \boldsymbol{C}^{'}\right|\,1^{'}$$

where |. | denotes the matrix of absolute values of the element-wise differences of the two square matrices (not the determinant of the matrix). In this vector, all elements for which F_i is zero are undefined (missing values).

The *assessee error* (EE) is a measure of divergence among peers in assessing a given subject's Artifact. Subject i's assesse error is defined as

$$\gamma_i = \begin{cases} \sum_{j=1}^{N} |\bar{c}_i - c_{ij}| \ \forall\, i \neq j \ if \ E_i = 1 \\ \varnothing \ if \ E_i = 0 \end{cases}$$

The column vector of assesse error measures is

$$\boldsymbol{\gamma} = (|\bar{\boldsymbol{c}}\mathbf{1} - \boldsymbol{C}| \circ \boldsymbol{Q})\ \mathbf{1}'$$

where $\boldsymbol{Q}_{N\times N}$ is a square matrix with zeros on the diagonal and ones off diagonal, the operator "$\circ$" denotes the Hadamard product of matrices (entry-wise product operator). In this vector, all elements for which E_i is zero are undefined (missing values).

The *average group assessor error* (AGER) is defined as

$$\bar{\delta} = \frac{\mathbf{1}\boldsymbol{\delta}}{\mathbf{1}\boldsymbol{f}'} = \frac{\sum_{i=1}^{N} \delta_i}{\sum_{i=1}^{N} F_i}$$

The *average group assessee error* (AGEE) is defined as

$$\bar{\gamma} = \frac{\mathbf{1}\boldsymbol{\gamma}}{\mathbf{1}\boldsymbol{e}'} = \frac{\sum_{i=1}^{N} \gamma_i}{\sum_{i=1}^{N} E_i}$$

It can be shown that corresponding AGER and AGEE are equal.

The *intra-group inter-observer reliability* (IGIOR) for any given group is defined as

$$y = \frac{\bar{\delta}_{div} - \bar{\delta}}{\bar{\delta}_{div} - \bar{\delta}_{con}}$$

where δ−δ− is the AGER of the given group,

$\delta-_{con}$δ−con is the AGER of the group with the *perfect convergence* among peers' evaluations of each other's Artifacts on the ordinal-scale (relative ranks),

$\delta-_{div}$δ−div is the AGER of the group with the *perfect divergence* among peers' evaluations of each other's Artifacts on the ordinal-scale (relative ranks).

The IGIOR can be interpreted as how far the given peer group as a whole is from the perfect convergence on ranking each other's Artifact. For a group with the perfect convergence, the IGIOR is equal 1; for a group with the perfect divergence, the IGIOR is equal 0. Note that since it can be shown that corresponding AGER and AGEE are equal, it does not matter whether either AGER or AGEE are used to compute the IGIOR.

Bias (*normalized ER*) is the ER adjusted for the chosen values of *C* so that it ranges between 0 and 1. Bias can be interpreted as a measure of a given subject's divergence from the rest of the peer group in assessing each peer's Artifact irrespective of the overall rank of the subject's Artifact and the maximum possible attainment score.

A given subject *i*'s *bias* is defined as

$$\hat{\delta}_i\left(\delta_i, r_i\right) = \frac{\delta_i - y\,\delta_{con}\left(r_i\left(\bar{c}_i\right)\right)}{\delta_{div}}$$

where r_i is the rank of the subject *i*'s Artifact among the rest of the Artifacts of the peer group based on the Artifact's attainment score (in other words, r_i corresponding to the larges value in the vector c–c– is equal to 1 and r_i corresponding to the smallest value in the vector c–c– is equal to *N*);

$\delta_{con}(r_i)$ is the ER corresponding to the rank r_i in the peer groups with the perfect convergence (IGIOR $y = 1$);

δ_{div} is the ER in the peer groups with the perfect divergence (IGIOR $y = 0$). In a peer groups with the perfect divergence, all peers have the same ER because no one is better in assessing others than the rest of the group. Computations of $\delta_{con}(r_i)$ and δ_{div} are explained in the appendix 2.

Controversy (*normalized EE*) is the EE adjusted for the chosen values of *C* so that it ranges between 0 and 1. Controversy can be interpreted as a measure of divergence among peers in assessment of a given subject's Artifact irrespective of the overall ranks their Artifacts and the maximum possible attainment score for the Artifact.

Controversy of an Artifact produced by a subject *i* is defined as

$$\widehat{\gamma}_i\left(\gamma_i, r_i\right) = \frac{\gamma_i - y \quad \gamma_{con}\left(r_i\left(\bar{c}_i\right)\right)}{\gamma_{div}}.$$

Bias and controversy are recorded in vectors δˆ and γˆ respectively. Normalization is necessary because while ER and EE are the same for all subjects in the peer group with the perfect divergence (because no one is better in assessing Artifacts than the rest of the group), in the peer group with the perfect convergence ER and EE have non-zero values and are non-linearly dependent on the rank r_i of the subject *i*'s Artifact among the Artifacts of the peer group based on the attainment score. In other words, in a peer group with non-equal attainment scores (that is when at least some convergence exists among peer on the quality of each submission and submissions can be ranked based on the score), each rank position is characterized by a systematic non-zero ER and EE just because of its place in the relative ranking of peers' Artifacts. This is due to the fact that subjects' own self-evaluations are not included in the computations of the attainment scores. A more detailed explanation of this phenomenon is given below.

This single-groups ranking model can be easily extended to multiple groups.

Appendix 2: Correcting Nonlinearity in Ranking (Computing AGER and AGEE for Peer Groups with Perfect Convergence and Perfect Divergence)

To determine IGIOR, bias and controversy, AGER and AGEE for the extreme special cases of the perfect convergence and the perfect divergence of a peer group need to be computed. In addition, ER for each relative rank position in a peer group needs to be computed for the case of perfect convergence.

First, let us consider the case of the *perfect intra-group inter-observer convergence*, that is, the summative assessment result for which IGIOR is equal 1. Suppose the row vector $\boldsymbol{s}_{1\times N} = [1, 2, \ldots, N]$ is the vector of latent ranks of potential attainment (*goodness*) of a given set of Artifacts. That is, we assume that each Artifact is of such *goodness* and each subject has such evaluation skills that when asked to rank-order these Artifacts, the subjects come to the perfect convergence on ranking of each Artifact (we also assume that all subjects submit their Artifacts). Under these assumptions, the latent ranks should be equal to the ranks generated by the system, that is $s_i = r_i$ for all i. However, since each Artifact's attainment is computed using ranks received from peers (ranging in $\{1, 2, \ldots, N-1\}$) and excluding subject's self-ranking of his own Artifact, despite the perfect convergence each subject will make an assessor error (ER) of a various degree. For example, in a peer group of six subjects, subject 1 will not be able to give his own Artifact the rank of 1 but would have to give it to the Artifact of subject 2. Similarly, subjects 2, 3, 4, and 5 will have to give the rank of 5 to the Artifact of subject 6. Consequently, each Artifact will also bear an assessee error (EE) of varying degree.

The matrix $\boldsymbol{A}_{con}$ of ranks given in a peer group with the perfect convergence is obtained from the vector $\boldsymbol{s}$ by the following transformation

$$\boldsymbol{A}_{con} = N\,\boldsymbol{I} + \left(1'\boldsymbol{s}\,\boldsymbol{H}_1\right) \circ \boldsymbol{T}_1 + \left(1'\boldsymbol{s}\,\boldsymbol{H}_1\boldsymbol{H}_2\right) \circ \boldsymbol{T}_2$$

where $\boldsymbol{I}_{N\times N}$ is the identity matrix, $\boldsymbol{1}_{1\times N}$ is a vector of ones,

$\boldsymbol{H}_{1\,N\times N}$ is a square shift matrix with all elements equal zero except for elements equal one just above the main diagonal,

$\boldsymbol{H}_{2\,N\times N}$ is a square shift matrix with all elements equal zero except for elements equal one just below the main diagonal,

$\boldsymbol{T}_{1\,N\times N}$ is a square matrix with all elements in the upper triangle above the main diagonal equal ones and all other equal zero,

$\boldsymbol{T}_{2\,N\times N}$ is a square matrix with all elements in the lower triangle below the main diagonal equal ones and all other equal zero, the operator "∘" denotes the Hadamard product of matrices (entry-wise product operator).

For the special case of $N = 6$, the matrix $\boldsymbol{A}_{con}$ looks as follows:

6	1	2	3	4	5
1	6	2	3	4	5
1	2	6	3	4	5
1	2	3	6	4	5
1	2	3	4	6	5
1	2	3	4	5	6

Using the rule for transforming ranks into scores the matrix of scores $\boldsymbol{C}_{con}$ is obtained as

$$\boldsymbol{C}_{con} = \boldsymbol{A}_{con}{'}\,\frac{(1-C)}{N-2} + 1{'}1\,\frac{C\,(N-1)-1}{N-2}$$

such that

$$c_{ji} = \begin{cases} 1 + (N - a_{ij} - 1)\,\frac{C-1}{N-2} = a_{ij}\frac{(1-C)}{N-2} + \frac{C(N-1)-1}{N-2}\; if\; a_{ij} \neq N \\ 0\; if\; a_{ij} = N \end{cases}$$

For the case of N = 6 and C = 5, the matrix $\boldsymbol{C}_{con}$ looks as follows:

0.00	5.00	5.00	5.00	5.00	5.00
5.00	0.00	4.00	4.00	4.00	4.00
4.00	4.00	0.00	3.00	3.00	3.00
3.00	3.00	3.00	0.00	2.00	2.00
2.00	2.00	2.00	2.00	0.00	1.00
1.00	1.00	1.00	1.00	1.00	0.00

The column vector of attainment scores for the peer group in the perfect convergence is

$$\bar{\boldsymbol{c}}_{con} = \frac{\boldsymbol{C}_{con}{'}}{1}\,1\boldsymbol{f}{'}$$

The column vector of ER is

$$\boldsymbol{\delta}_{con} = \left|(\bar{\boldsymbol{c}}_{con}1)' - \boldsymbol{C}'_{con}\right| 1'$$

where |. | denotes the matrix of absolute values of the element-wise differences of the two square matrices (not the determinant of the matrix).

The column vector of EE is

$$\boldsymbol{\gamma}_A = (|\bar{\boldsymbol{c}}_{con}1 - \boldsymbol{C}_{con}| \circ \boldsymbol{Q})\ 1'$$

where $\boldsymbol{Q}_{N\times N}$ is a square matrix with zeros on the diagonal and ones off diagonal, the operator "°" denotes the Hadamard product of matrices (entry-wise product operator).

The following table summarizes the attainment, assessor error (ER) and assessee errors (EE) scores for the case of the peer group with the perfect convergence where N = 6 and C = 5:

s = r	$\bar{c}_{con}$	$\delta_{con}(r)$	$y_{con}(r)$
1	5.00	2.00	0.00
2	4.20	1.20	1.60
3	3.40	0.80	2.40
4	2.60	0.80	2.40
5	1.80	1.20	1.60
6	1.00	2.00	0.00

The following diagram graphically represents attainment, ER and EE scores for the extreme special case of the peer group with the perfect convergence where N = 6 and C = 5:

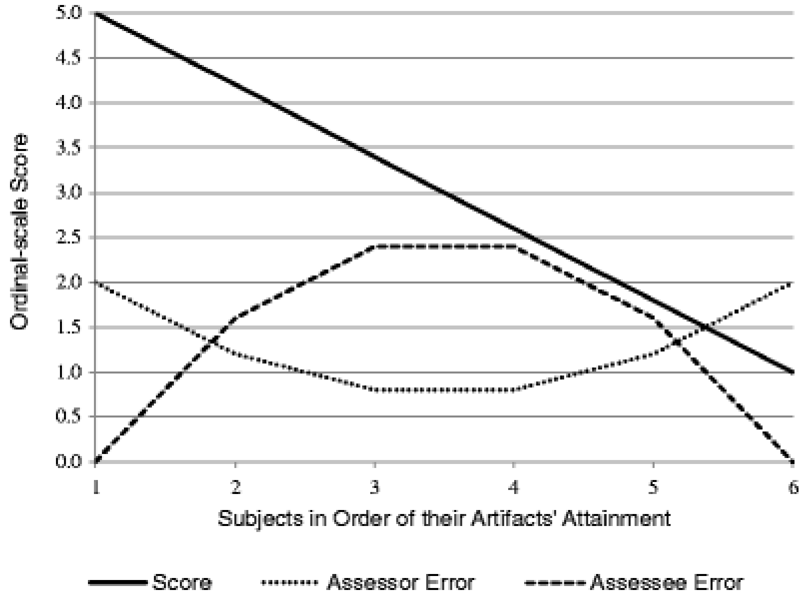

Figure. 13. Non-linear Error Behavior in the Perfect Convergence Group (N = 6, C = 5)

Thus, we obtained ER and EE scores for the extreme special case of the peer group with the perfect convergence.

The AGER for the peer groups with the perfect convergence is

$$\bar{\delta}_{con} = \frac{\mathbf{1}\boldsymbol{\delta}_{con}}{\mathbf{1}\boldsymbol{f}'} = \frac{\sum_{i=1}^{N} \delta_{con\,i}}{\sum_{i=1}^{N} F_i}.$$

The AGEE for the peer groups with the perfect convergence is

$$\bar{\gamma}_{con} = \frac{\mathbf{1}\boldsymbol{\gamma}_{con}}{\mathbf{1}\boldsymbol{e}'} = \frac{\sum_{i=1}^{N} \gamma_{con\,i}}{\sum_{i=1}^{N} E_i}.$$

For the peer group with the perfect convergence with $N = 6$ and $C = 5$, the AGER δ−conδ−conand the AGEE γ−conγ−con are both equal 4/3.

Thus, we obtained values of AGER and AGEE for each relative rank for the extreme case of the perfect convergence.

Now, let us consider the case of the *perfect intra-group inter-observer divergence*, that is, the result of mutual summative peer assessment, for which IGIOR is equal zero. In this case, the row vector $\boldsymbol{s}_{1\times N} = [1, 2, \ldots, N]$ does not reflect latent ranks of quality of a given set of Artifacts. That is, we assume that each Artifact is of such *goodness*, and each subject has such evaluation skills that when asked to rank-order these Artifacts the subjects are not be able to converge on the*goodness* of Artifacts and their ranking, resulting in the perfect divergence on ranking of each Artifact (we also again assume that all subjects submit the deliverable). In other words, the latent*goodness* of all Artifacts and evaluation skills of all subjects are assumed to be absolutely equivalent/homogenous. Under these assumptions, each subject's Artifact receives from his peers the entire range of possible ranks, and no one subject is better in assessing his peers than the rest of the group. The matrix $\boldsymbol{A}_{div}$ of ranks given in a peer group with the perfect divergence, therefore, is a Latin square – an $N \times N$ matrix filled with integers from $\{1, 2, \ldots, N\}$, each occurring exactly once in each row and exactly once in each column. One of the ways such matrix can be constructed is by the Cyclic Method (Bailey, 2008): Place $\boldsymbol{s}$ in reverse order (or (N+1)1−s(N+1)1−s) in the top row of $\boldsymbol{A}_{div}$; in the second row, shift all the integers to the right one place, moving the last symbol to the front; continue in this fashion, shifting each row one place to the right of the previous row.

For the special case of $N = 6$, the matrix $\boldsymbol{A}_{div}$ looks as follows:

6	5	4	3	2	1
1	6	5	4	3	2
2	1	6	5	4	3
3	2	1	6	5	4
4	3	2	1	6	5
5	4	3	2	1	6

Note that for the purpose of obtaining ER and EE scores for the case of the perfect divergence the method of obtaining $\boldsymbol{A}_{div}$ matrix does not matter as long as it is a Latin square with the diagonal elements equal N.

Similarly to the case of the perfect convergence, using the rule for transforming ranks into scores the matrix of attainment scores $\boldsymbol{C}_{div}$ is obtained as

$$\boldsymbol{C}_{div} = \boldsymbol{A}_{div}' \frac{(1-C)}{N-2} + 1'1 \frac{C(N-1)-1}{N-2}$$

such that

$$c_{ji} = \begin{cases} 1 + (N - a_{ij} - 1)\frac{C-1}{N-2} = a_{ij}\frac{(1-C)}{N-2} + \frac{C(N-1)-1}{N-2} \; if \; a_{ij} \neq N \\ 0 \; if \; a_{ij} = N \end{cases}$$

For the case of $N = 6$ and $C = 5$, the matrix $\boldsymbol{C}_{div}$ looks as follows:

0.00	5.00	4.00	3.00	2.00	1.00
1.00	0.00	5.00	4.00	3.00	2.00
2.00	1.00	0.00	5.00	4.00	3.00
3.00	2.00	1.00	0.00	5.00	4.00
4.00	3.00	2.00	1.00	0.00	5.00
5.00	4.00	3.00	2.00	1.00	0.00

The column vector of attainment scores for the group in the perfect convergence is

$$\bar{c}_{div} = \frac{{C_{div}}'}{1} 1 f'$$

The column vector of assessor errors is

$$\delta_{div} = \left| (\bar{c}_{div} 1)' - C'_{div} \right| 1'$$

The column vector of assessee errors is

$$\gamma_{div} = (|\bar{c}_{div} 1 - C_{div}| \circ Q)\ 1'$$

The following table summarizes attainment, bias and controversy scores for the case of the peer group with the perfect divergence where N = 6 and C = 5:

s = r	$\bar{c}_{div}$	$\delta_{div}(r)$	$\gamma_{div}(r)$
1	3.00	6.00	6.00
2	3.00	6.00	6.00
3	3.00	6.00	6.00
4	3.00	6.00	6.00
5	3.00	6.00	6.00
6	3.00	6.00	6.00

Thus, we obtained values of ER and EE for the extreme special case of the peer group with the perfect divergence. Note that all subjects' submissions in such group are characterized by equal attainment, ER and EE scores.

The AGER for the peer group with the perfect divergence is

$$\bar{\delta}_{div} = \frac{1\boldsymbol{\delta}_{div}}{1\boldsymbol{f}'} = \frac{\sum_{i=1}^{N-1} \delta_{div\,i}}{\sum_{i=1}^{N-1} F_i}.$$

The AGEE for the peer group with the perfect divergence is

$$\bar{\gamma}_{div} = \frac{1\boldsymbol{\gamma}_{div}}{1\boldsymbol{e}'} = \frac{\sum_{i=1}^{N-1} \gamma_{div\,i}}{\sum_{i=1}^{N-1} E_i}.$$

For the peer group with the perfect divergence where *N* = 6 and *C* = 5, AGER δ−divδ−divand AGEE γ−divγ−divare both equal 6.

Thus, we obtained values of AGER and AGEE for the extreme special case of the perfect divergence. The IGIOR for the case of the perfect convergence y_{con}=1 and for the case of the perfect divergence y_{div}=0.

References

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: a comparison of ratings and rankings. Public Opinion Quarterly, 49(4), 535–552. doi:10.1086/268949.

Amin, A., & Roberts, J. (2008). Knowing in action: beyond communities of practice. Research Policy, 37(2), 353–369. doi:10.1016/j.respol.2007.11.003.

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. Journal of Mathematical Psychology, 56(6), 452–469.

Bandura, A. (1962). Social learning through imitation. University of Nebraska Press.

Bandura, A. (1977). Self-efficacy: toward a unifying theory of behavioral change. Psychological Review, 84, 191–215. doi:10.1037/0033-295X.84.2.191.

Bandura, A., & Walters, R. H. (1963). Social Learning and Personality Development. New York, Holt, Rinehart and Winston [1963].

Barnett, W. (2003). The modern theory of consumer behavior: ordinal or cardinal? The Quarterly Journal of Austrian Economics, 6(1), 41–65.

Batchelder, W. H., & Anders, R. (2012). cultural consensus theory: comparing different concepts of Cultural truth. Journal of Mathematical Psychology, 56(5), 316–332.

Bentein, K., Vandenberghe, C., Vandenberg, R., & Stinglhamber, F. (2005). The role of change in the relationship between commitment and turnover: a latent growth modeling approach. Journal of Applied Psychology, 90(3), 468–482. doi:10.1037/0021-9010.90.3.468.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice, 5(1), 7–74. doi:10.1080/0969595980050102.

Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: a socio-technical perspective. Part I: the causes. MIS Quarterly, 1(3), 17–32.
Bouzidi, L., & Jaillet, A. (2009). Can online peer assessment be trusted? Educational Technology & Society, 12(4), 257–268.
Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32–42. doi:10.3102/0013189X018001032.
Brown, J. S., & Duguid, P. (2001). Knowledge and organization: a social-practice perspective. Organization Science, 12(2), 198–213. doi:10.1287/orsc.12.2.198.10116.
Brutus, S., & Donia, M. B. L. (2010). Improving the effectiveness of students in groups with a centralized peer evaluation system. The Academy of Management Learning and Education (AMLE), 9(4), 652–662.
Brutus, S., Donia, M. B. L., & Ronen, S. (2013). Can business students learn to evaluate better? Evidence from repeated exposure to a peer-evaluation system. Academy of Management Learning & Education Academy of Management Learning & Education, 12(1), 18–31.
Buchy, M., & Quinlan, K. M. (2000). Adapting the scoring matrix: a case study of adapting disciplinary tools for learning centred evaluation. Assessment & Evaluation in Higher Education, 25(1), 81–91. doi:10.1080/713611419.
Campbell, D. J. (1988). Task complexity: a review and analysis. Academy of Management Review, 13(1), 40–52.
Carterette, B., Bennett, P. N., Chickering, D. M., & Dumais, S. T. (2008). Here or there: preference judgments for relevance. Lecture Notes in Computer Science, 4956, 16–27.
Cho, K., Chung, T. R., King, W. R., & Schunn, C. D. (2008). Peer-based computer-supported knowledge refinement: an empirical investigation. Communications of the ACM, 51(3), 83–88. doi:10.1145/1325555.1325571.
Conklin, J. (2001). Wicked problems and social complexity. Dialogue Mapping:Building Shared Understanding of Wicked Problems Retrieved from http://www.ideapartnership.org/documents/wickedproblems.pdf.
Crooks, T. (2001). The validity of formative assessments. University of Leeds.
Curran, P., Bauer, D., & Willoughby, M. (2004). Testing main effects and interactions in latent curve analysis. Psychological Methods, 9(2), 220–237.
Cusinato, A., Della Mea, V., Di Salvatore, F., & Mizzaro, S. (2009). QuWi: Quality control in Wikipedia. In Proceedings of the 3rd workshop on information credibility on the web (pp. 27–34). New York, NY, USA: ACM. doi:10.1145/1526993.1527001.
Dai, H., Zhu, F., Lim, E.-P., & Pang, H. (2012). Detecting Anomalies in Bipartite Graphs with Mutual Dependency Principles. In IEEE 12th International Conference on Data Mining (ICDM) 2012 (pp. 171–180). IEEE. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6413905
Dede, C. (2008). A seismic shift in epistemology. Educause Review, 43(3).
Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and Co-assessment in higher education: a review. Studies in Higher Education, 24(3), 331–350.
Dorst, K. (2003). The problem of design problems. Expertise in Design, 135–147.
Douceur, J. R. (2009). Paper rating vs. Paper ranking. ACM SIGOPS Operating Systems Review, 43(2), 117–121.
Duncan, T. E. (1999). An introduction to latent variable growth curve modeling: Concepts, issues, and applications. Mahwah, N.J.:L. Erlbaum Associates Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=19332.
Edwards, K. (2001). Epistemic communities. Situated Learning and Open Source Software Development. Epistemic Cultures and the Practice of Interdisciplinarity Retrieved from http://orbit.dtu.dk/fedora/objects/orbit:51813/datastreams/file_2976336/content.
Falchikov, N. (1986). Product comparisons and process benefits of collaborative peer group and self assessments. Assessment and Evaluation in Higher Education, 11(2), 146–166.
Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: a meta-analysis. Review of Educational Research, 59(4), 395–430. doi:10.3102/00346543059004395.
Ford, E., & Babik, D. (2013). Methods and Systems for Educational On-Line Methods.
Gagne, R. M. (1985). The conditions of learning and theory of instruction. Holt Rinehart & Winston.
Gersick, C. J. G. (1991). Revolutionary change theories: a multilevel exploration of the punctuated equilibrium paradigm. Academy of Management Review, 16(1), 10–36. doi:10.5465/AMR.1991.4278988.
Gillespie, A., & Cornish, F. (2010). Intersubjectivity: towards a dialogical analysis. Journal for the Theory of Social Behaviour, 40(1), 19–46. doi:10.1111/j.1468-5914.2009.00419.x.
Habermas, J. (1981). New social movements.
Hardaway, D. E., & Scamell, R. W. (2012). Open knowledge creation: bringing transparency and inclusiveness to the peer review process. MIS Quarterly, 36(2).
Hargrave, T. J., & Van de Ven, A. H. (2006). A collective action model of institutional innovation. Academy of Management Review, 31(4), 864–888.
Heersmink, R. (2013). A taxonomy of cognitive artifacts: function, information, and categories. Review of Philosophy and Psychology, 4(3), 465–481.
Hermida, A. (2011). Social media is inherently a system of peer evaluation [blog]. Retrieved September 21, 2012, from http://blogs.lse.ac.uk/impactofsocialsciences/2011/06/27/social-media-is-inherently-a-system-of-peer-evaluation-and-is-changing-the-way-scholars-disseminate-their-research-raising-questions-about-the-way-we-evaluate-academic-authority/.
Huhta, A. (2008). Diagnostic and formative assessment. In B. Spolsky, & F. M. Hult (Eds.), The handbook of educational linguistics (pp. 469–482). Blackwell Publishing Ltd. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470694138.ch33/summary.
Hu, M., Lim, E.-P., Sun, A., Lauw, H. W., & Vuong, B.-Q. (2007). Measuring article quality in Wikipedia: Models and evaluation. In Proceedings of the sixteenth ACM conference on information and knowledge management (pp. 243–252). New York: ACM. doi:10.1145/1321440.1321476.
Jones, B. L., Nagin, D. S., & Roeder, K. (2001). A SAS procedure based on mixture models for estimating developmental trajectories. Sociological Methods & Research, 29(3), 374–393. doi:10.1177/0049124101029003005.
Joordens, S., Desa, S., & Paré, D. (2009). The pedagogical anatomy of peer assessment: dissecting a peerScholar assignment. Journal of Systemics, Cybernetics & Informatics, 7(5) Retrieved from http://www.iiisci.org/journal/CV$/sci/pdfs/XE123VF.pdf.
Kamsu-Foguem, B., Tchuenté-Foguem, G., & Foguem, C. (2014). Using conceptual graphs for clinical guidelines representation and knowledge visualization. Information Systems Frontiers, 16(4), 571–589. doi:10.1007/s10796-012-9360-2.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, 90(430), 773–795. doi:10.1080/01621459.1995.10476572.
Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. Journal of the American Statistical Association, 90(431), 928–934. doi:10.1080/01621459.1995.10476592.
King, A. (1989). Verbal interaction and problem-solving within computer-assisted cooperative learning groups. Journal of Educational Computing Research, 5(1), 15.
Kirsh, D. (2010). Thinking with external representations. AI & SOCIETY, 25(4), 441–454.
Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. Journal of Experimental Psychology: Human Learning & Memory Journal of Experimental Psychology: Human Learning & Memory, 6(2), 107–118.
Kreps, D. M. (1997). Intrinsic motivation and extrinsic incentives. American Economic Review, 87(2), 359–364.
Krosnick, J. A. (1999). Maximizing questionnaire quality. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), Measures of political attitudes (pp. 37–57). San Diego, CA US: Academic Press.
Krosnick, J. A., Thomas, R., & Shaeffer, E. (2003). How Does Ranking Rate?: A Comparison of Ranking and Rating Tasks. In Conference Papers – American Association for Public Opinion Research (p. N.PAG).
Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one’s own incompetence lead to inflated self-assessments. Journal of Personality and Social Psychology, 77(6), 1121–1134. doi:10.1037/0022-3514.77.6.1121.
Kruger, J., & Dunning, D. (2002). Unskilled and unaware–but why? A reply to Krueger and Mueller. Journal of Personality and Social Psychology, 82(2), 189–192.
Lauw, H. W., Lim, E.-P., & Wang, K. (2006). Bias and Controversy: Beyond the Statistical Deviation. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 625–630). ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1150478
Lauw, H. W., Lim, E.-P., & Wang, K. (2008). Bias and controversy in evaluation systems. IEEE Transactions on Knowledge and Data Engineering, 20(11), 1490–1504.
Leite, W. L., & Stapleton, L. M. (2011). Detecting growth shape misspecifications in latent growth models: an evaluation of fit indexes. The Journal of Experimental Education, 79(4), 361–381. doi:10.1080/00220973.2010.509369.
Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. OBHP Organizational Behavior and Human Performance, 26(2), 149–171.
Lindblom-ylänne, S., Pihlajamäki, H., & Kotkas, T. (2006). Self-, peer- and teacher-assessment of student essays. Active Learning in Higher Education, 7(1), 51–62.
Lin, S. S.., Liu, E. Z.., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. Journal of Computer Assisted Learning, 17(4), 420–432. doi:10.1046/j.0266-4909.2001.00198.x
Markus, M. L., & Robey, D. (1988). Information technology and organizational change: causal structure in theory and research. Management Science Management Science, 34(5), 583–598.
Matusov, E. (1996). Intersubjectivity without agreement. Mind, Culture, and Activity, 3(1), 25–45. doi:10.1207/s15327884mca0301_4.

McKenzie, C. R. (1997). Underweighting alternatives and overconfidence. YOBHD Organizational Behavior and Human Decision Processes, 71(2), 141–160.
Meyer, B. D. (1995). Natural and quasi-experiments in economics. Journal of Business & Economic Statistics, 13(2).
Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 63(2), 81.
Miranda, S. M., & Saunders, C. S. (2003). The social construction of meaning: an alternative perspective on information sharing. Information Systems Research, 14(1), 87–106.
Mizzaro, S. (2003). Quality control in scholarly publishing: a new proposal. Journal of the American Society for Information Science and Technology, 54(11), 989–1005.
Norman, D. A. (1992). Design principles for cognitive artifacts. Research in Engineering Design Research in Engineering Design, 4(1), 43–50.
Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. Assessment and Evaluation in Higher Education, 25(1), 23–38. doi:10.1080/02602930050025006.
Piaget, J., & Gabain, M. (1926). The language and thought of the child, by Jean Piaget...Preface by Professor E. Claparède. London, K. Paul, Trench, Trubner & co., ltd.; New York, Harcourt Brace & company, inc., 1926.
Pulford, B. D., & Colman, A. M. (1997). Overconfidence: feedback and item difficulty effects. PAID Personality and Individual Differences, 23(1), 125–133.
Raman, K., & Joachims, T. (2014). Methods for Ordinal Peer Grading. arXiv:1404.3656 [cs]. Retrieved from http://arxiv.org/abs/1404.3656
Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169. doi:10.1007/BF01405730.
Roos, M., Rothe, J., Rudolph, J., Scheuermann, B., & Stoyan, D. (2012). A Statistical Approach to Calibrating the Scores of Biased Reviewers: The Linear vs. the Nonlinear Model. In Proceedings of the 6th Multidisciplinary Workshop on Advances in Preference Handling. Retrieved from http://mpref2012.lip6.fr/proceedings/RoosMPREF2012.pdf
Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? Journal of Economic Psychology, 33(5), 1012–1031. doi:10.1016/j.joep.2012.06.003.
Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. Educational Assessment, 11(1), 1–31.
Sage, A. P., & Rouse, W. B. (1999). Information systems frontiers in knowledge management. Information Systems Frontiers, 1(3), 205–219.
Salazar-Torres, G., Colombo, E., Da Silva, F. S. C., Noriega, C. A., & Bandini, S. (2008). Design issues for knowledge artifacts. Knowledge-Based Systems, 21(8), 856–867. doi:10.1016/j.knosys.2008.03.058.
Sargeant, J., Mann, K., van der Vleuten, C., & Metsemakers, J. (2008). “Directed” self-assessment: practice and feedback within a social context. Journal of Continuing Education in the Health Professions, 28(1), 47–54.
Scheff, T. J. (2006). Goffman unbound!: A New paradigm for social science. Boulder, Colo.:Paradigm Publishers.
Schleicher, D. J., Bull, R. A., & Green, S. G. (2008). Rater reactions to forced distribution rating systems. Journal of Management, 35(4), 899–927. doi:10.1177/0149206307312514.

Schutz, A. (1967). The Phenomenology of the Social World. Translated by George Walsh and Frederick Lehnert. With an introd. by George Walsh. Evanston, Ill.:Northwestern University Press.
Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464. doi:10.1214/aos/1176344136.
Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013). A case for ordinal peer evaluation in MOOCs. NIPS Workshop on Data Driven Education Retrieved from http://lytics.stanford.edu/datadriveneducation/papers/shahetal.pdf.
Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. YOBHD Organizational Behavior and Human Decision Processes, 42(3), 271–283.
Shepard, L. A. (2007). Formative assessment: Caveat emptor. Erlbaum.
Sieck, W. R., & Arkes, H. R. (2005). The recalcitrance of overconfidence and its contribution to decision aid neglect. J. Behav. Decis. Making Journal of Behavioral Decision Making, 18(1), 29–53.
Sieck, W. R., Merkle, E. C., & Van Zandt, T. (2007). Option fixation: a cognitive contributor to overconfidence. Organizational Behavior and Human Decision Processes, 103(1), 68–83. doi:10.1016/j.obhdp.2006.11.001.
Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. The American Economic Review, 49(3), 253–283.
Simon, H. A. (1969). The sciences of the artificial. The MIT Press.
Simon, H. A. (1973). The structure of ill-structured problems. ARTINT Artificial Intelligence, 4(3), 181–201.
Simon, H. A. (1979). Information processing models of cognition. Annual Review of Psychology, 30(1), 363–396.
Slavin, R. E. (1992). When and why does cooperative learning increase academic achievement? Theoretical and empirical perspectives:Cambridge University Press.
Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J. J. G., & Bastiaens, T. J. (2002). The training of peer assessment skills to promote the development of reflection skills in teacher education. Studies in Educational Evaluation, 29(1), 23–42. doi:10.1016/S0191-491X(03)90003-4.
Sluijsmans, D., Brand-Gruwel, S., van Merrienboer, J., & Martens, R. (2004). Training teachers in peer-assessment skills: effects on performance and perceptions. Innovations in Education and Teaching International, 41(1), 59–78.
Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. Psychological Review, 112(4), 881–911.
Stone, E. R., & Opel, R. B. (2000). Training to improve calibration and discrimination: the effects of performance and environmental feedback. YOBHD Organizational Behavior and Human Decision Processes, 83(2), 282–309.
Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the Few and How collective wisdom shapes business, economies, societies, and nations. New York:Doubleday.
Sutton, D. C. (2001). What is knowledge and can it be managed? European Journal of Information Systems, 10(2), 80–88.

Taras, M. (2002). Using assessment for learning and learning from assessment. Assessment & Evaluation in Higher Education, 27(6), 501–510. doi:10.1080/0260293022000020273.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. Review of Educational Research, 68(3), 249–276. doi:10.3102/00346543068003249.

Topping, K. J. (2005). Trends in peer learning. Educational Psychology, 25(6), 631–645. doi:10.1080/01443410500345172.

Tung, W.-F., Yuan, S.-T., Wu, Y.-C., & Hung, P. (2014). Collaborative service system design for music content creation. Information Systems Frontiers, 16(2), 291–302. doi:10.1007/s10796-012-9346-0.

Uebersax, J. S. (1988). Validity inferences from interobserver agreement. Psychological Bulletin, 104(3), 405–416.

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: the influence of interpersonal variables and structural features. Educational Research Review, 4(1), 41–54. doi:10.1016/j.edurev.2008.11.002.

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. Learning and Instruction, 20(4), 280–290. doi:10.1016/j.learninstruc.2009.08.010.

Walsham, G. (2006). Doing interpretive research. European Journal of Information Systems, 15(3), 320–330.

Wang, M. (2011). Integrating organizational, social, and individual perspectives in web 2.0-based workplace E-learning. Information Systems Frontiers, 13(2), 191–205. doi:10.1007/s10796-009-9191-y.

Wang, S.-L., & Wu, P.-Y. (2008). The role of feedback and self-efficacy on web-based learning: the social cognitive perspective. Computers & Education, 51(4), 1589–1598. doi:10.1016/j.compedu.2008.03.004.

Yu, F. Y., Liu, Y. H., & Chan, T. W. (2005). A web-based learning system for question-posing and peer assessment. Innovations in Education and Teaching International, 42(4), 337–348.

Zheng, Z. (. E.)., Pavlou, P. A., & Gu, B. (2014). Latent growth modeling for information systems: theoretical extensions and practical applications. Information Systems Research, 25(3), 547–568. doi:10.1287/isre.2014.0528.

Zigurs, I., & Buckland, B. K. (1998). A theory of task: technology fit and group support systems effectiveness. MIS Quarterly, 22(3), 313–334.