BABIK, DMYTRO, Ph.D. Investigating Intersubjectivity in Peer-Review-Based, Technology-Enabled Knowledge Creation and Refinement Social Systems. (2015) Directed by Dr. Rahul Singh. 211 pp.

In peer-based knowledge creation domains, problem complexity and subjectivity of individual understanding impedes development of actors' competencies. Prior research remains ambivalent on whether interactions between peers lead to the development of shared, intersubjective, understanding about one's own and peers' competencies. On the one hand, actors may develop this shared understanding through social learning. On the other hand, due to the Dunning–Kruger effect, both less and more competent actors may persistently miscalibrate their own performance relative to peers.

This dissertation examines how creation and evaluation competencies in peer-based social knowledge creation communities, where complex-problem social knowledge artifacts are produced, change and interact over time. It hypothesizes the existence of latent classes of longitudinal trajectories of creation and evaluation competency development, and convergence of these trajectories over multiple interactions, as intersubjective understanding emerges; moreover, their trajectories may be affected by the openness of peer groups.

To investigate this research problem, a peer review system was designed, instantiated, and tested in a controlled experiment study. Findings support the existence of multiple latent longitudinal trajectories. Partial evidence of the peer group openness' effect on competency change over time was also found. Results indicate that longitudinal peer interaction patterns are very complex. Practical implications of these finding for various domains are discussed and directions for further investigation are proposed.

INVESTIGATING INTERSUBJECTIVITY IN PEER-REVIEW-BASED, TECHNOLOGY-ENABLED KNOWLEDGE CREATION AND REFINEMENT SOCIAL SYSTEMS

by

Dmytro Babik

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2015

Approved by

Rahul Singh, Ph.D.
Committee Chair

To the memory of the *Nebesna Sotnya*,

the fallen *Cyborgs*, all men and women,

who gave their lives to free Ukraine.

APPROVAL PAGE

This dissertation written by Dmytro Babik has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair Rahul Singh

Committee Members Aprille N. Black

Eric W. Ford

Christian D. Schunn

Xia Zhao

March 18, 2015
Date of Acceptance by Committee

March 18, 2015
Date of Final Oral Examination

ACKNOWLEDGEMENTS

This research tremendously benefited from the help and advice of my Dissertation Committee. I am very grateful to Dr. Rahul Singh for continued guidance in developing my systems thinking and conducting research; Dr. Christian Schunn for suggesting alternative perspectives and encouraging rigor; Dr. Xia Zhao for examining my mathematical apparatus and assuring its thoroughness; Dr. Aprille Black for helping me bridge my research to pedagogical theories and practical applications in education; Dr. Eric Ford for the partnership in developing Mobius SLIP, offering a 10,000-feet view of the research problem and pointing out exciting future research opportunities.

I would like to thank my colleagues and faculty at the Department of Information Systems and Supply Chain Management, Bryan School of Business and Economics, UNCG at large, and other institutions, with whom I have had an opportunity to collaborate during my doctoral study.

My special thanks are to my wife Iryna Babik who has been not only a supportive spouse but also a great research partner and consultant; Alexey and Irene Babik for enthusiastically and patiently working with me on Mobius SLIP, which played a fundamental role in this research. Without the support of my family and friends at Kharkiv Polytechnic Institute, CERGE-EI, Tulane University and outside academia, I would not have achieved this important milestone in my life.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Research Problem

Peer-based creation, review, and evaluation of knowledge artifacts (KAs) representing solutions to intricate or complex problems have been a common practice in multiple domains, including research, education, knowledge management, open-source innovation, and social media. Given complexity of such problems, multiple peer reviews offer several advantages over single-expert evaluations. First, individual's expertise limitations and subjectivity make reviews by multiple peers a more dependable and trustworthy source of evaluation and feedback vis-à-vis a single-expert's review (Reily, Finnerty, & Terveen, 2009). Second, solving complex problems takes higher-level cognitive skills, heuristic reasoning and judgment competencies, development of which requires observing and imitating (Bandura, 1962; Miller & Dollard, 1941), practice ("learning by doing") (Simon, 1969), and feedback from multiple sources (Prins, 2006). Third, through the practice of having to evaluate and being evaluated, actors develop shared meaning, norms, and expectations (Sutton, 2001). Finally, in certain learning and expertise development settings, such as, Massive Open Online Courses (MOOCs), evaluations and feedback from a single expert may not be feasible, and peer evaluations become the only efficient and scalable solution (Raman & Joachims, 2014). Thus, peer-based creation and evaluation play an important role in the knowledge systems.

Peer evaluation is gaining in popularity and prevalence within and beyond research and education thanks to modern technologies. Rapid advances in web 2.0-based social media have made creating, distributing, and evaluating digital KAs cheaper and easier and, consequently, more socially engaging and beneficial to participants. Actors in various online social groups create KAs to represent and share information and knowledge (Salazar-Torres, Colombo, Da Silva, Noriega, & Bandini, 2008). Digitally published articles, outlining current knowledge about a topic, are now co-created and co-evaluated by self-organized technology-enabled social knowledge creation communities (SKCC) (Dede, 2008). People gain quick access to rich information sources through social media, such as wikis, blogs and other public shared knowledge sites (e.g., eHow.com and About.com), and social networks, such as Facebook, LinkedIn and ResearchGate. Actors also use and evaluate KAs to pass along their insights on the degree to which these artifacts are useful and how they may be applied. Through evaluating peers' creations and being evaluated by peers, actors learn, advance their collective understanding of various topics, and improve their KAs. Such social knowledge creation has received increasing research attention under labels, such as "open innovation" (Chesbrough, Vanhaverbeke, & West, 2006), "commons-based peer production" (Benkler, 2006), "open access movement" (Hardaway & Scamell, 2012), and "the wisdom of crowds" (Surowiecki, 2004).

The very "openness" of SKCC, however, raises fundamental concerns about the validity, accuracy and utility of socially produced and socially evaluated knowledge artifacts. In particular, the interpretation of facts, opinions, values and beliefs contributed to the solution and that may or may not be shared among the community are open to question. A diligent user should be discerning because the quality, reliability, validity and usefulness of these KA varies. Whether a KA possesses these desired properties depends on the creator's competency and stance. Any communication of knowledge is made with an intention or purpose and at times may be misleading or malicious. On the other hand, the judgment of whether a KA is trustworthy or valuable also depends on the subjective characteristics of the evaluator such as domain competency, positionality, and intentionality (Kruglanski, 1989; Tetreault, 2012). "An intrinsic feature of intentionality is that it is 'aspectual', i.e., always from a perspective, 'point of view' or focus of interest" (Searle, 1992, p. 131). Interpretation of facts, opinions, values and beliefs, may not be shared by all actors in the community due to their varying backgrounds (Searle, 1992; Sutton, 2001; Walsham, 2006). Subjectivity in solving complex problem presents, on the one hand, a threat to peer-based knowledge creation, evaluation and refinement, and on the other hand, an opportunity for researchers.

Although the research on the phenomena occurring in peer evaluations has grown since the 1990s, many research problems in this area remain under-investigated (Dochy, Segers, & Sluijsmans, 1999; Topping, 1998, 2005, 2009). For example, there are problems of correctly determining the KA quality from multiple peer evaluations (Hardaway & Scamell, 2012) and improving KA quality in non-expert peer-based knowledge refinement (Cho, Chung, King, & Schunn, 2008; Cho & Schunn, 2007), assessing creative and evaluative competencies of contributors to social knowledge bases (Cusinato, Della Mea, Di Salvatore, & Mizzaro, 2009; Mizzaro, 2003), inferring evaluation validity from reliability (Uebersax, 1988), accounting for evaluator biases (Lee & Schunn, 2011; Piramuthu, Kapoor, Zhou, & Mauw, 2012), and improving competency measurement accuracy (Douceur, 2009; Raman & Joachims, 2014; Shah, Bradley, Parekh, Wainwright, & Ramchandran, 2013). These problems have stimulated a recent wave of research publications and design patents, indicating high demand for more accurate and convenient peer evaluation systems in various areas.

Previous research has focused primarily on socially constructed knowledge in technology-mediated, web-based communities (Brown, Collins, & Duguid, 1989; Miranda & Saunders, 2003). Very little research in the IS field has examined the intersubjective nature of competency development and knowledge refinement in the context of peer evaluation in SKCC. However, the question of how peer-evaluation interactions affect creation and evaluation competencies and KA qualities is important and under-investigated. Addressing this research problem from the SKCC perspective may provide better understanding of social knowledge creation dynamics, individual competency development, and artifact improvement. Further, it will inform better designs for peer-based technology-enabled knowledge creation and refinement systems.

In this dissertation, the changes and interactions of competencies are viewed through the lens of intersubjectivity. This research adopts an inclusive definition of intersubjectivity as the variety of relations between different perspectives (Gillespie & Cornish, 2010; Matusov, 1996). Intersubjectivity is viewed as both a process and an outcome of interactions among actors' perspectives and perceptions about KAs. Creation and evaluation competencies are theorized to evolve through intersubjective perspective taking in recurrent creator-evaluator interactions. Through these interactions, actors may achieve better understanding of other actors' views and expectations and, thus, become more aware of environmental demands. Actors may agree or disagree with these demands, but they are likely to adjust their approach to creating new artifacts and evaluating artifacts of others in the future. Consequently, their competencies change during peer evaluation process, which, in turn, may lead to changes in qualities of future KAs. This dissertation investigates whether actors' perceptions of their creation and evaluation competencies tend to converge towards perceptions of peer and expert evaluators. This research systematically explores general patterns of how actors in SKCC develop their creation and evaluation competencies, how these competencies interact, and how this process affects the creation and quality properties of new KAs' creation.

This dissertation investigates changes and interactions of competencies in SKCC that rely on intersubjective processes of peer evaluations in expertise development and knowledge building. Due to creators' and evaluators' subjective positions in addressing complex problems, peer evaluations are inherently intersubjective. Intersubjectivity as a process involves posing, framing, and exchanging ideas through the social discourse by sharing of KAs and receiving reactions from others (Hargrave & Van de Ven, 2006). Intersubjectivity as an outcome is a temporary settlement in the social discourse that emerges as mutual awareness of agreement or disagreement among actors and acknowledgement of multiple perspectives (Hargrave & Van de Ven, 2006; Matusov, 1996). Intersubjectivity exhibits situated, interactional and performative nature (Schegloff, 1982). Therefore, intersubjectivity is dynamic, meaning that over multiple interactions and under actors’ mutual influence, their perspectives may shift and competencies may change.

Research Questions

The purpose of this dissertation is to design an analytical method for exploring the phenomenon of intersubjectivity in peer-based knowledge creation, evaluation and refinement social system and to test its utility in a controlled experimental setting. This research explores the dynamics and interactions of the two types of competencies actors need to solve complex open-ended problems in SKCC – KA creation and evaluation. In this context, interactions mean information exchanges between creators and evaluators. Dynamics mean systematic changes of competencies of creators and evaluators over multiple interactions. Very simply, a creator presents a KA for evaluation or an evaluator provides feedback to a creator. These acts constitute information exchanges, that is, interactions between the creator and the evaluator. As these interactions repeat multiple times between steady or varying creator-evaluator dyads of actors, according to the social learning theory, their competencies change in a systematic manner (Bandura, 1962; Bigge & Shermis, 2003).

The dynamics and interactions of the two competency types are important for the following reason. Firstly, in a SKCC, to ensure a meaningful and sustainable discourse between a creator of a KA and the user audience, actors have to possess both types of competencies. That is, to create a KA valuable for the audience, a creator has to show certain competency and awareness of his audience’s needs and preferences. Furthermore, to make a meaningful and valid evaluation of other creator's artifacts, a user actor has to be a competent evaluator. However, anecdotally, not all creators are equally skillful in both types of competencies, i.e., not all creators are dependable evaluators and, vice-versa, not all evaluators are ingenious creators. Since peer reviews enable information exchange among actors, as well as learning and competency building through social interactions, the peer-based SKCC is studied as a social system.

Social cognitive theory suggests that social learning occurs over time with repeated social interactions between creators and evaluators (Bandura, 1962; Bigge & Shermis, 2003). Cognitive learning is promoted through modeling and performance feedback. This learning is expected to result in advancement of both types of competencies, as well as in creation of new or improved knowledge artifacts. Mastering these two types of competencies, however, requires different cognitive abilities – producing a new artifact as a solution to a problem demands creation and synthesis, whereas evaluation of the artifact requires critical thinking and analysis (Anderson et al., 2001; Bloom, Krathwohl, & Masia, 1956). Furthermore, different cognitive abilities, such as problem solving and critical thinking, may exert mutual influence. Finally, individual barriers to learning, such as personal and social biases or personality traits, may affect individual learning trajectories for creation and evaluation competencies. Inevitably, these competency dynamics affect KA development and intersubjective evaluations in SKCC.

This dissertation addresses the following research questions (RQs):

RQ 1: How do creation and evaluation competencies change and interact over multiple creator-evaluator interactions in peer-based Social Knowledge Creation Communities (SKCC)?

RQ 2: How do competency dynamics impact the Social Knowledge Artifact (SKA)?

To answer these research questions, first, the key concepts of the domain of interest are defined and a theoretical foundation for researching dynamics and interactions of creation and evaluation competencies through the lens of actors' intersubjective mutual assessments of original KA creations and peer critiques is built. The definitions of SKCC and SKA are developed based in these key concepts. Then, an analytical method for exploring the phenomenon of intersubjectivity in SKCC is designed, hypotheses that can be tested using this system are formulated, and an empirical study to test these hypotheses and to explain how competencies change, interact, and impact KAs in SKCC has been conducted.

Definitions

Knowledge Artifacts

To be communicated between actors, knowledge has to be explicated, i.e., codified and embodied in an object that holds and conveys usable representation of knowledge (Holsapple & Joshi, 2001; Nonaka & Von Krogh, 2009). Based on existing research literature, such an object is defined as a *knowledge artifact* (KA) (Salazar-Torres et al., 2008). KAs are creations that social groups use to represent, transmit, store, and share knowledge. The notion of KAs is based on the notion of cognitive artifacts – things that help people 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 representation of knowledge, such as cognitive artifacts, have multiple distributed cognition benefits and provide structured shareable referents to coordinate thought (Kirsh, 2010; Sutton, 2001). KAs facilitate information sharing in SKCC where KA creators, evaluators, and users collaboratively learn and develop the collective understanding of a topic (Salazar-Torres et al., 2008).

Complex Open-ended Problems

This work is concerned with evaluations of KAs that emerge as outcomes of *complex tasks* or solutions to *open-ended problems*, also labeled as 'ill-defined', 'ill-structured', or 'wicked problems' (Lynch, Ashley, Alven, & Pinkwart, 2006; Rittel & Webber, 1973). To define the concept of complex task, let's begin with the notion of a simple task. Simple tasks have a simple desired outcome, a single solution scheme, and no conflicting interdependence or solution scheme/outcome uncertainty (Zigurs & Buckland, 1998). More specifically, a simple task, as a well-structured problem, is characterized by the following properties (Voss, 2005, p. 322):

> (1) The goal is well-defined, and generally the solution is agreed upon by the members of the respective community.

> (2) Constraints are usually stated in the problem statement or are readily apparent.
> (3) Operators are frequently mathematical, logic-based, or in the case of some games, object moves.
> (4) The problem lends itself to computer simulation, because the number of states, the constraints and the operators are readily within computer simulation capabilities.

In contrast, complex tasks are characterized by various combinations of complexity attributes, such as outcome multiplicity, solution scheme multiplicity, conflicting interdependence, and solution scheme/outcome uncertainty (Campbell, 1988). Producing research and publishing academic articles, architectural plans, designs, laws, pieces of visual and language art, creating compositions are but a few examples of complex tasks from various domains. Complexity of a task results from a combination of the following characteristics (Voss, 2005, p. 323):

> (1) The goal is vaguely stated, and requires analysis and refinement in order to make the particular issue tractable.
> (2) The constraints of the problem typically are not in the problem statement; instead, the solver needs to retrieve and examine the constraints when appropriate during the solving process.
> (3) In most cases, the solver's solution is divided into a representation and a solution phase, as previously discussed. However, in contrast to well-structured problems, different solvers may vary considerably in the nature and contents of each of the phases. This is because ill-structured problems may be approached in different ways, according to the solver's knowledge, beliefs, and attitudes.
> (4) Solutions to ill-structured problems typically are not right or wrong, and not valid or invalid; instead, solutions usually are regarded in terms of some level of plausibility or acceptability. Furthermore, solution evaluation may be a function of the evaluator's knowledge and beliefs regarding the issue at hand.
> (5) When a solution is stated, it usually is justified by verbal argument that indicates why the solution will work as well as providing a rebuttal by attacking a particular constraint or barrier to the solution or by attempting to refute an anticipated opposing position. The solver's definition of the problem in the representation phase and presentation and justification of its solution demonstrate that this solution process is rhetorical in nature.

> (6) The solutions of ill-structured problems often are not 'final', in the sense that … the problem asked for is 'solved', but to know if it would 'really work' would require implementation and subsequent evaluation. When to terminate discussion of the solution is thus somewhat arbitrary.
> (7) The size of the database required for most ill-structured problems and the difficulties in accessing it make simulation difficult.

Solving such a class of problems, which for the purpose of this dissertation are called *complex open-ended problems*, requires framing (e.g., characteristics 2 and 3); solutions are often expressed as narrative (characteristics 5 and 6); and no simple deterministic algorithms exist for solving 'correctly' an open-ended problem, only heuristic techniques are available to guide the analysis (characteristics 1, 6, and 7) (Goldin, 2011). These aspects of complex open-ended problems present challenges to acquiring skills and competencies required to solve them, as well as to evaluate the quality of a solution and the competency of a solving actor.

Competencies

For the purpose of this dissertation, *competency* is defined as an actor's quality of being competent, i.e., being in possession of skills, knowledge, qualification or capacity of performing a task or solving a problem ("The Definition of Competence," n.d.), or, in simple words, the ability to do something well ("Competence," n.d.). This work is specifically concerned with the competencies of dealing with *complex open-ended problems.* Solving this kind of problems, as well as evaluating solutions or outcomes, requires heuristics approach, practice, and judgment, i.e., actual competence in a given field (Sutton, 2001). Solutions to such problems cannot be evaluated using objective, non-contradicting criteria and, therefore, for the purpose of evaluation, are usually subjected to multiple peer reviews.

This dissertation is concerned with two types of competencies – creation and evaluation. *Creation competency* represents creator's ability to produce a KA that possesses certain properties valuable or desirable to a certain audience of evaluators. *Evaluation competency* means evaluator's ability to accurately recognize the lack of those valuable or desirable properties of KAs and to provide constructive and actionable critique to the creator that can be used to improve the KA or produce a better new KA. Creation competencies are needed to produce KAs and to develop evaluation competencies needed to recognize a conceptualization of a complex open-ended problem that is not necessarily "correct", but that could be understood by all involved actors and that would facilitate action toward solving the problem rather than paralyze it (Simon, 1969).

Attainment, Refinement, Learning

Attainment reflects success in achieving something. *Artifact attainment* is the degree or level to which an actor succeeded in solving a particular open-ended problem or performing a specific complex task. In other words, artifact attainment reflects the degree to which a KA possesses some properties or values desired by the creator and users, such as efficacy, verity, accuracy, utility or style (Dorst, 2003; Simon, 1969). *Actor attainment* reflects the actor's success in achieving a certain level of competency of solving an open-ended problem or performing a complex task. Here, attainment is defined as a concept of competency achievement. In the operationalized research model, the term *attainment* refers to the construct of *goodness* as the holistic recognition of the desired artifact properties.

Technically speaking, actor attainment is latent, i.e., not directly observable or measurable, but is evidenced through the attainment of the artifacts produced by the actor. If the attainment of artifacts produced by a given actor is regarded as high, one can conclude that the attainment of the actor with respect to a particular competency (or set of competencies) needed to produce this artifact is also high. In other words, it can be assumed that actor's latent attainment is highly correlated with the observable (but subjective) attainment of the actor's artifacts.

Refinement refers to a perceived positive change in artifact attainment that may occur as a result of practice or additional work done on the artifact (note that in this context, artifact means not only a particular tangible object, such as an artwork piece, but also a method or an approach to solving a particular problem or performing a task. It is in the latter that the artifact attainment translates into actor attainment. Furthermore, *learning* is defined here as a favorable change in an actor's competency. For example, if an actor is able to perform a task better in the later instance than in an earlier instance, one may conclude that learning occurred between the two instances. *Learning* of complex task competencies is a favorable change in actor's attainment, and due to the fact that actor's attainment is latent, learning can be recognized only through the improvement of the artifact attainment over multiple performances.

Evaluation

Attainment is recognized through the act of *evaluation*. *Evaluation* is the process of assessing, i.e., analyzing, judging and documenting artifact's or actor's attainment. In some literature streams, evaluation is referred to as *assessment*; for consistency the former definition will be used throughout this dissertation, but where necessary to refer to the literature, these two terms will be considered synonyms. More specifically, this dissertation is concerned with *explicit evaluation*, i.e., assessment made by the evaluating actor in the form of documented mark, score or comment made by the actor to indicate his perception of the artifact's properties. In the psychology and education literature, explicit evaluation, or assessment, is usually categorized as either formative or summative (Black & Wiliam, 1998). The purpose of *formative assessment* is to improve actor's competency (i.e., to enhance the ability to reach a higher attainment) through behavior modification in the learning process (Crooks, 2001; Huhta, 2008). *Summative assessment* aims at monitoring learning outcomes and measuring and documenting attainment at a particular time (usually in the form of a score or a grade) (Shepard, 2007; Topping, 2009). Since the artifact attainment improvement is a learning process, the notion of formative and summative assessments can be applied more broadly to knowledge artifact development in SKCC.

For the purpose of this dissertation, three kinds of evaluations will be considered – peer, expert and self-evaluation. *Peer evaluation* (or *peer assessment*) is defined 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" (Topping, 1998, p. 250). Depending on its purpose, peer assessment may be either formative or summative. *Peer review* is referred to as a process and a product of providing a combination of formative assessment (critique) and summative assessment (benchmarking) by an evaluator to a creator. *Expert evaluation* is defined as an evaluation of an artifact by an actor of demonstrably higher competency or status than that of the artifact creator. In the education context, for example, an instructor will be considered the expert, whereas students will be considered peers to each other. *Self-evaluation* refers to the evaluation of an artifact by its own creator.

Bias and Controversy

Given the inevitable subjectivity of evaluations of complex problem solutions, *divergence of evaluations* a KA's attainment among peers is an interesting and important aspect of the evaluation intersubjectivity phenomenon (Gillespie & Cornish, 2010). Such divergence is manifested in two concepts: (1) a deviation of a given actor's evaluations from other actors' evaluations of the same set of KAs; following Lauw, Lim and Wang (2008), this deviation is referred to as evaluator *bias*; (2) an overall spread of evaluations of the same KA, which is referred to as artifact *controversy* (Lauw et al., 2008). This notions are inevitably present in evaluation systems, where evaluating actors have incomplete information about evaluated objects, i.e., where information asymmetry is present (Akerlof, 1970), or where evaluation are highly subjective due to complexity of evaluated objects, limited competencies of evaluating actors or social biases (Lauw et al., 2008; Lee & Schunn, 2011). Controversy and bias are closely coupled and inter-related concepts of divergence in perceptions and evaluations. *Controversy* of a KA captures a degree of *divergence* of evaluations among less biased evaluators; *bias* of an evaluator refers to a degree of *deviation* in the actor's evaluation from other actors' evaluators on less controversial objects. More controversial objects do not provide strong evidence of evaluator bias because higher controversy implies significant biases of all evaluators (Lauw et al., 2008).

Miscalibration, Overconfidence and Underconfidence

Another important aspect of evaluation is the alignment of creator's perception of their own KA with the perceptions of peer evaluators. For the purpose of this study, *miscalibration* refers to the dissimilarity of the creator's perception of own artifact attainment (i.e., self-evaluation) and other evaluators' perceptions about it (external evaluation) (Kruger & Dunning, 1999, 2002; Sadler & Good, 2006; Sargeant, Mann, van der Vleuten, & Metsemakers, 2008). Following Kruger and Dunning (1999), if self-evaluation exceeds external evaluations, such miscalibration is called *overconfidence*. In contrast, if self-evaluation is lower than external evaluations, such miscalibration is referred to as *underconfidence*.

Thus, the key concepts that will be used to study dynamics and interactions of creation and evaluation competencies in peer-based SKCC are introduced and defined. Now the approach to studying the phenomenon of interest will be explained.

Approach

To explore the phenomenon of the evaluation intersubjectivity in peer-based knowledge creation and refinement and to answer the posed research questions, this dissertation aims to achieve the following objectives. First, it proposes a theoretical framework of assessing creation and evaluation competencies in peer-review-based knowledge creation and social learning environment. Second, based on this framework, the analytical method, called the Double-Loop Mutual Assessment, or DLMA, is designed and implemented as an instantiation of a IT-enabled self-regulated social learning system, called Mobius Social Leaning Interaction Platform (SLIP), to facilitate competency enhancement and KA refinement through anonymous distributed online peer reviews (evaluations and critiques) of digital KAs in educational setting (Ford & Babik, 2013). Third, the hypotheses about the longitudinal dynamics and interactions between creation and evaluation competencies and their impact on KAs are formulated. Forth, the utility of the designed system is evaluated by testing these hypotheses in a controlled-experiment quantitative empirical study.

In the empirical study, students taking a Systems Analysis and Design course complete a series of open-ended assignments leading to the design of an information system. Assignments include three tasks: (1) creation of system models/diagrams and their descriptions (submissions of KAs), (2) anonymous mutual peer reviews (critiques and evaluations), and (3) anonymous mutual evaluations of critiques. An interaction describes an instance of information exchange between actors such as (a) a creator makes available his KA submission to a peer reviewer; (b) a reviewer provides summative evaluation and formative critique to a creators' KA; and (c) a creator provides summative evaluation of a received critique. In this study, each KA was subjected to evaluation and critiquing by several reviewers; and each reviewer's critiques were evaluated by several respective creators. Thus, each actor engaged in several interactions simultaneously. Such arrangement models environments in which each creator's KA is evaluated by multiple peer reviewers, and each reviewer evaluates and provides critiques to multiple artifacts by various creators. Moreover, it permits collecting evaluation data for each information object (such as original KA submission or its critique) exchanged in the peer review process. Importantly, each actor concurrently operates as a creator and an evaluator. In this sense, these creator-evaluator interactions are both social and intersubjective, even though they occur in virtual and anonymous environment. These social intersubjective interactions lead to social construction of knowledge as a state of understanding of a particular design issue, shared by multiple actors (Miranda & Saunders, 2003) and reached by exchanging original KAs, its critiques (formative assessment), and quantitative evaluations (aggregate summative assessment) (Cho et al., 2008).

For the sake of clarity, henceforward, two types of KA are distinguished: (A) creators' submissions of the original KAs (solutions to the assignment problem), called for simplicity *Submissions*; and (B) critiques of the original KA submissions, called for short *Critiques*. Note that while a Submission produced by a creator and reviewed by a reviewer is the same single whole document object (KA), a set of *Critiques given* by each reviewer to creators and a set of *Critiques received* by each creator are different composite document objects that consist of several individual critiques originating from a dyadic interaction between a creator, whose KA submission is being reviewed and a reviewer, who critiques and revaluates the KA submission.

In each assignment, self-, peer, and expert (instructor) evaluations of the KAs, both Submissions and Critiques, were collected. From these data, for each KA, measures of attainment perceived by the creator, peer evaluators, and the expert were calculated. (The detailed algebraic description of the method of computing various DLMA measures are given in appendix A). These measures also indicate different actors' perceptions of creation competencies of the KA's creator. The Critiques attainment is computed similarly to the Submission attainment based on self-, peer, and expert evaluations. Peer-evaluation attainment of a KA is based on the aggregate evaluations by several reviewers of the KA submission. I.e., to compute attainment from peer evaluations, scores given by multiple peer reviewers to the same artifact are averaged. The higher score given to an artifact by a particular peer evaluator indicates a higher preference or judgment of that evaluator towards that artifact. Under low controversy, the higher aggregated attainment scores obtained from multiple peers indicate a positive consensus view of several evaluators of the artifact attainment. In other words, a high aggregate attainment score indicates stronger intersubjective convergence among peer evaluators about the high attainment of the artifact, whereas, a low aggregate attainment score indicates stronger intersubjective agreement among peer reviewers about the low attainment of the artifact.

Under certain (strong) assumptions, expert evaluations are treated as proxy measures of the "true" underlying latent attainment, or quality of the artifact. Although peer-evaluation attainment of a KA may not coincide with the expert evaluation, it reflects the view dominating among peers. The difference between external (peer or expert) evaluations and creator's self-evaluation of the KA attainment is a respective measure of miscalibration.

Measures of individual and group divergence in peer evaluations were also calculated. Controversy characterizes divergence of peer group in evaluating a particular KA. Low variance of evaluations received an individual KA (i.e., low controversy) indicates convergence among peers in assessing attainment of the KA, or, in other words, high intersubjective agreement. In contrast, high variance (i.e., high controversy) indicates divergence, intersubjective disagreement, or, said differently, the lack of common understanding among evaluators about the KA's attainment. Measures of controversy were calculated for each actor's Submissions and Critiques. Evaluator bias characterizes divergence of an individual evaluator's judgment from evaluations given by the rest of the peer group. High bias indicates that an evaluator does not share perceptions of the KA attainment with the rest of the peer group, whereas low (or no) bias indicates the alignment of individual evaluator's perceptions with those of the rest of the peer group. The combination of miscalibration, controversy, and bias measures shows whether actors reach understanding of creation and evaluation competencies, and whether they share this understanding with their peers.

The goal of the analyses in this dissertation is to identify latent classes of individual actors' longitudinal trajectories of miscalibration, bias and controversy and to find relationships between these trajectories. This approach provides insights into how creation and evaluation competencies change over time, and how they interact, that is, to answer the posed research questions. This approach also permits to control for factors that are extraneous to studying competency development through peer interactions, but that may impact student competencies due to their involvement in other learning activities. The focus here is on higher-level creation and evaluation competencies and artifact qualities that emerge through intersubjective social interactions.

To examine utility of the proposed analytical method in researching longitudinal competency and intersubjectivity dynamics, controlled experiment was used. The experiment followed the randomized complete block (RCB) design with repeated measures within subjects. The empirical study compares two conditions: in the control group, subjects were placed randomly in peer groups in each assignments, i.e., in each assignment each subject interacted with a new random group of peer reviewers; in the treatment group, subjects were placed randomly in peer groups in the first assignment and remained with the same peer group for all consecutive assignments, i.e., they gave and received critiques and evaluations to and from the same peers. Subjects' overall education level (undergraduate or graduate) was treated as a control variable. (Further details of the experiment design are given in subsection "Research Design").

Scope of Research

The objective of this work is to explore the longitudinal dynamics and interactions of specific aspects of creation and evaluation competencies (creator and artifact attainment, artifact controversy, evaluator bias and miscalibration) in the environment of

SKCC where KA attainment and actor's competency can be assessed only intersubjectively. To achieve this objective, the scope of this study includes (1) development of research framework; (2) development of the analytical method to gauge these aspects of intersubjectivity and its instantiation of a software implementation; (3) development of the research hypotheses about the relationships of these intersubjectivity aspects; (4) validation of utility of the analytical method and its instantiation through testing the hypotheses in the controlled experimental study. These competency dynamics are analyzed quantitatively by means of explicit evaluations and interpreted through the lens of social cognitive and economic psychology theories. This study does not intend, however, to address the psychological causes of any sources of subjectivity.

Dissertation Organization

This dissertation is organized as follows. Chapter 2 presents a theoretical foundation for studying dynamics and interactions of creation and evaluation competencies in multiple double-looped mutual peer evaluation and their impact on KA attainment. Chapter 3 presents research methodology. Chapter 4 reports analyses results. Chapter 5 summarizes and discusses research findings, theoretical and practical contributions, limitations of the study, and directions for future research.

Throughout this dissertation, the singular androgynous pronouns “he”, “his”, “him” are used to refer to actors, such as creators, evaluators, reviewers, students, experts, instructors, of both genders not on the grounds of personal preference of the author but for the sake of simplicity, clarity and space.

CHAPTER II

THEORETICAL FOUNDATION AND HYPOTHESES

This chapter presents the system model, the conceptual model and theoretical foundations of competency dynamics in Social Knowledge Creation Community (SKCC).

System Model

The aim of this dissertation is to explore the phenomena of competency development and the evaluation intersubjectivity in peer-based knowledge creation communities. For the purpose of the dissertation, SKCC is defined as a collective of actors who share the common interest in the same topical domain and the goal of advancing knowledge in this topical domain. Domains that involve solving complex open-ended problems are of particular interest in this study. Solving these problems, such as, for example, creating artwork, conducting scientific research, developing business strategy, or designing fashion, requires creation of complex objects or performance of complex tasks (Campbell, 1988).

Complex open-ended problems have two important characteristics central to the phenomena of interest. The first characteristic is that solving such problems requires heuristics, subjective judgment, and practice (Brown & Duguid, 2001; Polanyi, 2009; Sutton, 2001). Therefore, the competency of solving such problems, i.e., creation competency, cannot be acquired by learning correct answers, but only through multiple trials and errors, interactions with other creators, by receiving and processing feedback, and gaining experience. This competency building process is a process of cyclical interactions between the creator, the artifact, and the users of the artifact (who also act as explicit or implicit reviewers and evaluators). In this development process achieving current goals in creating the artifact leads to feedback from evaluators to the creator, and, in turn, feedback suggests new goals.

The second characteristic is that assessing the degree of success in solving complex open-ended problems can also only be achieved through the use of heuristics and subjective judgment. Moreover, subjective judgments dynamically change with the experiences of the actor. “Exposure to new experiences is almost certain to change the criteria of choice, and most human beings deliberately seek out such experiences” (Simon, 1969, p. 186). Any individual opinion about the qualities of a solution is susceptible to personal biases or limitation of the expertise. Therefore, for practical purposes, the assessment of solutions to the complex open-ended problems is usually conducted by several judges or reviewers or who oftentimes have an equal status with the creator, that is, are creator’s peers.

An important aspect of evaluating solutions to complex-open ended problems is providing feedback that helps creator learn from practice, improve creation competency, and generate better solutions. Because of multiple and possibly conflicting criteria of evaluation, providing such feedback, e.g. in the form of a review or critiques, is also a complex open-ended problem by itself. Hence, it inherently has the characteristics outlined in the previous paragraph. Typically, a review consists of two components: (a) a critique that highlights strength and weaknesses of the solutions according to certain explicit or implicit criteria and provides recommendations with the aim of improving the solution (formative assessment), and (b) an evaluation that indicates the degree of attaining a certain level of success or goodness.

In peer review systems used to evaluate solutions to complex open-ended problems and to help creators develop their competencies, an interaction usually occurs between a creator and several reviewers/evaluators. Therefore, peer review is fundamentally a social process (van Gennip, Segers, & Tillema, 2009). For that reason, hereafter, peer evaluation is modeled as an intersubjective social interaction in a social system. Since a social system manifests itself through social interactions between persons acting in their roles, such as actions or communications (Luhmann, 1995; Parsons, 1991; Viskovatoff, 1999), this notion presents a suitable way of modeling competency development in peer-based SKCC.

Each SKCC consists of a number of actors who work on solving a complex open-ended problem, generate new knowledge, and communicate it to other peer users of a KA. Actors review and evaluate each other’s KAs and communicate back their judgments and recommendations in the form of quantifiable evaluations (summative assessment) and critiques (formative assessment) respectively. Generally, due to various possible process constraints, each KA is reviewed and evaluated by only a few actors in SKCC, not by all actors in the topical domain community.

The review process, focused on a single KA, takes place during a certain finite period of time, as a single interaction between the creator and the evaluators. In the peer-based SKCC modeled in this research, a complete creator-evaluator interaction in the social subsystem consists of the following processes. A creator communicates his original KA submission to peer evaluators (Figure 1, Step 1). At this step, creation competency is exposed to evaluators. Each evaluator reviews the creator's original KA submission, provides a critique (formative assessment), makes judgment about the attainment of the KA and provides summative assessment (Figure 1, Step 2). At this step, evaluation of creation competency is recorded, and evaluation competency is exposed to creators. The creator receives and reflects upon formative assessment from peer evaluators, makes judgment about the attainment of peers' critiques and provides their summative assessment (Figure 1, Step 3). At this step, evaluation of evaluation competency is recorded.

Thus, overall, the SKCC at all times consists of multiple "creator-KA-evaluators" subsystems around specific KAs. Note that each actor, in general, may enter multiple subsystems at the same time as either a creator or an evaluator. All KAs are intended to develop knowledge and solve problems in the same topical domain.

Peer evaluation systems that include the third step where creators reciprocally evaluate reviewers' (evaluators') critiques emerge increasingly in academia and education (Cho & Schunn, 2007; Goldin, 2011; Hamer, Ma, & Kwong, 2005; Sitthiworachart & Joy, 2004). In this dissertation, they are categorized as Double-Looped Mutual Assessment (DLMA) systems. The two primary advantages of the DLMA

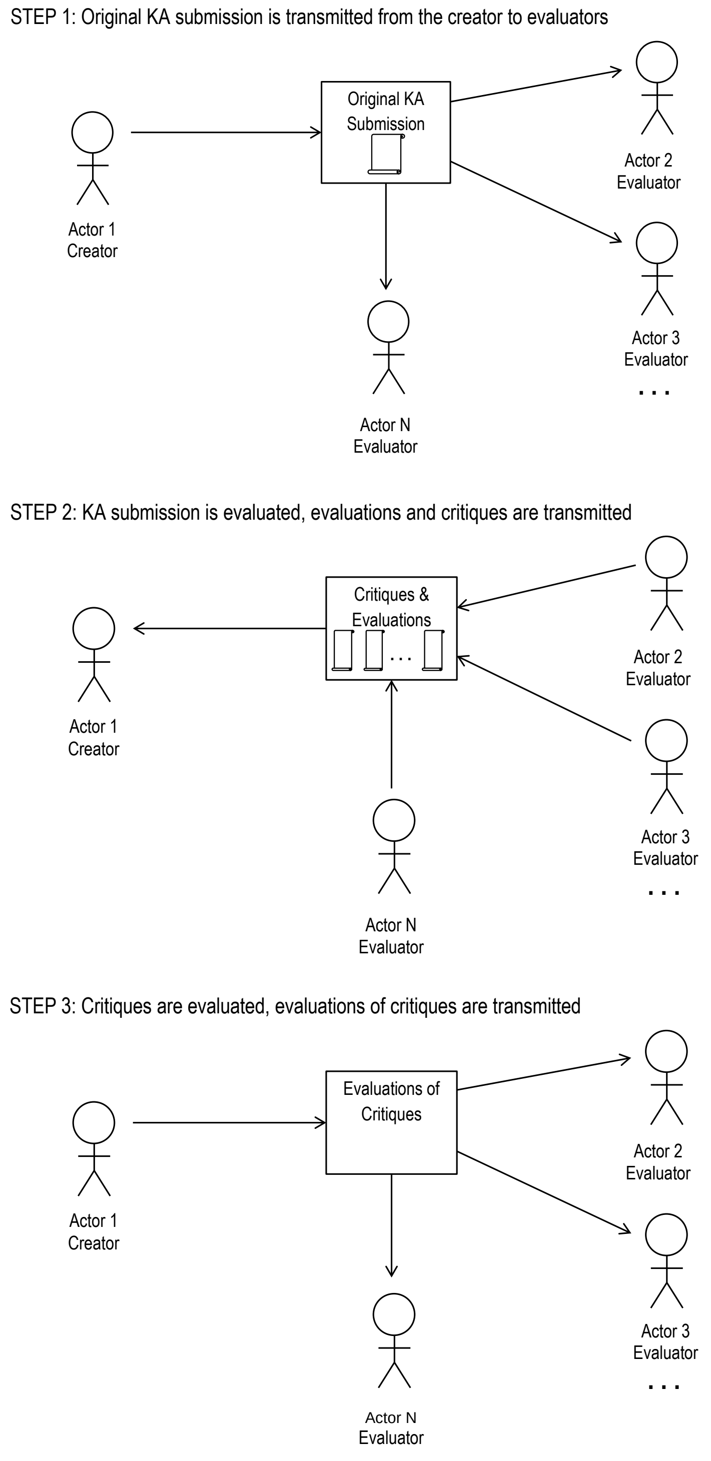


Figure 1. Knowledge Artifact Evaluation Interactions

systems over the single-looped peer evaluation systems are the following: (a) the fact that peer evaluators’ critiques are also evaluated by creators provides a stimulus for evaluators to give diligent evaluations and critiques; (b) the evaluations of evaluators’ critiques by creators provide a source of data on evaluators’ performance.

In addition to mutual evaluation of the original KAs and their critiques, the model presented in this dissertation also includes self-critiques and self-evaluations. In step 2, a creator self-evaluates his own original KA and formulates self-critiques. In step 3, an evaluator self-evaluates his critiques to the creator’s KA, i.e., makes judgment about its attainment.

The complete set of information exchanges of critiques and evaluations constitutes the core unit of creator-evaluator interaction around a single KA. It is a subsystem that consists of the KA, the creator of the KA, and several evaluators of the KA. Some practical details of this interaction unit are discussed in chapter III “Methodology”. As several actors may engage in solving the same complex open-ended problem at the same time, they may communicate their KA to each other and act as each other’s evaluators. This brings into existence a more complex social peer review interaction of actors, defined for the purpose of this dissertation as a Social Knowledge Artifact (SKA). SKA is modeled by superimposing several KAs that are solutions to the same complex open-ended problem, with their respective creator-evaluator interaction, so that creators evaluate each other’s and their own original KAs, provide critiques and self-critiques, and evaluate each other’s and their own critiques (Figure 2).

This dissertation suggests that SKA is the appropriate way of modeling and studying the phenomena of creation and evaluation competency development and the evaluation intersubjectivity in peer-based knowledge creation and refinement social systems because it represents a complete unit of knowledge and competency building, and captures all interactions necessary to model the domain and phenomena of interest. More specifically, it allows capturing attainment of KAs and critiques and, through them, attainment of creator and evaluator competencies.

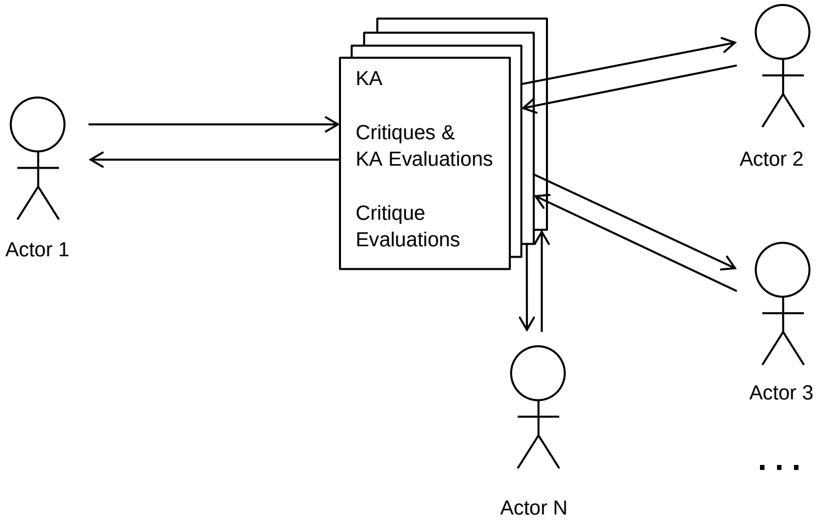


Figure 2. Social Knowledge Artifact and Social Knowledge Creation Community

Moreover, observing multiple instantiations of SKA produced by actors over time and measuring differences in attainment of KAs and critiques between different SKA at different times allows making inferences and epistemological claims about the system-level changes in creation and evaluation competencies of the actors in SKCC. If the change in creation and evaluation competencies over time is positive, one may conclude that the system produces the social learning effect.

Conceptual Model

To describe the system-level properties of SKA that result from the intersubjective interaction of information exchanges in the DLMA peer evaluations in a SKCC, the following conceptual model is proposed (Figure 3). In this model, intersubjectivity is conceptualized as a combination of concepts of miscalibration, controversy and bias surrounding evaluations of KAs within an SKA. Attainment is a concept that denotes the level of success of solving the problem by each of the KAs and derived from the systemic creator-evaluator interaction in the SKA around a set of KAs. Attainment can be assessed by actors within the systems – through peer evaluations and self-evaluations, or from outside the system – through expert evaluations. Peer-evaluation attainment is interrelated with evaluator biases and controversy of the KA, and miscalibration depends on self-evaluation by the creator and evaluations by other evaluators.

Miscalibration reflects the (mis)alignment between creator's self-perception and peer evaluators' perceptions of the KA attainment. Self-evaluation is a complex social activity that requires self-reflection and critical thinking (Lin, Liu, & Yuan, 2001; Sargeant et al., 2008). In the context of developing creation and evaluation competencies, self-assessment and self-reflection play dual roles: they stimulate creator's motivation and creativity to produce and refine new KAs; they also guide the creator to be responsive to external feedback and evaluations. Self-evaluation and self-regulation are activities 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).

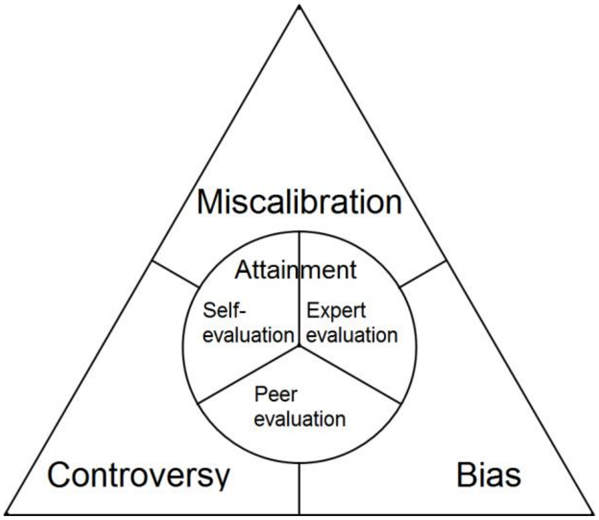


Figure 3. SKA Evaluation Intersubjectivity Model in One Single-Loop Iteration

Controversy and bias are interrelated and interdependent concepts of the departure of an evaluation of a KA from other evaluations (Lauw et al., 2008). Evaluator bias refers to a degree of deviation in the evaluators' assessments from other evaluators on less controversial KAs (more controversial KAs do not provide strong evidence of evaluator's bias because higher controversy implies significant biases of all evaluators). KA controversy refers a degree of divergence of evaluations of a specific KA among evaluators. For the purpose of this study, these notions are employed as manifestations of intersubjective congruence (or the lack of it) among peer evaluators' judgments about attainment of KAs. Thus, attainment is systematically influenced by two moving parts: (1) controversy as the lack of agreement among peers about a specific KA's attainment; and (2) bias as an individual evaluator's inability to reach consensus with other peer evaluators about the attainment of given KAs. While controversy reflects a level of agreement, or overall evaluation consensus, among evaluators about the attainment of a KA, it also intrinsically reflects the creator's ability to address their evaluating audience. In other words, KAs characterized by higher controversy are probably comprehended only by a part of their evaluating audience. In this sense, non-controversial high-attainment KAs reflect high creation competency. In addition to attainment of critiques, bias is also an important reflection of evaluation competency. An evaluator who has systematically low evaluation bias demonstrates high level of common understanding with the rest of evaluating audience. The level of bias indicates the extent to which a particular evaluator is reliable vis-à-vis other evaluators. A given evaluator's deviations on controversial KAs may be due to the controversy of these KAs. In contrast, if the KAs are non-controversial, any deviation would suggest idiosyncratic evaluator's bias. The deviations by less biased evaluators should be attributed to the KAs controversy because the deviations by the biased ones are likely to occur due to bias (Lauw et al., 2008).

Together, these four concepts – attainment, controversy, bias, and miscalibration – describe the system-level understanding, or knowledge, about a set of KAs on a specific complex open-ended problem among the actors solving the problem. Note that since in each interaction evaluators evaluate the original KAs and creators evaluate evaluators' critiques, the outcome of each creator-evaluator interaction is represented by two such triangles – one for the original KAs and another for the critiques (Figure 4).

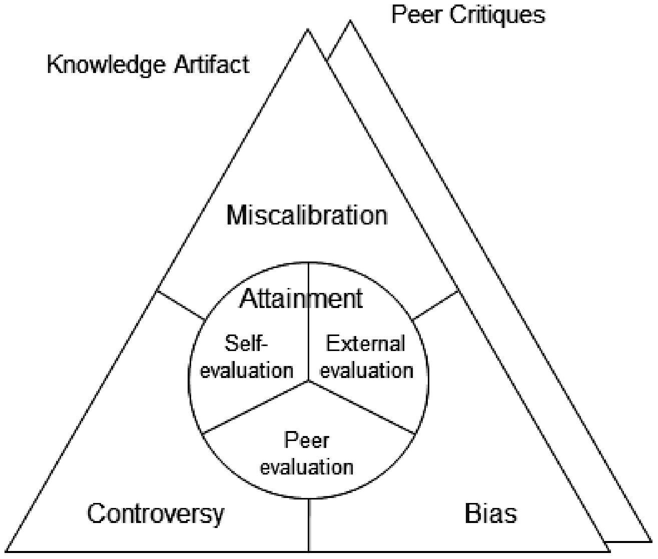


Figure 4. SKA Evaluation Intersubjectivity Model in One Double-Loop Iteration

Competency Dynamics

The SKA forms a basic unit in which unobserved creation and evaluation competencies are exposed and can be measured. Through the information exchange of original KAs, their critiques, and evaluations, creation and evaluation competencies of actors are revealed. As competencies interact, they reciprocally affect each other. Incremental changes in competencies affect the characteristics of the new KAs, critiques and evaluations produced by the actors in the next round of creator-evaluator interaction as the process repeats. Under the assumption of no other extraneous influences, the change of competencies through the practice of evaluating and being evaluated can be attributed to social learning. The system effect is the change of competency of the entire SKCC over time (Figure 5).

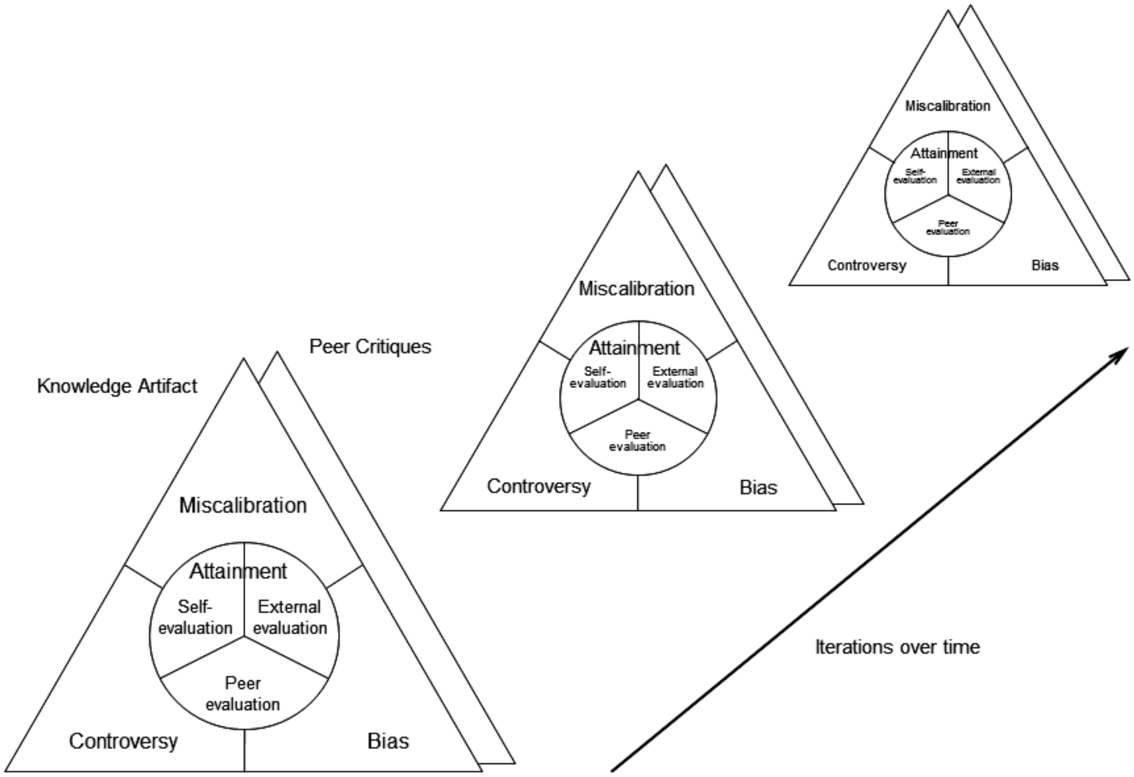


Figure 5. Conceptual Model of Competency Change over Time

The question remains open of "How exactly does this change happen?" Several perspectives may help answer this question. For competencies to be developed, retained and applied to new problems, actors have to engage in information-processing activities, such as rehearsal, organization, and elaboration (Gagne, 1985). These processes help cognitive structuring and re-construction (Reigeluth, 1983; Wittrock, 1978). Social Cognitive Theory suggests that social interactions provide opportunities for observing and imitating successful behaviors from models leading to changes in actors' levels of competencies and increasing chances of succeeding in solving a problem (Bandura, 1986). "…In order to acquire new tastes in music, a good prescription is to hear more music; in painting, to look at paintings; in wine, to drink good wines" (Simon, 1969, p. 186). "The idea of learning from examples can be extended to a method of learning 'by doing'" (Simon, 1969, p. 123). The interaction among peers focused on solving complex problems also facilitates learning of critical concepts (King, 1989). When peers interact in learning environment, inconsistent knowledge is exposed, opposing perceptions and ideas are explored, and inadequate logical reasoning and strategies may be challenged, leading to the better comprehension by actors (Piaget & Gabain, 1926; Slavin, 1992; Yu, Liu, & Chan, 2005). According to the social construction of knowledge perspective, intersubjective "meaning derives from interactive interpretation by multiple persons, not simply from the cognition of a single individual" (Miranda & Saunders, 2003, p. 88). Intersubjective understanding enables the SKCC to construct a richer interpretation of the complex problem-related information through taking different perspectives and generating a more comprehensive solution. "The best learning takes place when learners articulate their unformed and still developing understanding, and continue to articulate it throughout the process of learning. Articulating and learning go hand in hand, in a mutually reinforcing feedback loop" (Sawyer, 2008, p. 6).

According to the knowledge management perspective, specifically to the Dynamic Theory of Organizational Knowledge Creation, competencies that exist as tacit knowledge within individuals change through the cycle of socialization, externalization, combination, and internalization (Nonaka, 1994; Nonaka & Toyama, 2003; Nonaka & Von Krogh, 2009). In this view, knowledge creation and, therefore, competency building is a dialectical process, in which various contradictions are synthesized through dynamic interactions among individuals, the social system, and the environment. “What individuals learn always and inevitably reflects the social context in which they learn it and in which they put it into practice” (Brown & Duguid, 2001, p. 201). Together, these perspectives suggest that if the process of re-creation of SKA repeats over time, actors settle on certain views and perspectives in the form of agreement or shared understanding of disagreement (Matusov, 1996). From this perspective, actors’ competencies evolve in a cycle from individual knowledge to shared knowledge to value settlement to the new level of individual knowledge. Value settlement should lead to convergence of evaluations where attainments of KAs would improve, and bias, controversy, and miscalibration would attenuate.

The opposing perspective is that inferior competencies of some actors may increase dissonance and confusion among peers (Sluijsmans & Moerkerke, 1999). This specifically relates to the processes of self-evaluation and self-regulation that are intrinsic to creative pursuits and are complex social activities (Sargeant et al., 2008). In the context of SKA, self-evaluation and self-reflection play dual roles: they stimulate creator’s motivation and creativity to produce and refine new KAs; they also encourage the creator to be responsive to external critiques and evaluations. The behavioral economics literature, specifically, the research on the “unskilled-and-unaware” problem, indicated that actors with lower competency (the “unskilled”) tend to overestimate their performance, thus, showing overconfidence (Kruger & Dunning, 1999; Ryvkin, Krajč, & Ortmann, 2012). In contrast, individuals with higher competency levels (the "skilled") typically underestimate their performance, showing underconfidence. According to Kruger and Dunning (1999), the "unskilled" lack the metacognitive ability to realize their incompetence. They are afflicted by a "double curse" of the low skill and the low ability to recognize competence when presented with a KA. Further, the distributions of these biases in the population may not be normal if a subpopulation of the "unskilled" actors who overestimate their own performance is prevailing. Moreover, "unskilled" peers may introduce bias into evaluations within a SKCC when assessing other KAs' attainment which may compromise the reliability of the system. Based on this perspective, it can be theorized that evaluations in a SKCC may diverge or, at least, may not produce consistent patterns. Given that there may be two possible subpopulations of actors based on their competencies, the distribution of attainments may be bimodal, and its dynamics over time may not follow the same pattern. In addition, biases, controversies, and miscalibrations may be similar within subpopulations and different across them.

The two possible conjectures outlined above bracket the range of possible conflicting outcomes. Hence, theoretical predictions of the peer-based competency development outcomes in the DLMA SKCC are ambiguous. Therefore, the domain and phenomena of interest warrant further studies. In particular, it is interesting what factors contribute to convergence or divergence of shared understanding and, consequently, evaluation in a SKCC.

Previous research showed that miscalibration may be reduced by feedback (Brutus, Donia, & Ronen, 2013; Ryvkin et al., 2012). The reduction of miscalibration between expert and self-evaluations over multiple assignments reflects the learning effect. That is, if over time, creator's self-evaluation converges towards expert evaluation, one may conclude that feedback had positive effect on the actor's mastering of creation competencies in a particular domain. In addition, a reduction of miscalibration between peer and self-evaluations over multiple assignments reflects intersubjective shared understanding among actors. In other words, irrespective of expert's evaluations, creator's self-evaluation converging towards peer evaluation over time suggests that an actor shares understanding of creation competencies in a particular domain with his peers. The difference between peer and expert evaluations, in turn, indicates whether they are based on common understanding of creative competencies. The reduction of this difference over time suggests overall homogeneity of learning among actors.

One factor that may affect convergence in shared understating in SKCC and the evolution of the SKA over the multiple temporal iterations is whether the membership of SKCC is constant or dynamically changes. In other words, the question is how the openness of SKCC affects the dynamics of competency building and artifact development. In closed SKCCs the actor membership is constant, i.e., over time actors interact with each other within the same steady group. In contrast, in open SKCCs, the membership constantly changes, i.e., the creator-evaluator interaction may not be between the same actors. To model this environment, it is assumed that SKCC may consist of two types – *fixed-membership (steady) groups* or *randomly recombined peer* *groups*. In reality, the nature of SKCCs may be more intricate; however, for the purpose of this dissertation and to test efficacy of the proposed analytical method for examining the evaluation intersubjectivity, these two SKCC configurations are considered to be extreme special cases.

These configurations may have contradictory effects on intersubjectivity in SKCCs, formulated as the following propositions. If actors interact in the double-looped peer reviews and evaluations in *fixed-membership (steady) peer groups*,

Proposition A: Actors develop and share stable expectations and/norms for KAs and, therefore, aim to comply with them in order to achieve higher evaluation (higher attainment). Thus, the perceived competencies, intersubjectivity and reliability improve; i.e., *the social norming effect* aids learning and strengthens agreement, and the group achieves norm equilibrium;

Proposition B: As actors interact over multiple iterations, actors' compliance with expectations increases, the intersubjective perceptions of competencies within the group increase (or at least perceptions converge); consequently, actors are more perplexed in distinguishing relative attainment of KAs with similar goodness; differentiating and ranking KAs becomes more challenging; thus, while intersubjectivity improves, the level of ranking reliability drops. Thus, *the ranking confusion effect* aids learning but impedes intersubjective agreement; hence, in evaluation equilibrium, most KAs' controversy is high.

If actors interact in the double-looped peer reviews and evaluations in *randomly recombined peer groups,*

Proposition C: Actors learn from being exposed to many more good and bad KA examples than in steady groups; therefore, when producing their own KAs, they take into account strength and weaknesses they observe in others' work (implicit feedback); in addition, they learn from critiques received from many more peer reviewers (explicit feedback) and understand better their level of competency compared to the population. Thus, *the cross-pollination effect* aids learning and strengthens agreement;

Proposition D: As actors proceed over multiple iterations, they interact with ever-changing peer reviewers with varying degrees of competency and expectations; consequently, inconsistent or even contradicting explicit and implicit feedback confuses and disorients subjects about what constitutes a "good" KA. For this reason, actors do not necessarily improve attainment of their KAs and competency; intersubjective shared understanding is negatively affected. Thus, *the expectations perplexity effect* impedes learning and weakens agreement.

These four effects are mitigated by the metacognitive abilities of actors to recognize their own competence and that of others. Specifically, the Kruger-Dunning (1999) *unskilled-and-unaware* effect may amplify the bifurcation between actors with higher and lower competency:

Proposition E: The low-competency subjects ("the unskilled") overestimate their absolute and relative attainment (*the overconfidence effect*);

Proposition F: The high-competency subjects ("the skilled") underestimate their absolute and relative attainment (*the underconfidence effect)*.

These two effects (E and F) impact intersubjectivity and reliability of peer evaluation by adversely affecting the reduction of miscalibration. On the other hand, the unskilled-and-unaware problem can be reduced with feedback (Ryvkin, Krajč, Ortmann, 2012). Therefore, the four effects described in A, B, C and D are reciprocally affected by the initial competency distribution in the actor population through effects E and F.

Hypotheses

Based on the discussion above the following hypotheses are formulated (a summary of hypotheses is given in Table 1). Each hypothesis is stated for original KA submissions (marked *Ar*) and critiques to the original KA submissions (marked *Cr*). With regard to the hypothesized relative rate of change, two-tailed hypotheses are formulated (marked *A* and *B*) to reflect the ambiguity about dominating effect.

Table 1. Hypotheses Summary

Hypothesis		Knowledge artifact	
		Submission	Critiques
Attainment increases	over multiple iterations	H1Ar	H1Cr
	faster in recombined than in steady	H2ArA	H2CrA
	slower in recombined than in steady	H2ArB	H2CrB
Miscalibration decreases	over multiple iterations	H3Ar	H3Cr
	faster in recombined than in steady	H4ArA	H4CrA
	slower in recombined than in steady	H4ArB	H4CrB
Controversy decreases	over multiple iterations	H5Ar	H5Cr
	faster in recombined than in steady	H6ArA	H6CrA
	slower in recombined than in steady	H6ArB	H6CrB
Bias decreases	over multiple iterations	H7Ar	H7Cr
	faster in recombined than in steady	H8ArA	H8CrA
	slower in recombined than in steady	H8ArB	H8CrB

H1Ar: The *submission attainment increases* over multiple assignments (i.e., creation competency improves) (thanks to the social norming and the cross-pollination effects).

H1Cr: The *critique attainment increases* over multiple assignments (i.e., evaluation competency improves) (thanks to social norming and cross-pollination effects).

H2ArA: The *submission attainment increases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to the cross-pollination effect in recombined groups).

H2CrA: The *critique attainment increases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to the cross-pollination effect).

H2ArB: The *submission attainment increases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (due to expectations confusion effect in recombined groups).

H2CrB: The *critique attainment increases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (due to expectations confusion effect in recombined groups).

H3Ar: The *submission miscalibration decreases* over multiple assignments (i.e., evaluation competency improves) (thanks to the social norming effect).

H3Cr: The *critique miscalibration decreases* over multiple assignments (i.e., evaluation competency improves) (thanks to the social norming effect).

H4ArA: The *submission miscalibration decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to social norming effect in fixed-membership groups).

H4CrA: The *critique miscalibration decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to social norming effect in fixed-membership groups).

H4ArB: The *submission miscalibration decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to cross-pollination effect in fixed-membership groups).

H4CrB: The *critique miscalibration decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to cross-pollination effect in fixed-membership groups).

H5Ar: The *submission controversy decreases* over multiple assignments (i.e., creation competency improves) (thanks to the social norming and the cross-pollination effects).

H5Cr: The *critique controversy decreases* over multiple assignments (i.e., evaluation competency improves) (thanks to the social norming and the cross-pollination effects).

H6ArA: The *submission controversy decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (due to expectations confusion effect).

H6CrA: The *critique controversy decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (due to expectations confusion effect).

H6ArB: The *submission controversy decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (due to ranking confusion effect).

H6CrB: The *critique controversy decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (due to ranking confusion effect).

H7Ar: The *submission evaluation bias decreases* over multiple assignments (i.e., evaluation competency improves) (thanks to the social norming effect).

H7Cr: The *critique evaluation bias decreases* over multiple assignments (i.e., evaluation competency improves) (thanks to the social norming effect).

H8ArA: The *submission evaluation bias decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to social norming effect in fixed-membership groups).

H8CrA: The *critique evaluation bias decreases slower in randomly recombined peer groups* than in fixed-membership (steady) groups (thanks to social norming effect in fixed-membership groups).

H8ArB: The *submission evaluation bias decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (due to ranking confusion effect fixed-membership groups).

H8CrB: The *critique evaluation bias decreases faster in randomly recombined peer groups* than in fixed-membership (steady) groups (due to ranking confusion effect fixed-membership groups).

CHAPTER III

METHODOLOGY

To investigate temporal dynamics of creation and evaluation competencies in peer-based SKCC, a longitudinal experimental study was conducted with university students over the course of one semester. This chapter explains details of the conducted study, including the analytical method and IT-enabled instantiation, empirical research design (participants, protocol and procedure), collected data and analysis methodology.

Research Design

Participants

Participants were 97 students at a large public university in North Carolina, USA, taking a course in Systems Analysis and Design at the School of Business and Economics (out of 99 students initially enrolled, 97 students completed the course). The course was taught in fall 2014 and spanned 16 weeks. Fifty-six students majoring in Information Systems and Supply Chain Management were enrolled in the undergraduate face-to-face section; 43 students majoring in Information Technology Management were enrolled in the graduate online section. Both sections had the same study plan, content and assignments, and were taught by the same instructor. Age, gender or other demographic data were not collected.

Experiment Design

The experiment followed the randomized complete block (RCB) design with repeated measures within participants. Student participants were subjected to double-looped mutual peer reviews and evaluations under two alternative experimental conditions (treatments):

- Condition X_1: *Randomly recombined peer groups;*
- Condition X_2: *Steady (fixed-membership) peer groups.*

Since students of two different education levels participated in the study, these levels were treated as control blocks of observational conditions:

- Undergraduate (in the face-to-face section);
- Graduate: (in the online section).

Allocation of student participants to experimental conditions and control blocks is presented in Table 2. Prior to assignment 1, participants in the respective courses were randomly divided into the experimental conditions pools of equal size; participants were not informed what condition they were assigned to. In each assignment, participants in the condition X_1 were allocated into peer groups randomly, i.e., in each assignment each participant interacted with a new random set of peer reviewers. In the condition X_2, participants were placed randomly in peer groups in the first assignment and remained with the same peer group for all consecutive assignments, i.e., they gave to and received critiques and evaluations from the same set of peers.

Table 2. Allocation of Participants to Experimental Conditions and Control Blocks

		Experimental Conditions	
		Randomly recombined groups, X_1 (48 students)	Steady groups, X_2 (49 students)
Blocks	Undergraduate (56 students)	Assignment 1 (28)	Assignment 1 (28)
		2 (28)	2 (28)
		3 (28)	3 (28)
		4 (27)	4 (28)
		5 (27)	5 (28)
	Graduate (41 students)	1 (21)	1 (21)
		2 (20)	2 (21)
		3 (19)	3 (21)
		4 (19)	4 (19)
		5 (19)	5 (19)

This design is presented schematically in Table 3 as follows: *R* denotes randomization of subjects across blocks and treatments; *X* denotes treatment, *O* denotes observation/data collection; subscripts *U* and *G* denote blocks of undergraduate and graduate students respectively.

Table 3. Randomized Complete Block (RCB) Design with Five Repeated Measures

R_U	X_1	O_1	X_1	O_5	X_1	O_9	X_1	O_{13}	X_1	O_{17}
R_U	X_2	O_2	X_2	O_6	X_2	O_{10}	X_2	O_{14}	X_2	O_{18}
R_G	X_1	O_3	X_1	O_7	X_1	O_{11}	X_1	O_{15}	X_1	O_{19}
R_G	X_2	O_4	X_2	O_8	X_2	O_{12}	X_2	O_{16}	X_2	O_{20}

The peer group size was chosen to be five. This peer group size was chosen for the following reasons: on the one hand, about five peer evaluations give better grounds for statistical reliability of aggregated 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 & Krosnick, 1985; G. A. Miller, 1956). The group size of four to six peers is also advocated in other online peer review systems (Cho et al., 2008; Joordens, Desa, & Paré, 2009) (while the specific choice of the peer group size may affect the results of evaluation, this issue is outside the scope of this study). In practice, due to the actual number of participants, drop-outs, and the size of block and treatment pools, the peer group size varied between four and six participants.

To subdue the effect of social and inter-personal biases that typically result from non-anonymity, such as "friendship" or "hostility" evaluations, all reviews and evaluations were double-blind, i.e., the identities of the reviewers were not revealed to the recipients of Critiques and vice-versa at any time (Bamberger, 2005; Howard, Barrett, & Frick, 2010; Lu & Bol, 2007). Since creators and reviewers conduct mutual evaluations, to prevent students from engaging in retaliation behavior when evaluating peers' submissions and critiques, no summative results of performance are revealed to students prior to the assignment completion, and groups are randomly re-assigned for the next assignment (Cho & Kim, 2007; Goldin, 2011).

Protocol and Procedure

The following protocol and procedure were used to model the creator-evaluator interaction of the SKCC using a university course. During one semester, participants created knowledge artifacts in the form of digital documents in a series of assignments. In the conducted experiment, KA that participants produced as solutions to a complex open-ended problem were Submissions of “conceptual blueprints of systems that support management strategies and enable business processes” (Course Syllabus), i.e., documents that explicate system analysis and design. The course included five assignments for which students were required to turn in Submissions, Reviews, and Reactions. Specifically, participants were given the following assignments:

Assignment 1: System requirements specification;

Assignment 2: Use case diagram and description;

Assignment 3: Work-flow diagrams and description;

Assignment 4: Domain class diagram and description;

Assignment 5: Refined domain class diagram, systems level class diagram, and description.

Each consecutive assignment built on previous assignments. Participants worked on their Submissions individually and independently. Assignment descriptions were the same for all participants. In addition to turning in their Submissions, participants reviewed (critiqued and evaluated) each other’s Submissions in peer groups, and evaluated each other’s Critiques. That is, in each assignment each participant completed the following steps combined in three tasks:

Task 1: Submission (corresponds to the Step 1 in the system model:

Composed and turned in a Submission (as a pdf document);

Task 2: Review (corresponds to the Step 2 in the system model):

Received several Submissions for review;

Provided Critiques and holistic evaluations to several other peers' Submissions;

Provided self-critique and self-evaluation of his own Submission;

Task 3: Reaction (corresponds to the Step 3 in the system model):

Received several peer Critiques of own Submission;

Provided holistic evaluations of received Critiques;

Provided self-evaluation of own Critiques;

Received peer evaluations of own Critiques.

To guide holistic evaluations of Submissions and Critiques, in each assignment all participants were given the same problem-specific rubrics for Submission and Critiques evaluation (same within control blocks and experimental conditions). In each assignment, of the Submission evaluation rubrics pertained to the given assignment but the verbiage consistently addressed completeness and presentation quality of diagrams and descriptions in the Submission in that assignment. For all assignments, Critique evaluation rubrics were uniform. A sample assignment is presented in Appendix C.

Participants completed these assignments as part of their course work. Students were not offered any monetary or social incentives to participate in the experiment. The scores resulting from peer evaluations of both Submissions and Critiques were included as a significant component of students' grades to assure diligent responses to all tasks. A threat of the 50% assignment score reduction penalty for not turning in the Critique evaluations (Reaction) was announced to provide an incentive to fully complete each assignment. The overall response rate was 97% for Submissions, 94% for Reviews, and

92% for Reactions; 467 usable longitudinal records were obtained across five assignments.

All Submissions, Critiques and evaluations were turned in, redistributed, collected online through Mobius SLIP online application (described in the next subsection). Instructions on how to complete each assignment, including the requirements for the Submission, Critiques, and evaluations, were provided to participants by their professor. In addition, technical support for the use of software was provided to participants by the researcher through online user guides and through the help desk. No individual instructor feedback was provided to participants; only general uniform feedback on each assignment upon its completion was provided to the entire block of participants (either during a face-to-face class or online teleconference). After the completion of the assignment, individual-, group-, and class-level performance metrics were presented to students online (Figure 7).

Tackling Challenges of Gauging Peer Evaluation Intersubjectivity: Design and Implementation of the Online Peer Review System

Capturing Performance of Creators and Evaluators

Mobius Social Learning Interaction Platform (SLIP), a web-based peer review system, is used in this study to facilitate and monitor students' creator-evaluator interactions and to collect quantitative evaluations of Submissions and Critiques. The system was designed and developed at UNCG to promote learning through complex assignments, creative problem solving, critical thinking, communication and collaboration in large face-to-face and online classes. It is an online environment that combines challenging complex problem-solving assignments with iterative, anonymous peer reviews, formative and summative peer and self-assessments of both problem solution Submissions and Critiques, as well as provides analytics for targeted instructor feedback.

Mobius SLIP is built around the Double-Loop Mutual Assessment (DLMA) analytical method (Ford & Babik, 2013), presented in appendixes A and B. In DLMA, participants' Submissions are reviewed by several peers whose Submissions address the same problem; i.e., participants who created KAs solving a specific problem assess and critique each other's KAs. Then, each participant assesses a set of peer critiques received to his own Submission. Then, they assess each other's critiques in a similar fashion. In other words, participants evaluate not only their own and their peers' creations but also their own and their peers' critiques. Self-assessment of Submissions and Critiques is also captured. These multiple anonymous double-blind individual peer and self-assessments of Submissions and Critiques then are aggregated into a set of scores – peer-evaluation attainment, self-evaluation attainment, miscalibration, controversy, and bias. The process repeats over multiple assignments.

Approaches similar to the DLMA were proposed by other authors (Gehringer, 2001; Hamer, Ma, & Kwong, 2005; Cho & Schunn, 2007; Sitthiworachart & Joy, 2004). Important advantages of such approaches are two-fold. Firstly, they provide an opportunity to obtain evaluation data for both Submissions and Critiques from participants who attempt to solve the same problem. Therefore, such data inherently contains information about creator and evaluator competency, as well as about peer groups' overall evaluation intersubjectivity, which is the focus of this research. Secondly, mutual evaluations of Critiques and data on quality of Critiques and characteristics of peer and self-evaluations fed back to participants create an important motivating stimulus for the reviewers to exert their best effort in providing constructive Critiques to their peers and in evaluating Submissions and Critiques as candidly and rigorously as they can by holding them accountable for the goodness of both Submission and Critiques, and goodness of evaluations of Submission and Critiques . This changes the overall "motivational structure" of the SKCC transforming it into a self-regulating and self-norming social system (Berkowitz, 2004; Boyd & Richerson, 2002; Perkins & Berkowitz, 1986; Tinapple, Olson, & Sadauskas, 2013).

The Review task of each assignment produces two types of information – Critiques of Submissions that are directed to the creators, and Submission evaluation data collected by the system. From the Review task, the following pieces of the Submission evaluation data are collected:

(a) Ranking of each Submission in the peer group by each of the peer reviewers (evaluators) in the group, who provided their Critiques and evaluations of Submissions;

(b) Self-ranking by each creator, who provided Critiques and evaluations of peers' and their own Submissions;

(c) Rating of each Submission in the peer group by each of peer reviewers (evaluators) in the group, who provided their Critiques and evaluations of Submissions;

(d) Self-rating by each creator, who provided Critiques and evaluations of peers' and their own Submissions.

The Reaction task serves the purpose of collecting the evaluations of Critiques received by the creators from the evaluators. From the Reaction task, the following pieces of data are collected:

(a) Ranking of each Critique set in the peer group by each of the creators in the group, who provided evaluations of Critiques;

(b) Self-ranking by each reviewer (evaluator), who provided evaluations of Critiques, of their own set of Critiques given;

(c) Rating of each Critique set in the peer group by each of creators in the group, who provided evaluations of Critiques;

(d) Self-rating by each reviewer (evaluator), who provided evaluations of Critiques, of their own set of Critiques given.

Submissions and Critiques were also evaluated by two experts (an instructor and a teaching assistant) to generate external expert evaluations of Submissions and Critiques. Inter-rater reliability between the two experts was computed and then evaluations were averaged to obtain a measure of attainment produced outside the peer evaluation system as possible proxy of the "underlying goodness" of KAs. Expert evaluations of students' KA submissions and critiques were collected through the instructor interface of Mobius SLIP using both ranking and rating scale.

These pieces of peer and self-evaluation data are then aggregated into variables of peer- and self-evaluation attainment, controversy and bias further analysis. In summary, attainment can be gauged in the following ways – as self-evaluation attainment (i.e., what is the goodness of a KA perceived by its creator?), peer-evaluation attainment (i.e., what is the goodness of a KA expressed as the aggregation of peer evaluators' perceptions?) or expert-evaluation attainment (i.e., what is the goodness of a KA perceived by a higher-status evaluator?).

This dissertation places specific focus on the development of the analytical methods of metrics that capture intersubjectivity in such mutual evaluations and the design choices being made in implementing the DLMA as an information system supporting peer-based knowledge creation, evaluation and refinement. In the following subsections, the choice of analytical metrics and measurement scales for capturing the evaluation intersubjectivity in the DLMA are explained.

Choosing Measurement Scales

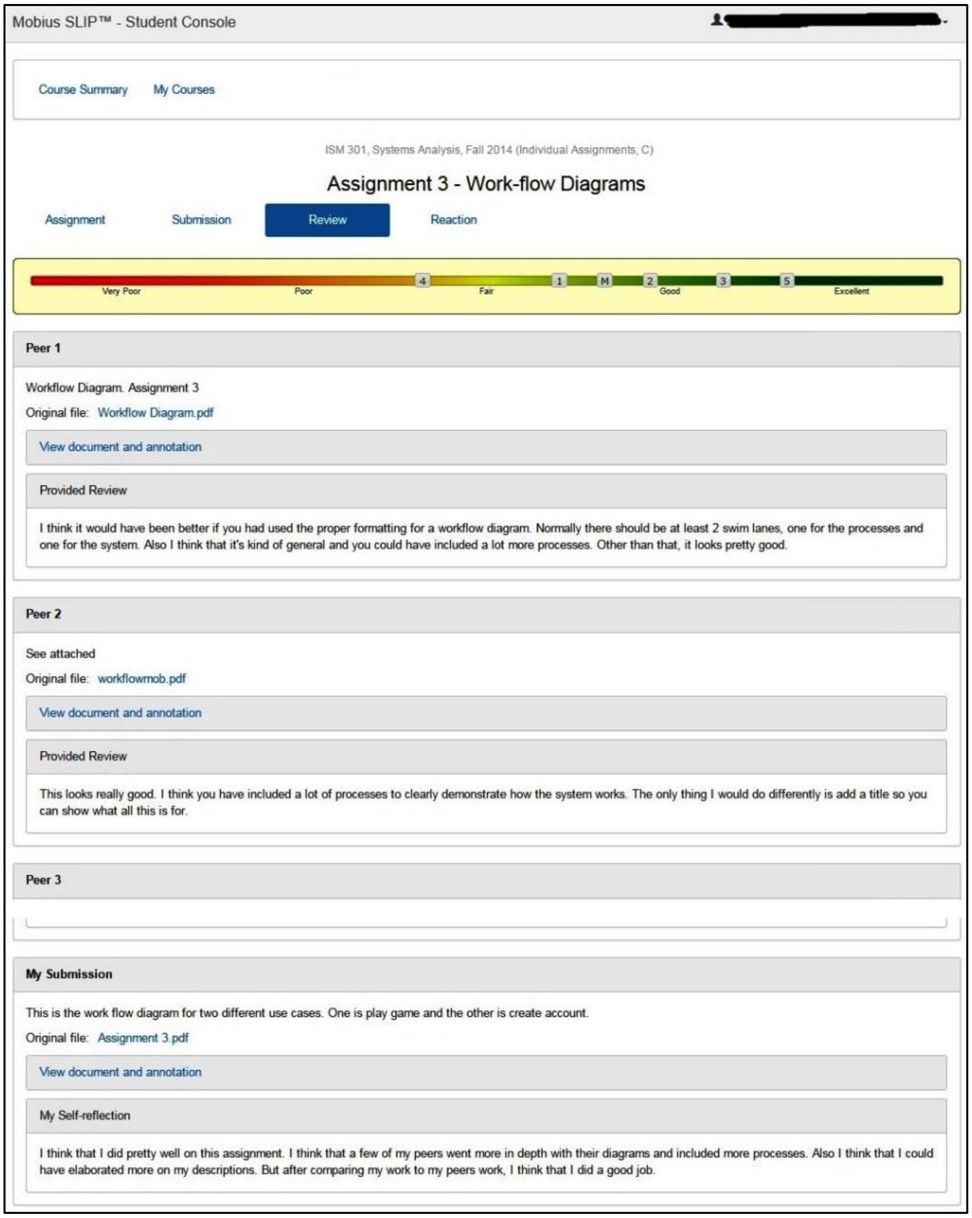
In general, summative evaluations may be conducted using either ranking or rating (Douceur, 2009). *Rating* refers to the comparison of different items using a common absolute, or *cardinal*, scale. *Ranking*, sometimes also called forced-distribution rating, means comparing different items directly one to another on a relative, or *ordinal*, scale (Schleicher, Bull, & Green, 2008). Both ranking and rating have their strengths and weaknesses, and there is still little consensus as to which has a greater predictive validity (Alwin & Krosnick, 1985; Krosnick, 1999; Krosnick, Thomas, & Shaeffer, 2003). Generally, they are expected to correlate, but some studies have demonstrated that ordinal (i.e., ranking-based) evaluations contain significantly less noise than cardinal (rating-based) evaluations (Shah et al., 2013; Waters, Tinapple, & Baraniuk, 2015).

Cardinal scale in the context of peer evaluations is also susceptible to score inflation, whereas ordinal scale is immune to this problem (Douceur, 2009). When cardinal scale is used, an evaluator may “smokescreen” his preferences by giving all evaluated artifacts the same rating, and may severely inflate scores by giving all artifact the same high ratings (similarly, he can severely degrade scores by giving all artifacts the same low ratings). Thus, cardinal scale is very vulnerable to personal and social biases or idiosyncratic shocks, such as mood or personal variation in evaluation style (e.g., never give the highest rating). When ordinal scale is uses, an evaluator must make an explicit and transitive choice of preferring each artifact (based on its perceived goodness) over others (Slovic, 1995). This makes the evaluation more robust. Psychological evidence suggests that evaluators are better at making comparative judgment than absolute one (Spetzler & Stael Von Holstein, 1975; Wang, Dash, & Druzdzel, 2002).

The ordinal scale also has its drawbacks. It forces evaluators to discriminate between artifacts that may be perceived to have very similar *goodness* as much as between the artifacts which qualities may be far apart. Some ordinal scales may implicitly emphasize items earlier in the list and lead to their higher ranking. Evaluating on ordinal scales places more cognitive load on the evaluator because it requires him to compare multiple items against each other.

To avoid making an explicit design choice between ranking and rating and reduce cognitive evaluation biases due to the choice of the interface controls, in Mobius SLIP, summative evaluations are captured using the SLIP Slider GUI control (Figure 6, Figure 7). The SLIP Slider control displays a color-coded and labeled bar that represent the continuum from *Very poor* to *Excellent,* on which numbered handles corresponding to each particular Submission (or Critique) can be positioned according to the evaluator's judgment about their attainments based on the provided rubric. The M-handle represents self-evaluation ("me/my"). Importantly, handles could not be overlapped to indicate equivalent attainment levels. That is, judgments about attainment of any two KAs had to be at least marginally distinct. Therefore, the SLIP Slider forces participants to make judgments about merits of KAs relative to each other.

Thus, in this study, data on summative evaluations of KAs are captured as both ranking and rating. For measurement purposes, rating is recorded as an integer between 1 and 100 reflecting a position of the handle on the continuum from *Very poor* to *Excellent* irrespectively of the positions of other handless. Ranking is recorded as an integer between 1 (the highest rank) to group size minus 1 (*N-1*, the lowest) reflecting a relative position of the handle among other handles in the group. Note that participant's self-evaluation is not included in the computation of attainment by peer-evaluation, but instead is recorded as a separate self-evaluation attainment measure.



ISM 301, Systems Analysis, Fall 2014 (Individual Assignments, C)

Assignment 3 - Work-flow Diagrams

Peer 1

Workflow Diagram. Assignment 3

Original file: Workflow Diagram.pdf

Provided Review

I think it would have been better if you had used the proper formatting for a workflow diagram. Normally there should be at least 2 swim lanes, one for the processes and one for the system. Also I think that it's kind of general and you could have included a lot more processes. Other than that, it looks pretty good.

Peer 2

See attached

Original file: workflowmob.pdf

Provided Review

This looks really good. I think you have included a lot of processes to clearly demonstrate how the system works. The only thing I would do differently is add a title so you can show what all this is for.

Peer 3

My Submission

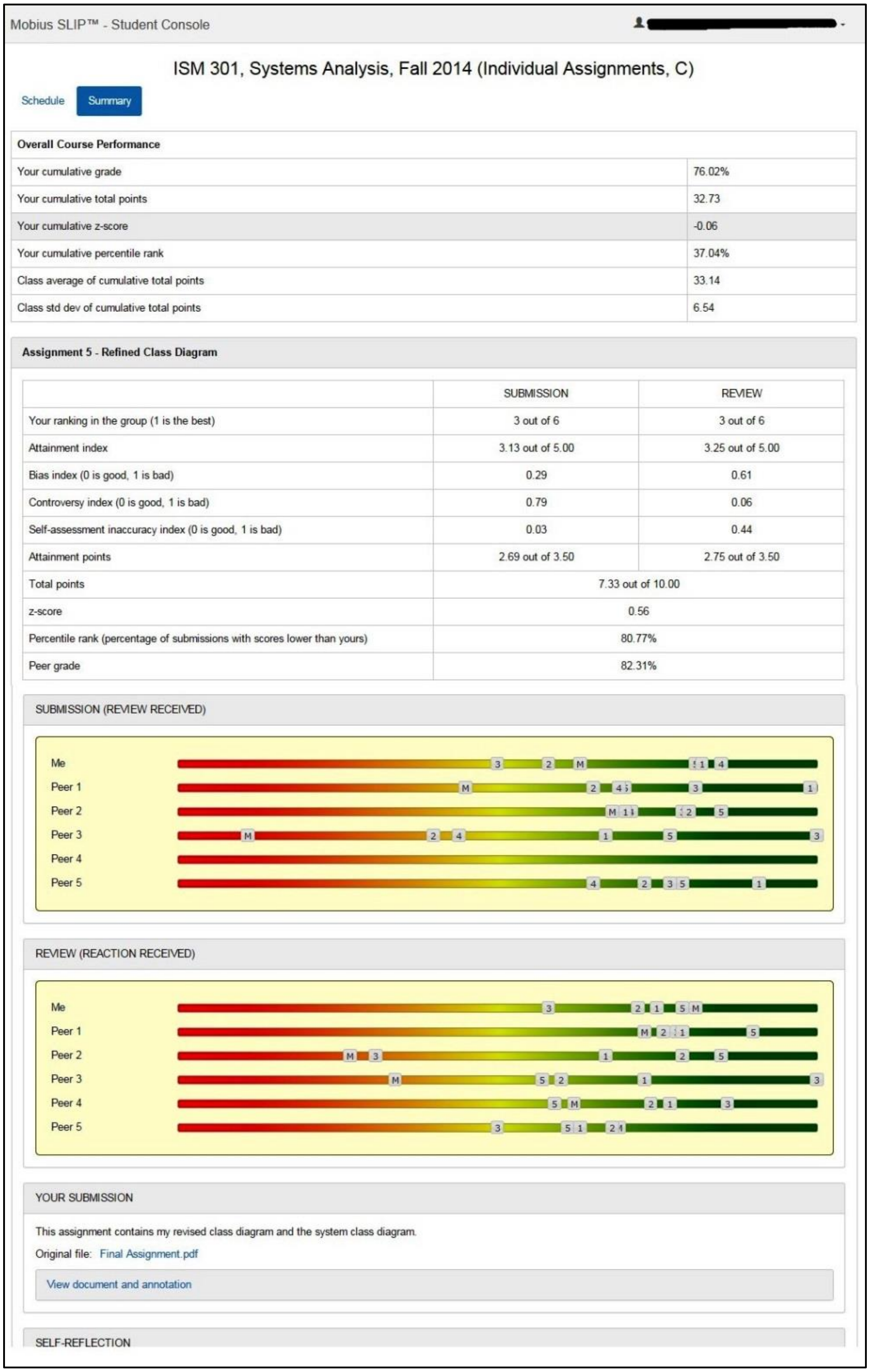
This is the work flow diagram for two different use cases. One is play game and the other is create account.

Original file: Assignment 3.pdf

My Self-reflection

I think that I did pretty well on this assignment. I think that a few of my peers went more in depth with their diagrams and included more processes. Also I think that I could have elaborated more on my descriptions. But after comparing my work to my peers work, I think that I did a good job.

Figure 6. The Mobius SLIP Assignment Review Interface



Mobius SLIP™ - Student Console [redacted]

ISM 301, Systems Analysis, Fall 2014 (Individual Assignments, C)

Overall Course Performance	
Your cumulative grade	76.02%
Your cumulative total points	32.73
Your cumulative z-score	-0.06
Your cumulative percentile rank	37.04%
Class average of cumulative total points	33.14
Class std dev of cumulative total points	6.54

Assignment 5 - Refined Class Diagram

	SUBMISSION	REVIEW
Your ranking in the group (1 is the best)	3 out of 6	3 out of 6
Attainment index	3.13 out of 5.00	3.25 out of 5.00
Bias index (0 is good, 1 is bad)	0.29	0.61
Controversy index (0 is good, 1 is bad)	0.79	0.06
Self-assessment inaccuracy index (0 is good, 1 is bad)	0.03	0.44
Attainment points	2.69 out of 3.50	2.75 out of 3.50
Total points	7.33 out of 10.00	
z-score	0.56	
Percentile rank (percentage of submissions with scores lower than yours)	80.77%	
Peer grade	82.31%	

SUBMISSION (REVIEW RECEIVED)

REVIEW (REACTION RECEIVED)

YOUR SUBMISSION

This assignment contains my revised class diagram and the system class diagram.

Original file: Final Assignment.pdf

View document and annotation

SELF-REFLECTION

Figure 7. The Student Interface Showing Self- and Peer Evaluation Results

Operationalizing Intersubjectivity Using Ordinal Scale

To operationalize the concepts of intersubjectivity comprising our research model, the measures of attainment, controversy, bias, and miscalibration were developed. The detailed algebraic representation of the DLMA analytical method is presented in appendixes A and B. This subsection explains the intuition behind each of the variables and illustrates their calculations based on the ordinal scale.

Mutual peer and self-evaluation data, collected in the Review and the Reaction tasks, can be represented as square matrices, where the row index identifies the recipient of evaluation, and the column index identifies the provider of evaluation. Attainment scores are computed by inverting ranks; 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. This inversion and transformation of rank into scores is necessary to assure that attainment scores do not depend on the peer group size because in two groups with slightly different number of actors numeric values of highest or lowest ranks may vary. Aggregate peer-evaluation 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; any peer-evaluation ranks below self-rank are shifted one notch up.

A simple example illustrates this (Figure 8). Consider the following matrix of the mutual peer-evaluation attainment scores in a group of five actors acting as both creators of KAs and evaluators. 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-evaluation from attainment calculations. The higher numbers signify the higher ranking. As can be seen from Figure 8, actor I receives the aggregate attainment score of 1, and the actor IV receives the aggregate attainment score of 4.

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 8. Example Scenario of Mutual Peer Ranking

Self-evaluation attainment score is computed by inverting self-rank similarly to converting individual peer-evaluation ranks. *Expert evaluation attainment* scores in each peer group in each assignment are also computed by inverting ranks given by each expert to each KA in the peer group in the assignment and averaging attainment scores produced by the experts for each KA.

Operationalizing the concept of miscalibration is straight forward: it is a deviation of self-evaluation of a KA from an external evaluation (such as peer or expert). *Miscalibration with respect to (WRT) peer evaluation* is computed as a difference between the self-evaluation attainment score and the aggregate peer-evaluation attainment score. Similarly, *miscalibration with respect to expert evaluation* is computed as a difference between the self-evaluation attainment score and the expert-evaluation attainment score. Miscalibration captures two aspects – how far is self-evaluation from the external evaluation (magnitude, or size), and the direction (or sign) of miscalibration (overconfidence or underconfidence). For illustration, consider the same scenario as in Figure 8, but now with self-assessment scores given on the main diagonal (Figure 9). Obviously, actor V shows very low miscalibration (his aggregate peer-evaluated attainment score is 4, and his self-assessment attainment score is also 4, hence, miscalibration is zero); whereas, actor I shows very high overconfidence (his aggregate peer-evaluated attainment score is 1, and her self-assessment attainment score is also 4, thus, miscalibration is negative 3).

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 9. Mutual Peer Evaluation and Self-evaluation Ranking

Miscalibration with respect to expert evaluation captures the dissonance between perceptions of a student and those of the expert, and can be interpreted as the actor's inability to evaluate adequately the "true" attainment of his KA. Therefore, a larger miscalibration with respect to the expert evaluation is treated as poor evaluation competency. Miscalibration with respect to peer evaluation captures the dissonance between perceptions of an actor and those of his peers and can be interpreted as the intersubjective disagreement about the attainment of the KA between the actor and his peer. *Miscalibration between peer evaluation and expert evaluation* is measured as the difference between peer-evaluated attainment and expert-evaluated attainment. This measure captures the dissonance between perceptions of a peer evaluator group and the expert, and is indicative of dominating poor evaluation competencies in the peer group. Operationalization of controversy and bias is not as trivial as that of miscalibration. There are many more considerations to take into account that affect design choices. The first consideration is whether controversy and bias should be computed as *deviation from mean* (DFM) or as *deviation from co-evaluators* (DFC). Lauw, Lim, and Wang (2006) argued that the deviation-from-mean approach is disadvantageous because it is more likely to produce deviation values close to zero. According to them, with a larger number of evaluators of the same KA, the distribution of evaluation scores is likely to peak at or near mean; deviation from mean would therefore approach zero. Consequently, the ratio among deviation values determines the computation outcome, and small changes in absolute deviation value may lead to a large change in ratio, making the system too sensitive to small changes. Deviation from co-evaluators produces larger deviation values and, therefore, is not likely to be very sensitive to small changes. In this dissertation, the hypotheses are tested using both operationalizations of controversy and bias.

The second consideration is whether controversy and bias computed based on the ordinal and cardinal scales capture the same phenomena. The ordinal scale assumes that the distances between neighboring values are the same across the scale. Therefore, a larger controversy means larger spread of rankings, and a larger bias means a more substantial misalignment of rankings given by a particular evaluator with the rest of co-evaluators. When using the cardinal scale, deviation values will contain information not only on misalignment of ranking but also on how an evaluator used the scale, i.e., the order, the average rating given, and the spread of ratings given. Therefore, considering the scope of this problem, this dissertation focuses on controversy and bias computed using the ordinal scale.

The third consideration is the mutual dependency of controversy and bias (Dai, Zhu, Lim, & Pang, 2012; Lauw et al., 2006, 2008). Both bias and controversy aggregate deviations among assessments of multiple artifacts by multiple evaluators. The input data for computing bias and controversy greatly overlaps. The difference is that controversy focuses on an artifact and captures the spread of its evaluations, whereas bias focuses on an evaluator and captures dissimilarity between his evaluations and those of other evaluators. A simple and straightforward approach to estimate controversy and bias is to aggregate deviations – either deviations from mean or deviations from co-evaluators. In this approach, evaluators whose assessments deviate substantially from other co-evaluators are considered to be biased, and artifacts that generate much deviation are considered to be controversial. Lauw, Lim, and Wang (2008) called this approach the *naïve* approach. They argued that the weakness of the naïve approach is that it ignores controversy when determining bias and vice versa; it treats an evaluator's deviations in assessing different artifacts equally without considering possibly different controversies of these artifacts (i.e., it ignores the fact that a higher bias may be reasonable if the artifact is controversial). To overcome the weaknesses of the naïve approach, Lauw, Lim, and Wang (2008) proposed the *Inverse Reinforcement* (IR) model and the *Evidence* model that account for mutual dependency and determine the degree of support for computed bias and controversy. While incorporating these models into the study of the evaluation intersubjectivity in SKCC is an interesting and promising avenue for future research, this dissertation applies the naïve approach to establish the base line for further investigation and design research of peer-based KA evaluation systems.

The forth consideration is the mutual dependency of attainment and controversy computed based on ordinal scale. When the forced distribution evaluation (ranking) is used, while the low-controversy artifacts fall into one of the three attainment categories (high, medium, and low), the high-controversy artifacts tend to fall into the medium attainment category because averaged inversed ranks in this case gravitate towards the median inverse rank. One of the two ways can be used to overcome this problem. One way is to rely in the external measure, such as expert evaluation, as a source of information on the underlying "true" artifact goodness. This way is straightforward and may be applied under certain circumstances, for example, in educational settings where the expert's competencies is demonstrably and reliably higher than peers; competencies. This may not work, however, in SKCC where there are no higher-competency experts and peer evaluations are the sole source of goodness assessment. This may be further complicated by a high degree of complexity and diversity of the problems that a SKCC attempts to solve by creating and evaluating KAs – in such environment, any actor's creation and evaluation competency is limited and cannon be relied on as a source of "underlying truth". The second way is, therefore, to refine methodologies of determining bias, controversy and to derive validity of evaluations through inter-observer reliability (Uebersax, 1988).

In this dissertation, the naïve approach is applied and controversy and bias are computed as both deviations from mean and deviations from co-evaluators. The following necessary adjustments are made in the computations: (a) for the number of submitted artifacts and participating evaluators; (b) for the peer group size to make these scores comparable across different groups; and (c) for the bias and controversy nonlinearity due to the use of ranking and the exclusion of self-assessment from the attainment computation).

Controversy of a KA as DFM is computed as the aggregate absolute value of deviations between attainment scores given to the KA by each evaluator and the average attainment score given by the rest of evaluators (excluding creator's self-evaluation). *Controversy of a KA as DFC* is computed as the aggregate absolute value of pair-wise differences between co-evaluators' attainment scores given to the KA by each participating evaluator (excluding creator's self-evaluation).

Figure 8 illustrates a scenario where each KA has zero controversy, i.e., all KAs were assigned the same ordinal positions (ranks) by all evaluators. Consider now the following scenario on Figure 10. Artifacts III, IV, and V show little variation in received ranks; the aggregate DFM and DFC of four respective attainment scores of each of these KAs is not large, and, hence, these KAs 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 aggregate DFM and DFC of four respective attainment scores of this KA 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 KA is also more controversial than artifacts III, IV and V.

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 10. Controversy in Mutual Peer Evaluation

Bias of an evaluator as DFM is the aggregate absolute value of deviations between the attainment scores given to every KA by the evaluator and the average attainment score given to these KAs by the rest of evaluators (excluding creator's self-evaluation). *Bias of an evaluator as DFC* is the aggregate absolute value of pair-wise differences between attainment scores given by the evaluator to all evaluated KAs and attainment scores given by all other co-evaluators to all respective KAs. Figure 8 illustrates a scenario where each evaluator shows zero bias (with respect to other evaluators), i.e., all evaluators assigned all KAs the same ordinal positions (ranks). Consider now the following scenario on Figure 11. Actors I, II, III, and IV assigned all KAs the same ranks (adjusted for exclusion of their self-evaluations). Thus, they are in implicit agreement about attainment of all KAs. Actor V, however, assigned ranks to all KAs in the reverse order; thus, the aggregate deviations of evaluations of actor V from the four evaluations by actors of each respective KAs is large, and, hence, actor V can be considered a highly biased evaluator (irrespective of the sources of his psychological or social biases).

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 11. Bias in Mutual Peer Evaluation

Operationalizing Intersubjectivity Using Cardinal Scale

Operationalization of attainment in the case of the cardinal scale is simpler than that based on the ordinal scale. Rating-based attainment score is directly equal to the position of the handle on the SLIP Slider bar, i.e., is to an integer between 1 and 100. The average rating-based peer-evaluation attainment score is the average ratings given to a particular KA by evaluators who reviewed and assessed it. Self-evaluation attainment score is equal to the position of the M-handle on the SLIP slider. Expert-evaluation attainment score is equal to the average of the expert-evaluation ratings given to the KA by two experts.

Miscalibration is operationalized the difference between self-evaluation rating and external evaluation rating (e.g., peer or expert). Miscalibration with respect to peer evaluation is computed as a difference between self-evaluation rating and peer-evaluation average rating received. Miscalibration with respect to peer evaluation is computed as a difference between self-evaluation rating and expert-evaluation average rating received.

The Ford and Babik (2013)'s DLMA analytical method was originally developed for the ordinal scale. Taking into account considerations explained in subsection and the scope of the analysis of rating-based data, this study focuses foremost on evaluating the ranking-based design and only addresses the analysis of miscalibration based on the cardinal scale. Specifically, controversy and bias measures based on the cardinal scale are left outside the scope of this dissertation, are currently research in progress, and will be explored in the future studies.

Data Analysis

Dependent Variables

To measure constructs that represent the conceptual elements of our theoretical model, data collected from participants' creator-evaluator interactions in Mobius SLIP are aggregated in the following variables (Table 4, Figure 12). Each variable is computed separately for each Submission and Critiques set produced by each student in each assignment. Descriptive statistics of dependent variables are presented in Tables 5 – 10, correlations are presented in Tables 11 and 12.

To establish evaluation benchmarks of KA that approximate their "underlying true goodness", external to self- and mutual peer evaluation interactions, expert evaluations of the original KA submissions and critiques were obtained (Table 8, Table 9). Since expert evaluations of complex open-ended problem solutions may also be very subjective and may not always be more accurate in evaluating complex KAs than a group of peer reviewers (Vista, Care, & Griffin, 2015), two instructors independently evaluated both Submissions and Critiques in each assignment. One instructor evaluator was a teaching assistant in the courses; the second instructor evaluator was the researcher of this dissertation. The entire set of Submissions and Critiques for the graduate students block was evaluated by both instructor evaluators. In the undergraduate student block, each of the two instructor evaluators assessed one entire experimental-condition subsample, and a half of the other experimental condition subsample; thus, the overlap between assessed subsamples for both instructor evaluators was about 50% of the undergraduate students block. The overlapping subsample was randomly chosen, and the instructor evaluators swopped the experimental conditions subsamples between even and odd assignments. The instructor evaluators did not discuss their approaches to evaluations prior to evaluating, nor the post-factum results of evaluations. The instructor evaluators were reasonably unfamiliar with the participants' names or prior backgrounds, and the interface used during evaluations sufficiently obscured the authorship of each Submission and Critique.

Table 4. Measures and Dependent Variables

Scale		Ordinal		Cardinal	
KA		Submission	Critiques	Submission	Critiques
Measures	Self-evaluation	ScrkStuAr_Self	ScrkStuCr_Self	RatgStuAr_Self	RatgStuCr_Self
	Peer evaluation	RankStuAr_	RankStuAr_	RatgStuAr_AvR	RatgStuCr_AvR
	Expert evaluation	RankInsAr_Exp	RankInsCr_Exp	RatgInsAr_Exp	RatgInsCr_Exp
Variables	Self-evaluation attainment	ScrkStuAr_Self	ScrkStuCr_Self	RatgStuAr_Self	RatgStuCr_Self
	Peer-evaluation attainment	ScrkStuAr_Attm	ScrkStuCr_Attm	RatgStuAr_Attm	RatgStuCr_Attm
	Expert-evaluation attainment	ScrkInsAr_Exp	ScrkInsCr_Exp	RatgInsAr_Exp	RatgInsCr_Exp
	Miscalibration WRT peer eval.	ScrkStuAr_Misc	ScrkStuCr_Misc	RatgStuAr_Misc	RatgStuCr_Misc
	Miscalibration WRT expert eval.	ScrkInsAr_Misc	ScrkInsCr_Misc	RatgInsAr_Misc	RatgInsCr_Misc
	Controversy DFM	ScrkStuAr_ContDFM	ScrkStuCr_ContDFM		
	Controversy DFC	ScrkStuAr_ContDFC	ScrkStuCr_ContDFC		
	Bias DFM	ScrkStuAr_BiasDFM	ScrkStuCr_BiasDFM		
	Bias DFC	ScrkStuAr_BiasDFC	ScrkStuCr_BiasDFC		

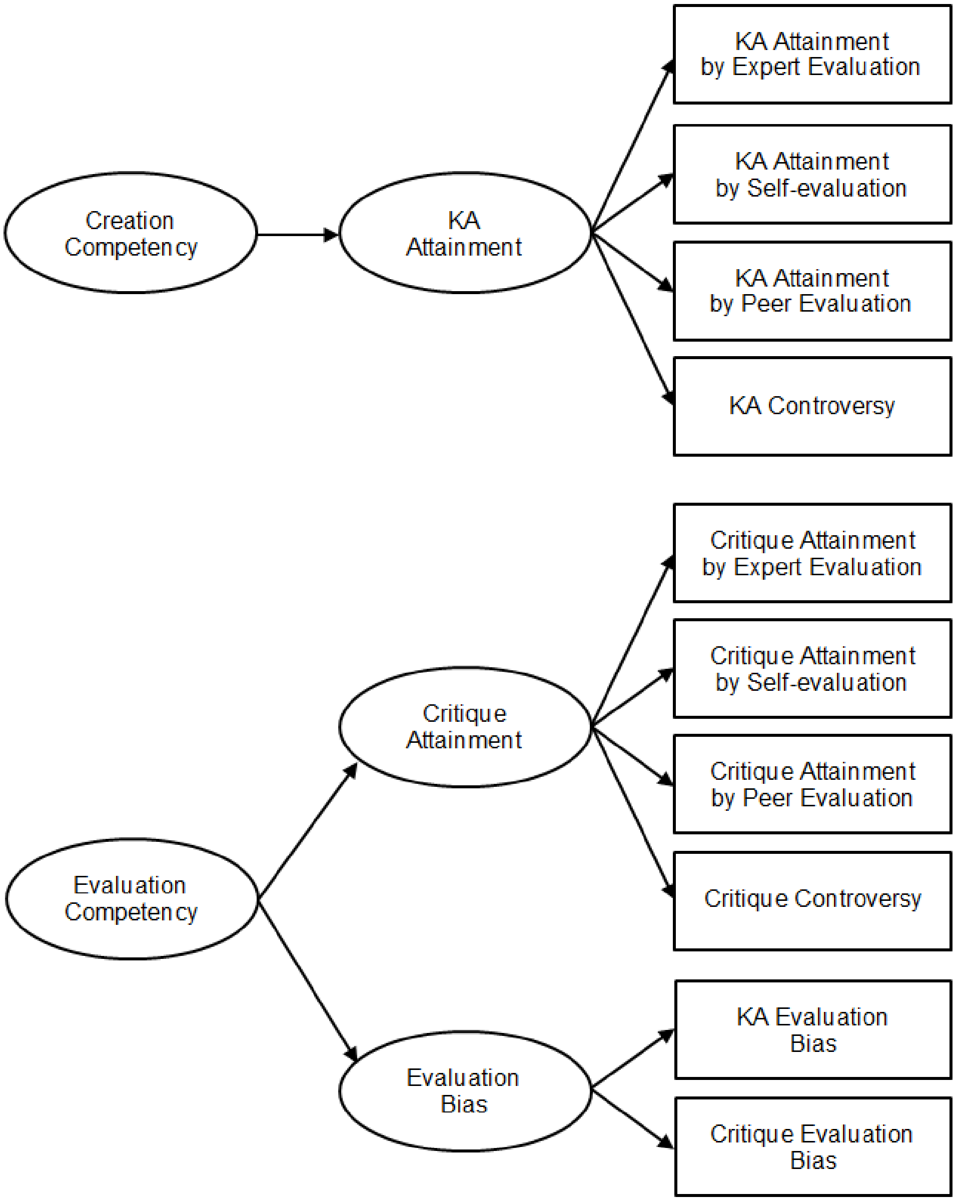

KA Attainment by Expert Evaluation
KA Attainment by Self-evaluation
Creation Competency
KA Attainment
KA Attainment by Peer Evaluation
KA Controversy
Critique Attainment by Expert Evaluation
Critique Attainment by Self-evaluation
Critique Attainment
Critique Attainment by Peer Evaluation
Evaluation Competency
Critique Controversy
KA Evaluation Bias
Evaluation Bias
Critique Evaluation Bias

Figure 12. Variables of Creation and Evaluation Competencies

Table 5. Descriptive Statistics of Attainment by Self-evaluation

Scale	Sub sample	Assign.	Submissions					Critiques				
			N obs	Mean	StDev	Min	Max	N obs	Mean	StDev	Min	Max
Ordinal	Entire sample	1	90	3.41	1.32	1	5	89	2.95	1.61	1	5
		2	90	3.44	1.31	1	5	81	3.47	1.33	1	5
		3	88	3.34	1.37	1	5	86	3.13	1.39	1	5
		4	89	3.38	1.36	1	5	89	3.19	1.34	1	5
		5	87	3.39	1.42	1	5	86	3.24	1.56	1	5
		OA	444	3.39	1.35	1	5	431	3.19	1.42	1	5
	Undergraduate	1	51	3.41	1.33	1	5	50	2.80	1.40	1	5
		2	52	3.38	1.29	1	5	48	3.47	1.30	1	5
		3	51	3.54	1.33	1	5	49	3.25	1.35	1	5
		4	52	3.29	1.38	1	5	52	3.25	1.33	1	5
		5	50	3.40	1.38	1	5	49	3.32	1.56	1	5
		OA	256	3.40	1.33	1	5	248	3.22	1.40	1	5
	Graduate	1	39	3.40	1.33	1	5	39	3.15	1.53	1	5
		2	38	3.53	1.34	1	5	33	3.48	1.39	1	5
		3	37	3.05	1.40	1	5	37	2.96	1.45	1	5
		4	37	3.51	1.33	1	5	37	3.11	1.37	1	5
		5	37	3.39	1.50	1	5	37	3.14	1.58	1	5
		OA	188	3.38	1.38	1	5	183	3.16	1.46	1	5
Cardinal	Entire sample	1	90	75.28	14.28	31	100	89	76.58	12.84	30	100
		2	90	75.12	14.76	42	100	81	78.57	12.03	45	100
		3	88	77.91	13.39	39	100	86	79.64	12.56	50	100
		4	89	76.65	15.29	10	100	89	79.04	14.46	10	100
		5	87	77.72	14.96	13	100	86	78.77	14.32	37	100
		OA	444	76.52	14.53	10	100	431	78.51	13.28	10	100
	Undergraduate	1	51	72.41	14.83	31	100	50	73.00	13.65	30	100
		2	52	70.73	14.82	45	100	48	75.79	13.07	45	100
		3	51	75.57	13.71	39	100	49	77.84	13.79	50	100
		4	52	73.31	14.16	31	100	52	76.13	13.24	28	100
		5	50	76.70	16.87	13	100	49	74.88	15.86	37	100
		OA	256	73.72	14.95	13	100	248	75.52	13.93	28	100
	Graduate	1	39	79.03	12.75	50	99	39	81.18	10.15	52	100
		2	38	81.13	12.54	42	100	33	82.61	9.09	58	96
		3	37	81.14	12.40	55	98	37	82.03	10.41	55	98
		4	37	81.35	15.75	10	100	37	83.14	15.26	10	100
		5	37	79.11	12.00	48	100	37	83.92	10.05	61	100
		OA	188	80.34	13.05	10	100	183	82.56	11.16	10	100

Table 6. Descriptive Statistics of Attainment by Peer Evaluation

Scale	Sub sample	Assign.	Submissions					Critiques				
			N obs	Mean	StDev	Min	Max	N obs	Mean	StDev	Min	Max
Ordinal	Entire sample	1	97	3.00	1.16	1.00	5.00	91	3.02	1.05	1.00	5.00
		2	97	3.02	1.01	1.00	5.00	90	3.02	0.94	1.00	5.00
		3	91	3.02	1.07	1.00	5.00	88	3.01	0.95	1.00	5.00
		4	90	3.00	1.03	1.00	5.00	89	3.00	0.99	1.00	4.67
		5	88	3.00	1.00	1.00	5.00	90	3.01	0.98	1.00	5.00
		OA	463	3.01	1.05	1.00	5.00	448	3.01	0.98	1.00	5.00
	Undergraduate	1	56	3.00	1.22	1.00	5.00	51	3.04	1.07	1.00	5.00
		2	56	3.02	1.04	1.00	5.00	52	3.01	0.91	1.00	5.00
		3	53	3.02	1.05	1.00	5.00	51	3.01	0.92	1.00	5.00
		4	53	3.01	1.05	1.00	5.00	52	3.00	1.02	1.00	4.67
		5	51	3.00	0.94	1.00	4.67	53	3.01	1.02	1.00	4.75
		OA	269	3.01	1.06	1.00	5.00	259	3.01	0.98	1.00	5.00
	Graduate	1	41	3.00	1.09	1.00	5.00	40	3.00	1.04	1.00	4.60
		2	41	3.01	0.97	1.20	5.00	38	3.03	0.99	1.00	5.00
		3	38	3.02	1.10	1.00	5.00	37	3.00	1.01	1.00	5.00
		4	37	3.00	1.03	1.00	5.00	37	3.00	0.94	1.00	4.67
		5	37	3.00	1.08	1.00	5.00	37	3.00	0.95	1.00	5.00
		OA	194	3.00	1.05	1.00	5.00	189	3.01	0.98	1.00	5.00
Cardinal	Entire sample	1	97	67.93	15.10	12.00	90.00	91	73.27	12.06	38.50	95.33
		2	97	67.04	15.53	3.75	92.25	90	71.63	11.75	36.67	98.50
		3	91	70.08	14.88	26.75	96.67	88	74.48	10.46	47.50	95.00
		4	90	69.74	13.16	15.50	91.75	89	75.00	10.22	50.67	93.33
		5	88	70.81	11.43	43.80	93.50	90	74.44	10.48	48.20	98.00
		OA	463	69.07	14.16	3.75	96.67	448	73.76	11.05	36.67	98.50
	Undergraduate	1	56	64.22	15.66	12.00	88.00	51	71.54	10.91	38.50	86.50
		2	56	60.95	15.02	3.75	84.33	52	68.87	10.18	47.75	90.25
		3	53	66.74	14.24	32.00	91.25	51	72.02	8.33	50.50	85.67
		4	53	66.98	11.68	41.50	91.75	52	72.34	9.51	50.80	86.60
		5	51	67.24	10.38	43.80	87.25	53	70.75	10.24	48.20	89.00
		OA	269	65.15	13.73	3.75	91.75	259	71.10	9.88	38.50	90.25
	Graduate	1	41	73.00	12.81	41.00	90.00	40	75.49	13.20	52.67	95.33
		2	41	75.36	12.07	47.25	92.25	38	75.41	12.80	36.67	98.50
		3	38	74.73	14.69	26.75	96.67	37	77.88	12.16	47.50	95.00
		4	37	73.70	14.27	15.50	89.75	37	78.74	10.14	50.67	93.33
		5	37	75.73	11.08	46.00	93.50	37	79.73	8.46	62.00	98.00
		OA	194	74.49	12.94	15.50	96.67	189	77.41	11.54	36.67	98.50

Table 7. Miscalibration with Respect to Peer Evaluation

Scale	Sub sample	Assign.	Submissions					Critiques				
			N obs	Mean	StDev	Min	Max	N obs	Mean	StDev	Min	Max
Ordinal	Entire sample	1	90	0.41	1.13	-2.33	3.00	89	-0.08	1.58	-2.67	4.00
		2	90	0.37	1.36	-4.00	3.00	81	0.41	1.53	-3.33	4.00
		3	88	0.26	1.24	-3.00	3.00	86	0.10	1.46	-3.33	3.33
		4	89	0.36	1.32	-3.00	4.00	89	0.19	1.49	-3.33	3.60
		5	87	0.39	1.27	-2.67	3.73	86	0.21	1.59	-3.33	3.67
		OA	444	0.36	1.26	-4.00	4.00	431	0.16	1.53	-3.33	4.00
	Undergraduate	1	51	0.40	1.05	-2.33	2.67	50	-0.24	1.37	-2.67	2.67
		2	52	0.27	1.14	-2.80	2.80	48	0.43	1.42	-2.00	4.00
		3	51	0.47	1.11	-2.40	3.00	49	0.21	1.22	-1.67	2.40
		4	52	0.26	1.22	-3.00	2.67	52	0.25	1.56	-3.00	3.60
		5	50	0.40	1.17	-1.67	3.73	49	0.27	1.52	-3.00	3.33
		OA	256	0.36	1.13	-3.00	3.73	248	0.18	1.43	-3.00	4.00
	Graduate	1	39	0.42	1.25	-2.33	3.00	39	0.13	1.81	-2.67	4.00
		2	38	0.50	1.62	-4.00	3.00	33	0.39	1.69	-3.33	3.33
		3	37	0.02	1.36	-3.00	2.33	37	0.04	1.74	-3.33	3.33
		4	37	0.51	1.44	-2.67	4.00	37	0.11	1.40	-3.33	3.33
		5	37	0.39	1.41	-2.67	2.67	37	0.14	1.70	-3.33	3.67
		OA	188	0.36	1.42	-4.00	4.00	183	0.14	1.66	-3.33	4.00
Cardinal	Entire sample	1	90	7.46	14.70	-31.00	39.20	89	3.41	16.47	-36.50	46.50
		2	90	7.37	15.67	-27.50	48.50	81	5.82	14.42	-36.00	33.80
		3	88	7.16	15.13	-23.00	47.50	86	5.12	14.62	-22.25	37.00
		4	89	6.80	15.52	-47.75	54.33	89	4.04	17.03	-78.25	44.50
		5	87	6.79	17.08	-55.50	52.20	86	4.02	16.92	-40.75	44.00
		OA	444	7.12	15.57	-55.50	54.33	431	4.45	15.91	-78.25	46.50
	Undergraduate	1	51	8.34	14.85	-31.00	35.75	50	1.68	17.20	-36.50	46.50
		2	52	8.62	16.77	-24.00	48.50	48	6.31	15.88	-36.00	33.80
		3	51	8.23	14.40	-22.50	47.50	49	5.84	15.32	-21.25	37.00
		4	52	6.19	16.76	-47.75	42.00	52	3.79	16.04	-53.75	44.50
		5	50	9.32	18.17	-55.50	52.20	49	3.88	19.77	-40.75	44.00
		OA	256	8.13	16.15	-55.50	52.20	248	4.28	16.86	-53.75	46.50
	Graduate	1	39	6.31	14.62	-26.25	39.20	39	5.62	15.41	-27.00	30.00
		2	38	5.66	14.07	-27.50	37.25	33	5.11	12.19	-25.75	28.75
		3	37	5.68	16.16	-23.00	45.25	37	4.15	13.78	-22.25	33.67
		4	37	7.65	13.79	-16.33	54.33	37	4.40	18.54	-78.25	37.33
		5	37	3.37	15.06	-21.25	38.00	37	4.19	12.43	-26.00	32.67
		OA	188	5.74	14.66	-27.50	54.33	183	4.69	14.57	-78.25	37.33

In keeping with the approach typically taken in psychometric literature, the inter-observer reliability of the two instructor evaluator was assessed as Spearman's *ρ* pairwise correlation coefficient for ordinal data (rankings) and as Pearson's *r* pairwise correlation coefficient for cardinal data (ratings) (Cho, Schunn, & Wilson, 2006; Haaga, 1993; Li, Liu, & Zhou, 2012). Inter-observer reliability for ranking data varied in the range between 0.27 and 0.58 for separate assignments (or 0.49 for the overall sample) in evaluating Submissions; and in the range between 0. 71 and 0.88 for separate assignments (or 0.80 for the overall sample) in evaluating Critiques. Inter-observer reliability for rating data varied in the range between 0.23 and 0.65 for separate assignment (or 0.47 for the overall sample) in evaluating Submissions; and in the range between 0.55 and 0.89 for separate assignment (or 0.74 for the overall sample) in evaluating Critiques. While arguably the values of the inter-observer reliability were unacceptably low, with the exceptions of some assignments the experts were able to reach reliability of about 55% for Submissions and 75% for Critiques. Evidently, complex-problem solutions were difficult to evaluate objectively even for more competent actors. The reliability data demonstrated that, despite the availability of rubrics and reasonable proficiency of the instructor evaluators in the subject matter, one or both of them either experienced some idiosyncratic shocks in their perceptions of attainment of Submissions and Critiques, or demonstrated systematic social or psychological biases when evaluating specific assignments. The expert evaluation for each KA was then computed as the arithmetic average of the two instructor evaluations wherever both were available or as the instructor evaluation for the KAs evaluated by a single instructor.

Table 8. Descriptive Statistics of Attainment by Expert Evaluation (Ordinal Scale)

KA	Subsample	Assign.	N obs	Instructor 1				Instructor 2				Reliability*		Aggregate	
				Mean	StDev	Min	Max	Mean	StDev	Min	Max	N obs	Stat	Mean	StDev
Submissions	Entire sample	1	78	3.00	1.43	1	5	3.00	1.43	1	5	60	0.55	3.00	1.33
		2	80	3.00	1.41	1	5	3.00	1.41	1	5	63	0.58	3.00	1.30
		3	76	3.00	1.42	1	5	3.04	1.41	1	5	61	0.49	3.02	1.29
		4	74	3.04	1.42	1	5	3.01	1.42	1	5	58	0.27	3.04	1.22
		5	73	3.00	1.43	1	5	3.00	1.43	1	5	58	0.52	3.00	1.31
		OA	381	3.01	1.41	1	5	3.01	1.41	1	5	300	0.49	3.01	1.28
	Undergraduate	1	37	3.00	1.46	1	5	3.00	1.45	1	5	19	0.27	3.00	1.36
		2	39	3.00	1.40	1	5	3.00	1.40	1	5	22	0.75	3.00	1.36
		3	38	3.00	1.41	1	5	3.00	1.41	1	5	23	0.43	3.00	1.32
		4	37	3.09	1.39	1	5	3.02	1.39	1	5	21	0.30	3.08	1.27
		5	36	3.00	1.43	1	5	3.00	1.43	1	5	21	0.28	3.00	1.31
		OA	187	3.02	1.40	1	5	3.00	1.40	1	5	106	0.42	3.01	1.32
	Graduate	1	41	3.00	1.43	1	5	3.00	1.43	1	5	41	0.68	3.00	1.30
		2	41	3.00	1.42	1	5	3.00	1.42	1	5	41	0.50	3.00	1.23
		3	38	3.00	1.45	1	5	3.08	1.42	1	5	38	0.53	3.04	1.25
		4	37	3.00	1.46	1	5	3.00	1.46	1	5	37	0.27	3.00	1.16
		5	37	3.00	1.46	1	5	3.00	1.46	1	5	37	0.66	3.00	1.33
		OA	194	3.00	1.43	1	5	3.02	1.42	1	5	194	0.53	3.01	1.24
Critiques	Entire sample	1	74	3.00	1.45	1	5	3.00	1.46	1	5	57	0.71	3.00	1.41
		2	73	3.00	1.43	1	5	3.00	1.43	1	5	58	0.73	3.00	1.36
		3	75	3.00	1.43	1	5	3.00	1.44	1	5	60	0.84	3.00	1.39
		4	73	3.00	1.43	1	5	2.99	1.40	1	5	58	0.88	2.99	1.37
		5	73	3.04	1.42	1	5	3.00	1.43	1	5	57	0.82	3.01	1.37
		OA	368	3.01	1.42	1	5	3.00	1.42	1	5	290	0.80	3.00	1.38
	Undergraduate	1	34	3.00	1.49	1	5	3.00	1.51	1	5	17	0.65	3.00	1.48
		2	35	3.00	1.44	1	5	3.00	1.42	1	5	20	0.76	3.00	1.38
		3	38	3.00	1.41	1	5	3.00	1.43	1	5	23	0.90	3.00	1.41
		4	36	3.00	1.43	1	5	2.98	1.35	1	5	21	0.95	2.98	1.37
		5	36	3.07	1.39	1	5	3.00	1.42	1	5	20	0.93	3.02	1.39
		OA	179	3.01	1.41	1	5	3.00	1.41	1	5	101	0.84	3.00	1.40
	Graduate	1	40	3.00	1.43	1	5	3.00	1.43	1	5	40	0.73	3.00	1.33
		2	38	3.00	1.45	1	5	3.00	1.45	1	5	38	0.71	3.00	1.34
		3	37	3.00	1.46	1	5	3.00	1.46	1	5	37	0.78	3.00	1.38
		4	37	3.00	1.46	1	5	3.00	1.46	1	5	37	0.85	3.00	1.40
		5	37	3.00	1.46	1	5	3.00	1.46	1	5	37	0.76	3.00	1.38
		OA	189	3.00	1.44	1	5	3.00	1.44	1	5	189	0.77	3.00	1.35

* Spearman's ρ

Table 9. Descriptive Statistics of Attainment by Expert Evaluation (Cardinal Scale)

KA	Subsample	Assign.	N obs	Instructor 1				Instructor 2				Reliability*		Aggregate	
				Mean	StDev	Min	Max	Mean	StDev	Min	Max	N obs	Stat	Mean	StDev
Submissions	Entire sample	1	78	88.47	14.05	0	100	78.94	16.19	0	100	60	0.61	83.65	14.05
		2	80	85.81	18.67	0	100	78.71	15.71	45	100	63	0.41	82.01	16.90
		3	76	88.92	9.68	50	100	82.04	12.13	60	100	61	0.39	85.51	10.60
		4	74	82.97	20.09	0	100	76.68	14.78	20	97	58	0.23	78.84	15.04
		5	73	81.92	12.34	50	100	74.44	15.66	5	100	58	0.65	78.03	13.74
		OA	381	85.87	15.73	0	100	78.21	15.11	0	100	300	0.47	81.67	14.46
	Undergraduate	1	37	92.81	6.90	75	100	75.34	12.78	50	96	19	0.32	83.83	11.77
		2	39	90.38	17.69	0	100	75.00	17.51	45	99	22	0.75	82.12	18.75
		3	38	91.92	7.42	75	100	76.32	12.67	60	99	23	0.65	84.56	11.70
		4	37	83.19	20.34	0	100	70.32	12.76	40	95	21	0.47	76.02	15.69
		5	36	81.36	12.80	50	100	65.28	14.03	5	92	21	0.54	74.50	14.49
		OA	187	88.40	14.82	0	100	72.55	14.54	5	99	106	0.54	80.31	15.20
	Graduate	1	41	84.56	17.44	0	100	82.27	18.35	0	100	41	0.77	83.41	16.81
		2	41	81.46	18.75	0	100	82.24	13.04	60	100	41	0.59	81.85	14.23
		3	38	85.92	10.78	50	100	87.76	8.37	70	100	38	0.68	86.84	8.80
		4	37	82.76	20.11	0	100	83.03	14.05	20	97	37	0.17	82.89	13.24
		5	37	82.46	12.03	50	98	83.35	11.56	65	100	37	0.75	82.91	11.05
		OA	194	83.43	16.23	0	100	83.69	13.58	0	100	194	0.56	83.56	13.18
Critiques	Entire sample	1	78	79.26	17.45	30	100	84.65	12.73	60	100	57	0.68	81.36	13.59
		2	80	70.64	25.25	15	100	82.89	14.30	25	100	58	0.55	76.13	19.62
		3	76	71.92	26.20	10	100	88.22	10.86	60	100	60	0.85	79.43	19.58
		4	74	68.86	26.64	10	100	81.89	13.31	50	100	58	0.89	74.37	21.14
		5	73	71.49	27.61	10	100	78.86	13.36	50	100	57	0.78	73.82	21.05
		OA	381	72.45	25.02	10	100	83.29	13.27	25	100	290	0.74	77.03	19.32
	Undergraduate	1	37	80.85	13.23	60	100	77.74	11.78	60	95	17	0.71	79.12	12.19
		2	39	68.89	26.14	15	100	80.22	13.28	58	100	20	0.65	74.13	20.97
		3	38	66.03	28.80	10	100	86.75	10.45	60	100	23	0.85	76.25	21.38
		4	37	64.50	31.06	10	98	80.57	14.08	50	99	21	0.92	71.67	23.78
		5	36	64.92	31.67	10	100	74.86	14.19	50	95	21	0.89	69.19	23.44
		OA	187	68.85	27.57	10	100	80.03	13.32	50	100	102	0.77	74.02	20.95
	Graduate	1	41	77.90	20.44	30	100	90.53	10.44	60	100	40	0.84	84.21	14.86
		2	41	72.26	24.63	15	95	85.50	14.94	25	100	38	0.53	78.88	17.49
		3	38	77.97	22.01	20	100	89.65	11.20	60	100	37	0.85	83.81	16.05
		4	37	73.11	21.07	20	100	83.22	12.55	60	100	37	0.87	78.16	16.31
		5	37	78.05	21.32	20	98	82.86	11.30	60	100	37	0.66	80.46	15.00
		OA	194	75.87	21.86	15	100	86.41	12.47	25	100	189	0.71	81.14	15.99

** Pearson's r

Table 10. Miscalibration with Respect to Expert Evaluation

Scale	Sub sample	Assign	Submissions					Critiques				
			N obs	Mean	StDev	Min	Max	N obs	Mean	StDev	Min	Max
Ordinal	Entire sample	1	90	0.42	1.34	-4.00	3.60	89	-0.06	1.67	-4.00	4.00
		2	90	0.38	1.53	-4.00	4.00	81	0.40	1.63	-3.20	4.00
		3	88	0.27	1.25	-2.67	3.00	86	0.12	1.41	-2.67	4.00
		4	89	0.32	1.68	-4.00	4.00	89	0.20	1.68	-4.00	4.00
		5	87	0.37	1.38	-2.00	4.00	86	0.21	1.47	-2.80	4.00
		OA	444	0.35	1.44	-4.00	4.00	431	0.17	1.57	-4.00	4.00
	Undergraduate	1	51	0.40	1.32	-4.00	3.00	50	-0.19	1.37	-4.00	2.67
		2	52	0.28	1.39	-4.00	3.20	48	0.41	1.68	-3.20	4.00
		3	51	0.48	1.17	-2.00	3.00	49	0.23	1.36	-2.40	4.00
		4	52	0.19	1.67	-4.00	3.50	52	0.27	1.73	-3.00	4.00
		5	50	0.36	1.49	-2.00	4.00	49	0.27	1.53	-2.80	4.00
		OA	256	0.34	1.41	-4.00	4.00	248	0.20	1.54	-4.00	4.00
	Graduate	1	39	0.44	1.39	-3.00	3.60	39	0.10	1.99	-4.00	4.00
		2	38	0.53	1.71	-2.00	4.00	33	0.38	1.58	-2.50	4.00
		3	37	-0.03	1.31	-2.67	2.67	37	-0.04	1.47	-2.67	4.00
		4	37	0.51	1.69	-2.67	4.00	37	0.11	1.62	-4.00	4.00
		5	37	0.39	1.22	-1.50	4.00	37	0.14	1.40	-2.67	3.00
		OA	188	0.37	1.48	-3.00	4.00	183	0.13	1.62	-4.00	4.00
Cardinal	Entire sample	1	90	-8.18	17.65	-50.00	64.00	89	-5.00	15.97	-43.00	38.00
		2	90	-7.55	18.11	-50.50	50.00	81	1.62	22.52	-49.50	79.00
		3	88	-8.23	12.55	-51.00	21.00	86	0.31	21.42	-46.00	70.00
		4	89	-2.85	18.80	-54.00	65.00	89	4.67	24.96	-82.00	65.00
		5	87	-0.89	17.50	-55.00	44.00	86	4.41	23.13	-55.50	80.00
		OA	444	-5.57	17.27	-55.00	65.00	431	1.18	21.98	-82.00	80.00
	Undergraduate	1	51	-11.53	17.73	-50.00	40.00	50	-6.05	15.06	-41.00	25.00
		2	52	-12.51	18.78	-50.50	50.00	48	1.17	25.32	-49.50	79.00
		3	51	-9.66	12.30	-51.00	21.00	49	1.89	24.42	-46.00	70.00
		4	52	-3.79	20.52	-54.00	65.00	52	4.46	26.33	-67.00	65.00
		5	50	1.26	19.79	-55.00	44.00	49	5.12	26.71	-55.50	80.00
		OA	256	-7.29	18.67	-55.00	65.00	248	1.33	24.10	-67.00	80.00
	Graduate	1	39	-3.79	16.76	-38.50	64.00	39	-3.65	17.16	-43.00	38.00
		2	38	-0.76	14.86	-36.50	38.00	33	2.29	18.05	-23.00	48.00
		3	37	-6.26	12.80	-32.50	19.50	37	-1.78	16.73	-34.50	34.00
		4	37	-1.54	16.26	-32.50	43.00	37	4.97	23.27	-82.00	60.00
		5	37	-3.80	13.52	-28.50	29.00	37	3.46	17.62	-31.50	42.50
		OA	188	-3.22	14.90	-38.50	64.00	183	0.98	18.80	-82.00	60.00

Table 11. Correlations (Ordinal-scale Data) (N obs = 431)

		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]	[17]
ScrkStuAr_Self	[1]	1.00																
ScrkStuAr_Attm	[2]	0.47	1.00															
ScrkStuAr_Misc	[3]	0.68	-0.33	1.00														
ScrkInsAr_Exp	[4]	0.42	0.54	0.00	1.00													
ScrkInsAr_Misc	[5]	0.57	-0.04	0.64	-0.51	1.00												
ScrkStuAr_ContDFM	[6]	0.02	0.01	0.02	0.03	-0.01	1.00											
ScrkStuAr_ContDFC	[7]	0.09	0.18	-0.05	0.12	-0.02	0.75	1.00										
ScrkStuAr_BiasDFM	[8]	-0.03	-0.08	0.03	-0.09	0.05	0.41	0.18	1.00									
ScrkStuAr_BiasDFC	[9]	-0.04	-0.03	-0.01	-0.03	0.00	0.17	0.39	0.64	1.00								
ScrkStuCr_Self	[10]	0.40	0.16	0.29	0.17	0.22	0.06	0.02	0.00	-0.04	1.00							
ScrkStuCr_Attm	[11]	0.14	0.24	-0.06	0.21	-0.06	0.09	0.12	-0.12	-0.09	0.23	1.00						
ScrkStuCr_Misc	[12]	0.29	-0.01	0.31	0.02	0.25	0.00	-0.06	0.08	0.02	0.78	-0.43	1.00					
ScrkInsCr_Exp	[13]	0.28	0.35	0.01	0.37	-0.07	0.04	0.07	-0.09	-0.08	0.37	0.60	-0.04	1.00				
ScrkInsCr_Misc	[14]	0.12	-0.16	0.25	-0.17	0.26	0.02	-0.04	0.08	0.03	0.58	-0.32	0.75	-0.54	1.00			
ScrkStuCr_ContDFM	[15]	0.05	0.05	0.01	0.07	-0.02	0.05	0.07	0.02	-0.02	0.03	0.11	-0.04	0.13	-0.09	1.00		
ScrkStuCr_ContDFC	[16]	0.05	0.05	0.01	0.09	-0.03	-0.03	0.22	-0.12	0.19	0.03	0.24	-0.13	0.21	-0.16	0.69	1.00	
ScrkStuCr_BiasDFM	[17]	0.03	0.00	0.04	-0.03	0.05	0.05	0.02	0.06	0.00	0.01	-0.12	0.09	-0.09	0.09	0.35	0.15	1.00
ScrkStuCr_BiasDFC	[18]	0.07	0.04	0.04	0.04	0.03	-0.05	0.22	-0.11	0.27	0.00	-0.03	0.02	-0.04	0.03	0.08	0.43	0.57

Table 12. Correlations (Cardinal-scale Data) (N obs = 431)

		[19]	[20]	[21]	[22]	[23]	[24]	[25]	[26]	[27]	[28]	[29]
RatgStuAr_Self	[19]	1.00										
RatgStuAr_AvR	[20]	0.40	1.00									
RatgStuAr_Misc	[21]	0.57	-0.52	1.00								
RatgInsAr_Exp	[22]	0.26	0.51	-0.21	1.00							
RatgInsAr_Misc	[23]	0.63	-0.07	0.65	-0.58	1.00						
RatgStuAr_EE	[24]	-0.05	-0.09	0.03	-0.05	0.00	1.00					
RatgStuAr_ER	[25]	0.04	-0.11	0.14	-0.14	0.15	0.02	1.00				
RatgStuCr_Self	[26]	0.65	0.35	0.29	0.19	0.40	-0.04	0.04	1.00			
RatgStuCr_AvR	[27]	0.09	0.41	-0.28	0.13	-0.03	0.02	-0.12	0.14	1.00		
RatgStuCr_Misc	[28]	0.48	0.02	0.44	0.07	0.35	-0.04	0.12	0.74	-0.56	1.00	
RatgInsCr_Exp	[29]	0.07	0.32	-0.22	0.22	-0.12	-0.09	-0.05	0.13	0.49	-0.22	1.00
RatgInsCr_Misc	[30]	0.33	-0.07	0.37	-0.08	0.34	0.06	0.07	0.49	-0.34	0.64	-0.80

Treatment Variables

This study examined the impact of two experimental conditions (X_1 and X_2) on the dependent variables over time (multiple assignments). These conditions are represented by the dummy variable *RC* and coded as follows:

$$RC = \begin{cases} 0\ if\ experimental\ condition\ X_1 \\ 1\ if\ experimental\ condition\ X_2 \end{cases}$$

Subjects in both the experimental condition pools (X_1 and X_2) were instructed by the same professor. Only uniform feedback was provided to all participants by presenting more and less successful examples upon the completion of an assignment. No individual feedback was given by the instructor to any participants. “Time” is included in the analysis models as the assignment sequence number.

Control Variables

Students of two categories participated in the experiment – undergraduate and graduate. To control for the differences in participants of these two categories, the dummy variable *GRAD* is included in the analysis models and coded as follows:

$$GRAD = \begin{cases} 0\ if\ undergraduate\ student \\ 1\ if\ graduate\ student \end{cases}$$

Subjects in both the control pools (undergraduate and graduate) were instructed by the same professor. Only uniform feedback was provided to all participants by presenting more and less successful examples upon the completion of an assignment. No individual feedback was given by the instructor to any participants.

Longitudinal Analysis

To answer the research questions, evidence is needed on how creation and evaluation competencies change over time. Since creation and evaluation competencies are not directly observable, they are assessed based on the changes in the goodness of the KAs produced by participants and evaluations by both participants and experts. The results of the pilot study (presented in the proposal of this dissertation's topic) indicated that peer-evaluation and self-evaluation attainment, as well as miscalibration with respect to peer evaluation do not change uniformly for all participants (Babik, Singh, Zhao, & Ford, n.d.). Different participants seemed to have different dynamics (or trajectories) of changes in the results of their Submissions' and Critiques' evaluations. Therefore, it can be hypothesized that actors have diverse dynamics of creation and evaluation competencies. Since longitudinal changes are associated with evaluations of Submissions and Critiques produced by a participant, an actor is the unit of analysis.

Behavior of creation competency is explored by analyzing patterns of Submission attainment calculated from peer, self-, and expert evaluations. In addition, Submission controversy indicates whether peer evaluators in SKCC were unanimous in the attainment evaluation. It can be interpreted, to a certain degree, as whether the creator was able to address effectively his SKCC audience, conditional on whether the audience consisted of competent evaluators.

Changes over time in evaluation competency are studied by analyzing patterns of Critique attainment calculated from peer, self- and expert evaluations, Submission and Critiques evaluation biases, and miscalibration. The bias variables are obtained for evaluations of the original Submissions, as well as Critiques of peers' Submissions. Bias is interpreted as a reflection of an individual peer's deviation from otherwise unanimous evaluation of attainment by other peers.

Based on the conceptual model presented in this study and the literature review on peer assessment and social construction of knowledge, it was hypothesized that changes in creation competency do not follow a common temporal pattern. Therefore, changes in creation and evaluation competencies in the SKCC cannot be detected with conventional statistical techniques, such as t-test or linear regression with respect to time. Instead, unobserved (latent) classes in the actor population may exist that have distinct temporal trajectories of individual understanding of KA attainment in the SKA structure. Moreover, these temporal trajectories may be non-linear. Therefore, Latent Growth Modeling (LGM) is the appropriate technique to study individual longitudinal patters on changes in competencies.

LGM is a longitudinal statistical technique used in the Structural Equation Modeling (SEM) framework to estimate growth trajectory over time (Chan, 2002; Jung & Wickrama, 2008). Latent Growth Models represent repeated measures of dependent variables as a function of time and other measures. In this dissertation, the LGM method is used to identify latent classes of actors that demonstrate different development patterns of creation and evaluation competencies. Evidence of existence of such classes provides a basis for explaining different patterns of competency development over multiple creator-evaluator interactions. This motivates our choice of LGM to investigate trajectories of attainment, controversy, bias, and miscalibration of Submissions and Critiques.

The TRAJ procedure in the STATA 11.1 statistical analysis software was used to identify the number of latent growth classes in the trajectories of dependent variables (Jones, Nagin, & Roeder, 2001). TRAJ estimates a discrete mixture model for longitudinal data groupings (latent classes) that may represent distinct subpopulations in the data. The Bayesian information criterion (BIC) was used to identify the number of classes in the model (Schwarz, 1978). Specifically, 2ΔBIC, i.e., 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 2ΔBIC is approximately equal to $2\ln B_{10}$, where B_{10} is the Bayes factor (Kass & Raftery, 1995). Statistically, $2\ln B_{10}$ 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 & Wasserman, 1995).

As the next step of the analysis, plots of the average values of each dependent variable for each latent growth class, dummy-coded experimental condition, and dummy-coded control block were visually inspected to form preliminary expectations of the dependent variables' behaviors.

Further, on the basis of the number on latent growth classes for each dependent variable obtained from the LGM and coded as dummy variables, Hierarchical Linear

Modeling (HLM) was applied to test for the most parsimonious multi-level models that describe longitudinal patterns of dependent variables. Analysis started with the most generic model, which included up to the cubic trend of time at the 1st level, and the dummy-coded variables of latent classes, experimental conditions, control blocks, and their interactions at the 2nd level:

Level-1 model:

$$y_{iT} = \pi_{0i} + \pi_{1i}{*}T + \pi_{2i}{*}T^2 + \pi_{3i}{*}T^3 + \varepsilon_{iT}$$

Level-2 model:

$$\begin{aligned}
\pi_{0i} = \beta_{00} &+ \beta_{01}{*}GRAD_i + \beta_{02}{*}RC_i + \beta_{03}{*}Class_i + \\
&+ \beta_{04}{*}(GRAD_i{*}RC_i) + \beta_{05}{*}(GRAD_i{*}Class_i) + \beta_{06}{*}(RC_i{*}Class_i) + \\
&+ \beta_{07}{*}(GRAD_i{*}RC_i{*}Class_i) + \delta_{0i} \\
\pi_{1i} = \beta_{10} &+ \beta_{11}{*}GRAD_i + \beta_{12}{*}RC_i + \beta_{13}{*}Class_i + \\
&+ \beta_{14}{*}(GRAD_i{*}RC_i) + \beta_{15}{*}(GRAD_i{*}Class_i) + \beta_{16}{*}(RC_i{*}Class_i) + \\
&+ \beta_{17}{*}(GRAD_i{*}RC_i{*}Class_i) + \delta_{1i} \\
\pi_{2i} = \beta_{20} &+ \beta_{21}{*}GRAD_i + \beta_{22}{*}RC_i + \beta_{23}{*}Class_i + \\
&+ \beta_{24}{*}(GRAD_i{*}RC_i) + \beta_{25}{*}(GRAD_i{*}Class_i) + \beta_{26}{*}(RC_i{*}Class_i) + \\
&+ \beta_{27}{*}(GRAD_i{*}RC_i{*}Class_i) + \delta_{2i} \\
\pi_{3i} = \beta_{30} &+ \beta_{31}{*}GRAD_i + \beta_{33}{*}RC_i + \beta_{33}{*}Class_i + \\
&+ \beta_{34}{*}(GRAD_i{*}RC_i) + \beta_{35}{*}(GRAD_i{*}Class_i) + \beta_{36}{*}(RC_i{*}Class_i) + \\
&+ \beta_{37}{*}(GRAD_i{*}RC_i{*}Class_i) + \delta_{3i}
\end{aligned}$$

where T denotes respective assignment, and for a student i y_{iT} represents a value of the respective dependent variable in the assignment T, $Class_i$ is the latent-class dummy variable for a respective dependent variable, RC_i is the experimental condition dummy variable, $GRAD_i$ is the control dummy variable.

The model was then gradually reduced by eliminating insignificant interactions and variables until the most parsimonious model describing longitudinal patterns of a respective dependent variable was obtained. Cross-tabulation latent growth classes underlying the individual variability of creation and evaluation competencies was used where necessary, to gain additional insights.

CHAPTER IV

RESULTS

The preceding chapter described the methodology used in this study to analyze longitudinal dynamics of the evaluation intersubjectivity constructs in social knowledge creation communities (SKCC). The presentation of results in this chapter is organized in the following manner. First, the LGM results are reported for all dependent variables of interest. Then, the HLM results are presented for both KA types (submissions and critiques) along several dimensions such as dynamics of attainment, miscalibration, controversy, and bias. Results for attainment by self-, peer and expert evaluations are reported for the ordinal and cardinal measurement scales. Results for miscalibration with respect to peer evaluation and with respect to expert evaluation are also reported for the ordinal and cardinal scales. Controversy and bias were investigated only on the ordinal scale; however, they are reported for two alternative computations approaches – *deviation from mean* (DFM) and *deviation from co-evaluators* (DFC).

For every variable and KA type, first, the plots of experimental data for each experimental condition and control block are presented. The results of visual inspection of these plots, as the preliminary step in assessing the data and discovering patterns, are followed by the results of estimations and respective plots for the most parsimonious HLM models. The HLM model building was conducted through a sequence of models including the unconditional means model, the unconditional growth model, the full level-1 model, and, finally, the full level-1 and level-2 model (Singer & Willett, 2003).

A model comparison framework was then used to reduce statistically non-significant fixed effects in the model, beginning with higher order interactions and working down to lower order interactions and main effects (Appelbaum & Cramer, 1974; Cramer & Appelbaum, 1980). The results of cross-tabulating latent classes, experimental conditions and control blocks are presented when they help interpretation or offer additional insights. The summary of key findings is presented in Table 13.

Latent Growth Modeling Results

The numbers of latent growth classes in the dependent variables were tested using LGM methodology. The refinement process through which the most reasonable number of classes was selected for each variable is summarized in Table 14. The 2ΔBIC criterion suggested that the best fitting models have two latent growth classes underlying attainment by self-evaluation, peer evaluation, expert evaluation, miscalibration with respect to peer evaluation, and miscalibration with respect to expert evaluation. This result holds for both ordinal- and cardinal-scale data.

Although in the pilot study three latent growth classes underlying attainment by peer evaluation and miscalibration, and four classes underlying attainment by self-evaluation were found, only two latent growth classes were found in the experimental study in this dissertation (Babik et al., n.d.). The evidence of two latent growth classes, however, is robust with respect to the use of the alternative measurement scales. The major differences between the current sample and the pilot study sample are the number of participants (98 and 435 respectively), and the number of longitudinal observations (five and nine respectively). Despite the smaller number of participants and observation points, the experimental study provided greater control over various sources of variance in the current sample, such as peer group size, student level, assignment type, and rubrics.

Exploratory LGM analysis suggested significant linear, but not quadratic or cubic, trends of change in attainment by self-evaluation, peer evaluation, and expert evaluation, as well as miscalibration with respect to peer evaluation and expert evaluation. This result is also inconsistent with the findings of significant cubic trends in the pilot study, but can be explained by the smaller number of longitudinal observations.

The LGM analysis revealed no strong evidence of the existence of latent growth classes underlying evaluation controversy and evaluator bias. This result holds for controversy and bias computed (using the ordinal scale data) as deviation from mean as well as deviation from co-evaluators. The lack of evidence for latent growth classes underlying controversy and bias is consistent with the findings of the pilot study. Developing methodology for computing controversy and bias based on the cardinal scale and testing the hypotheses regarding latent growth classes for these variables is left outside the scope of this dissertation because of the reasons explained in chapter III.

The LGM analysis provided information on the number on latent growth classes for each dependent variable, classified participants in these latent classes, and allowed further investigation of trajectories of attainment and miscalibration with the Hierarchical Linear Modeling (HLM). In the following subsection, dynamics of attainment of the original KA submissions and their critiques evaluated by their authors, peers and experts are considered with the aim to compare and contrast these three different perspectives at the goodness of Submissions and Critiques and to model longitudinal patterns of changes in creation competencies.

Table 13. Summary of Key Findings

Finding		Submissions	Critiques
Number of latent classes			
Attainment by	Self-evaluation	2	2
	Peer evaluation	2	2
	Expert evaluation	2	2
Miscalibration	Peer evaluation	2	2
Controversy		No latent classes	No latent classes
Bias		No latent classes	No latent classes
Change over time			
Miscalibration classes (ordinal scale)	Underconfident	Linear increase	Linear decrease
	Overconfident	Linear increase	Linear decrease
Miscalibration classes (cardinal scale)	Calibrating	Flat	Flat
	Overconfident	Flat	Flat
Controversy		Non-linear non-monotonic	Non-linear non-monotonic
Bias		Non-linear non-monotonic	Non-linear non-monotonic
Effect of experimental conditions			
Attainment			
	Self-evaluation	No effect	No effect
	Peer evaluation	No effect	No effect
	Expert evaluation	No effect	No effect
Miscalibration	Peer evaluation	No effect	No effect
Controversy		Faster decrease in random	Faster decrease in random
Bias		Faster decrease in random	Faster decrease in random

Table 14. Tabulated BIC and 2ΔBIC for Attainment, Miscalibration, Controversy, and Bias

Submissions, ordinal scale (ranking)

N of classes	Peer evaluation		Self-evaluation		Misc. peer eval.		Expert evaluation		Misc. expert eval.		Bias DFM		Controversy DFM		Bias DFC		Controversy DFC	
	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC
1	-718.44		-764.83		-741.19		-807.52		-803.88		-164.54		-215.04		-115.16		-268.86	
2	-707.09	**22.70**	-764.83	**0**	-725.40	**31.58**	-801.82	**11.40**	-789.21	**29.34**	-174.91	-20.74	-222.08	-14.08	-122.69	-15.06	-274.76	-11.80
3	-711.77	-9.36	-764.83	0	-728.53	-6.26	-805.35	-7.06	-794.23	-10.04	-184.85	-19.88	-231.09	-18.02	-125.59	-5.80	-281.59	-13.66
4	-713.66	-3.78	-764.83	0	-735.36	-13.66	-810.65	-10.60	-800.32	-12.18	-195.08	-20.46	-239.73	-17.28	-135.19	-19.20	-291.48	-19.78
N obs	98		95		95		98		95		96		98		96		98	

Critiques, ordinal scale (ranking)

N of classes	Peer evaluation		Self-evaluation		Misc. peer eval.		Expert evaluation		Misc. expert eval.		Bias DFM		Controversy DFM		Bias DFC		Controversy DFC	
	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC
1	-657.15		-752.33		-803.27		-795.23		-822.01		-166.87		-217.09		-185.44		-256.28	
2	-641.79	**30.72**	-737.72	**29.22**	-794.56	**17.42**	-760.40	**69.66**	-816.82	**10.38**	-174.07	-14.40	-224.48	-14.78	-192.31	-13.74	-256.72	-0.88
3	-646.97	-10.36	-730.44	14.56	-799.98	-10.84	-760.33	0.14	-822.66	-11.68	-182.99	-17.84	-	-	-196.94	-9.26	-264.61	-15.78
4	-651.73	-9.52	-733.78	-6.68	-805.19	-10.42	-765.82	-10.98	-824.67	-4.02	-192.83	-19.68	-	-	-204.14	-14.40	-273.42	-17.62
N obs	96		95		95		96		95		97		96		97		96	

Submissions, cardinal scale (rating)

N of classes	Peer evaluation		Self-evaluation		Misc. peer eval.		Expert evaluation		Misc. expert eval.	
	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC
1	-1888.92		-1799.90		-1855.09		-1857.37		-1895.80	
2	-1858.06	**61.72**	-1745.09	**109.62**	-1816.56	**77.06**	-1843.65	**27.44**	-1877.81	**35.98**
3	-1854.17	7.78	-1740.85	8.48	-1803.97	25.18	-1846.25	-5.20	-1870.43	14.76
4	-1858.03	-7.72	-1728.69	24.32	-1802.66	2.62	-1842.13	8.24	-1866.51	7.84
N obs	98		95		95		98		95	

Critiques, cardinal scale (rating)

N of classes	Peer evaluation		Self-evaluation		Misc. peer eval.		Expert evaluation		Misc. expert eval.	
	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC	BIC	2ΔBIC
1	-1716.97		-1703.27		-1810.40		-1945.70		-1945.30	
2	-1680.93	**72.08**	-1635.59	**135.36**	-1772.95	**74.90**	-1884.99	**121.42**	-1909.19	**72.22**
3	-1676.37	9.12	-1626.18	18.82	-1767.06	11.78	-1888.43	-6.88	-1907.43	3.52
4	-1681.61	-10.48	-1630.35	-8.34	-1762.07	9.98	-1895.28	-13.70	-1903.19	8.48
N obs	96		95		96		96		95	

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

Dynamics of Attainment

Visual examinations of the plots of average attainment calculated on the basis of self-, peer, and expert evaluations of Submissions and Critiques, using both ordinal and cardinal scales, for the experimental conditions and control blocks were conducted first with the purpose of tentative interpretation. Then, the HLM was then used to identify the most parsimonious models that describe longitudinal trajectories of changes in attainment. Although exploratory LGM analysis suggested significant linear, but not quadratic or cubic, trends of change in attainment by self-evaluation, peer evaluation, and expert evaluation, as well as miscalibration with respect to peer evaluation and expert evaluation, the order of mixture model trajectories was rigorously re-tested starting with the third-order polynomial function.

Self-evaluation

Visual examination of average attainment by self-evaluation on the ordinal scale revealed no evident upward or downward trend for all combinations of experimental conditions and control blocks (Figure 13 and Figure 14). This result holds for both Submissions and Critiques. Average ordinal self-evaluation attainment score fluctuated between 3 and 4 (on the 5-point scale) for Submissions, and slightly gravitated toward 3 for Critiques. There is some visual evidence for average ordinal self-evaluation attainment score being slightly greater for the participants in steady groups compared to randomly recombined groups for both undergraduate and graduate student participants.

Visual examination of average attainment by self-evaluation on the cardinal scale suggested similar tendencies, although a higher average self-evaluation attainment of Submissions in steady groups was more accentuated, particularly for graduate students (Figure 15 and Figure 16). This result was not found in average self-evaluation attainment of Critiques; moreover, undergraduate students in steady groups on average rated their Critiques lower than undergraduate students in randomly recombined groups.

Analysis of Submission attainment by self-evaluation on the ordinal scale produced the most parsimonious model with a significant linear, but not quadratic or cubic, trend and two latent growth classes: the *lower performance*, or the “pessimist”, class comprising around 54% of participants, who self-assess below the median and show declining self-assessment tendency over time; and the *higher performance*, or the “optimist”, class comprising 46% of participants, who self-asses starting near median and tend to self-assess increasingly over time (Table 15 and Figure 17).

The experimental condition variable *RC* had a significant effect on the level-1 model’s intercept (despite the fact that participants were randomly assigned to the experimental conditions, but no significant effect on the slope. *GRAD* had no significant effect on either intercept or slope, and was, therefore, removed from the model. The latent growth class dummy variable had significant effects on both the intercept and slope of the level-1 model. This model indicates that the latent growth classes of attainment by self-evaluation tend to diverge over time, and the mode of group allocation does not have any significant effect on this process.

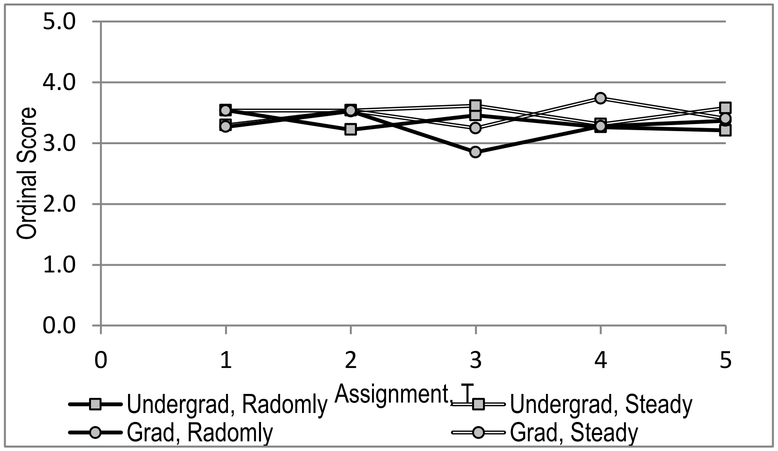


Figure 13. Average Submission Attainment by Self-evaluation on Ordinal Scale

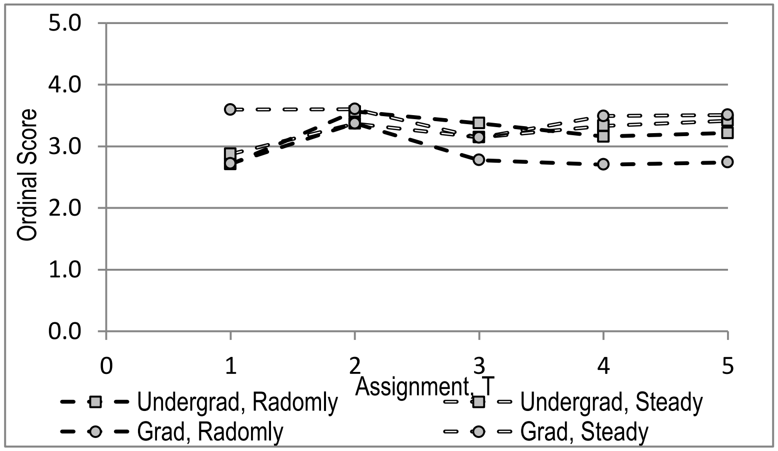


Figure 14. Average Critiques Attainment by Self-evaluation on Ordinal Scale

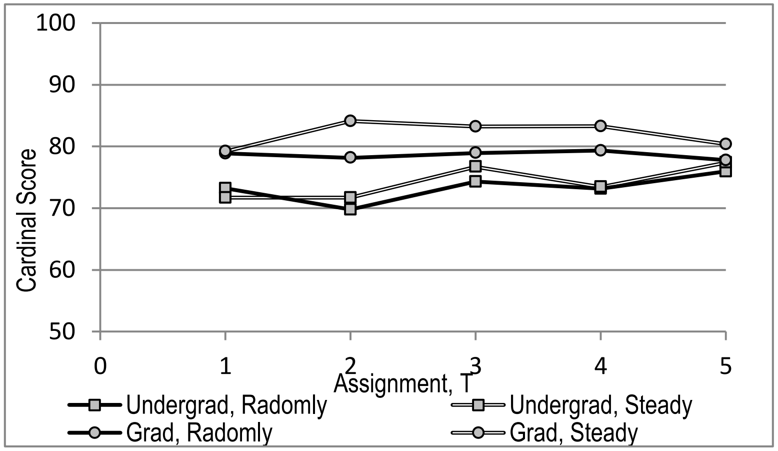


Figure 15. Average Submission Attainment by Self-evaluation on Cardinal Scale

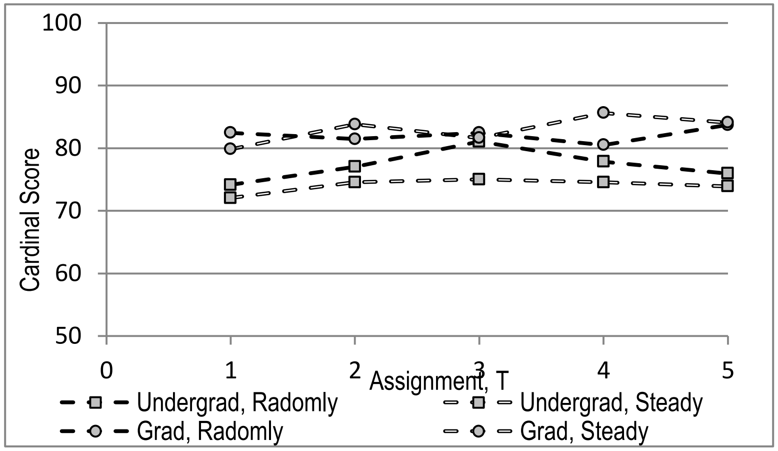


Figure 16. Average Critiques Attainment by Self-evaluation on Cardinal Scale

Analysis of Critiques attainment by self-evaluation on the ordinal scale produced the model with a significant cubic trend and two latent classes: the *lower performance* class comprising around 59% of participants (who self-assess below the median) and the *higher performance* class comprising 41% of participants (who self-asses above median) (Table 16 and Figure 18). Neither class showed any noticeable sloped trend.

Table 15. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Self-evaluation on Ordinal Scale

```
Level-1 Model
ScrkStuAr_Self = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*ScrkStuAr_Self_Class + B02*RC
Π1 = B10 + B11*ScrkStuAr_Self_Class
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            2.759      0.159       17.364        439    0.000
    ...Class, B01            1.056      0.199        5.299        439    0.000
          RC, B02            0.215      0.099        2.176        439    0.030
T slope, Π1
    INTRCPT2, B10           -0.129      0.056       -2.318        439    0.021
    ...Class, B11            0.259      0.075        3.452        439    0.001
```

Table 16. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Self-evaluation on Ordinal Scale

```
Level-1 Model
ScrkStuCr_Self = Π0 + Π1*T + Π2*T2+ Π3* T3+ E
Level-2 Model
Π0 = B00 + B01*ScrkStuCr_Self_Class + R0
Π1 = B10, Π2 = B20, Π3 = B30
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            1.764      0.146       12.100         93    0.000
    ...Class, B01            1.702      0.131       12.985         93    0.000
T slope, Π1
    INTRCPT2, B10            0.666      0.263        2.536        426    0.012
T2 slope, Π2
    INTRCPT2, B20           -0.384      0.171       -2.251        426    0.025
T3 slope, Π3
    INTRCPT2, B30            0.059      0.029        2.017        426    0.044

Random Effect                St Dev   Var Comp        DF     Chi-sq     P-value
INTRCPT1, R0                 0.330      0.109        93     125.768       0.013
  level-1, E                 1.151      1.326
```

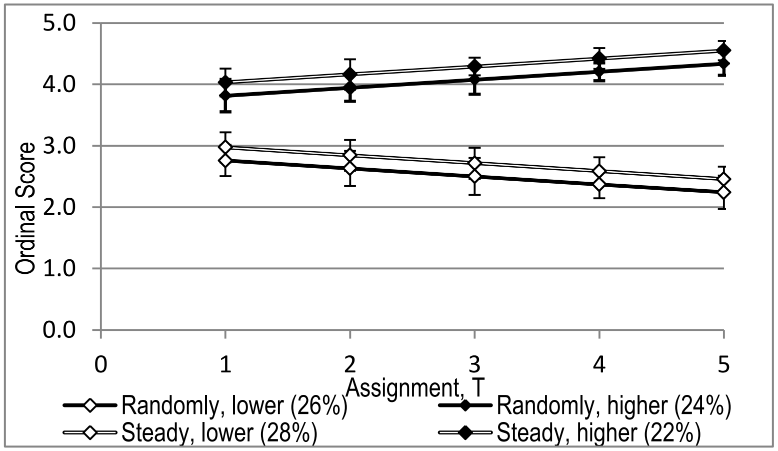


Figure 17. Latent Growth Classes of Submission Attainment by Self-evaluation on Ordinal Scale (with Indicated Standard Errors)

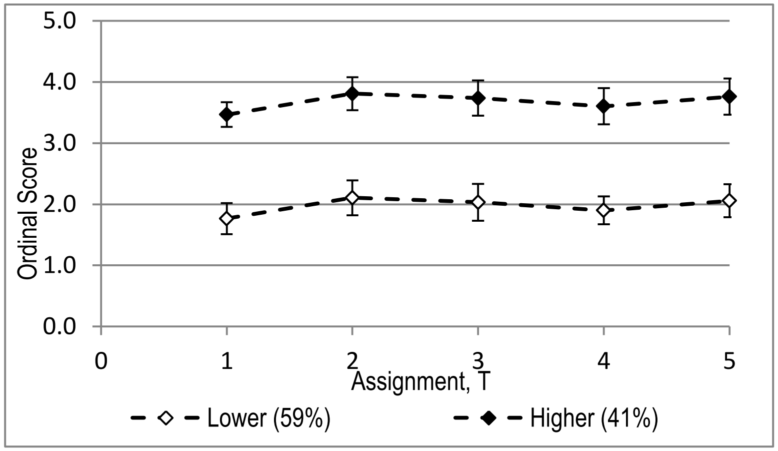


Figure 18. Latent Growth Classes of Submission Attainment by Self-evaluation on Ordinal Scale (with Indicated Standard Errors)

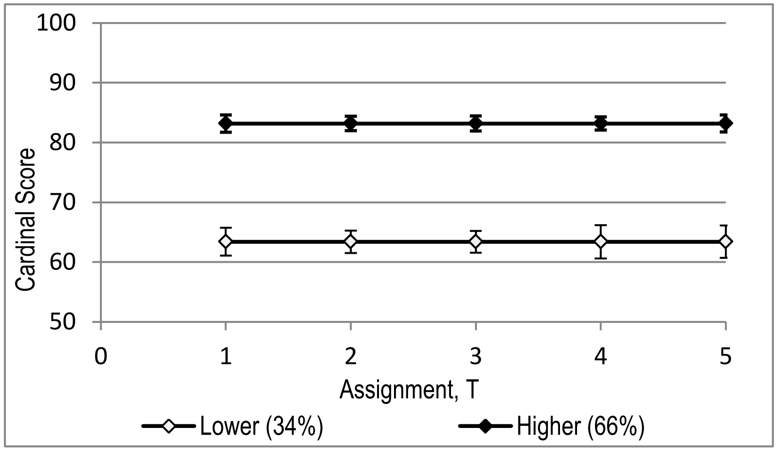


Figure 19. Latent Growth Classes of Submission Attainment by Self-evaluation on Cardinal Scale (with Indicated Standard Errors)

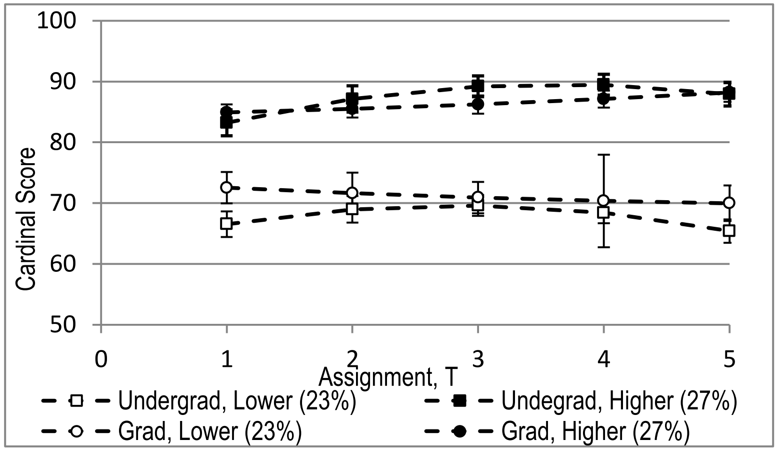


Figure 20. Latent Growth Classes of Critiques Attainment by Self-evaluation on Cardinal Scale (with Indicated Standard Errors)

Analysis of Submission attainment by self-evaluation on the cardinal scale produced the most parsimonious model with two latent classes: the *lower performance* class comprising around 34% of participants with average self-assessment near 63 points, and the *higher performance* class comprising 66% of participants with average self-assessment near 82 points (Table 17 and Figure 19). The latent class variable had no significant effect on the slope of the lavel-1 model. Time, as well as the dummy variables of experimental conditions *RC* and control blocks *GRAD*, had no statistically significant effect on the dependent variable, thus, and were removed from the model. In addition, there was a statistically significant random effect indicating substantial variation in individual attainment trajectories among participants within latent classes.

Analysis of Critiques attainment by self-evaluation on the cardinal scale produced the most parsimonious model with a significant quadratic trend and two latent growth classes: the *lower performance* class comprising around 46% of participants with average self-assessment near 67 points, and the *higher performance* class comprising 54% of participants with average self-assessment near 84 points (Table 18 and Figure 20). The experimental conditions dummy variable had no statistically significant effect on any parameters of the level-1 model and was removed. The control dummy variable *GRAD* had statistically significant effects on the intercept and both slopes. The spread between the higher performance and lower performance classes is larger for undergraduate than graduate students. Over time, this spread widens for both categories of participants. In addition, there was a statistically significant random effect indicating substantial variation in individual attainment trajectories among participants within latent classes.

Table 17. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Self-evaluation on Cardinal Scale

```
Level-1 Model
RatgStuAr_Self = Π0 + E
Level-2 Model
Π0 = B00 + B01*RatgStuAr_Self_Class + R0
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           63.397      1.243       50.997         93    0.000
    ...Class, B01           19.780      1.490       13.276         93    0.000

Random Effect               St Dev   Var Comp        DF        Chi-sq  P-value
INTRCPT1, R0                 4.950     24.504        93       200.632    0.000
  level-1, E                 9.925     98.502
```

Table 18. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Self-evaluation on Cardinal Scale

```
Level-1 Model
RatgStuCr_Self = Π0 + Π1*T + Π2*T2+ E
Level-2 Model
Π0 = B00 + B01*GRAD + B02*RatgStuCr_Self_Class + B03*(GRAD*...Class) + R0
Π1 = B10 + B11*GRAD + B12*RatgStuCr_Self_Class
Π2 = B20 + B21*GRAD
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           66.538      1.893       35.150         91    0.000
        GRAD, B01            5.985      2.482        2.411         91    0.018
    ...Class, B02           16.692      2.367        7.053         91    0.000
GRAD*  Class, B03           -4.307      2.422       -1.778         91    0.078
T slope, Π1
    INTRCPT2, B10            3.351      1.172        2.860        422    0.005
        GRAD, B11           -4.321      2.070       -2.088        422    0.037
   ... Class, B12            1.458      0.833        1.750        422    0.080
T2 slope, Π2
    INTRCPT2, B20           -0.908      0.259       -3.511        422    0.001
        GRAD, B21            0.992      0.426        2.327        422    0.020

Random Effect               St Dev   Var Comp           DF     Chi-sq  P-value
INTRCPT1, R0                 4.671     21.816           91    217.342    0.000
  level-1, E                 8.432     71.092
```

Peer Evaluation

Expectedly, the plot of average attainment by peer evaluation on the ordinal scale is not informative in terms of the level or the rate of change of attainment because the average of reversed ranking scores is always equal to the median of the scale, in this case, to 3 (Figure 21 and Figure 22). It is reported here for consistency and completeness.

Visual examination of average attainment by peer evaluation on the cardinal scale suggests several interesting observations (Figure 23 and Figure 24). Firstly, on average peer evaluations of Submission in the graduate course are higher than those in the undergraduate course, just as average attainment by self-evaluation are higher in the graduate course. Moreover, the difference in average peer evaluations of Submissions between graduate and undergraduate courses seem to be around 10 points, which is close to the difference in average self-evaluations between graduate and undergraduate courses. Secondly, received peer evaluations of Submissions, on average, occupy lower range of the cardinal scale than self-evaluations, again, by about 5-10 points. Finally, average attainment by peer evaluation of Submissions appears to be higher in the steady groups than in randomly assigned groups. In contrast, average attainment by peer evaluation of Critiques appears to be lower in the steady groups than in randomly assigned groups. Overall, average attainment by peer evaluation of both Submissions and Critiques indicated no strong upward or downward trend over time.

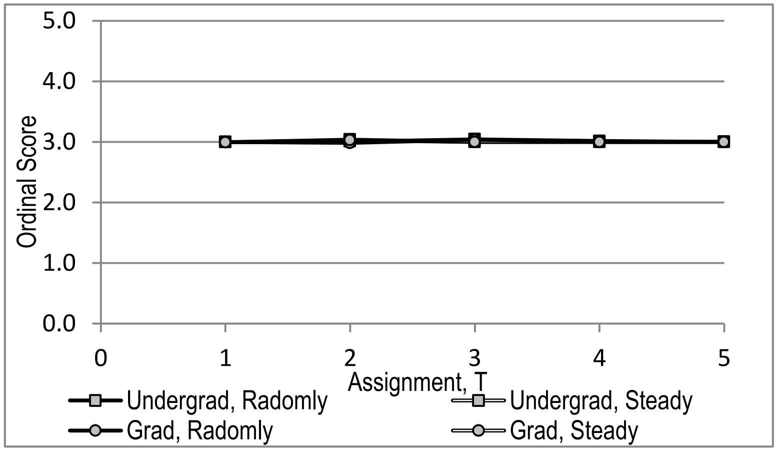


Figure 21. Average Submission Attainment by Peer Evaluation on Ordinal Scale

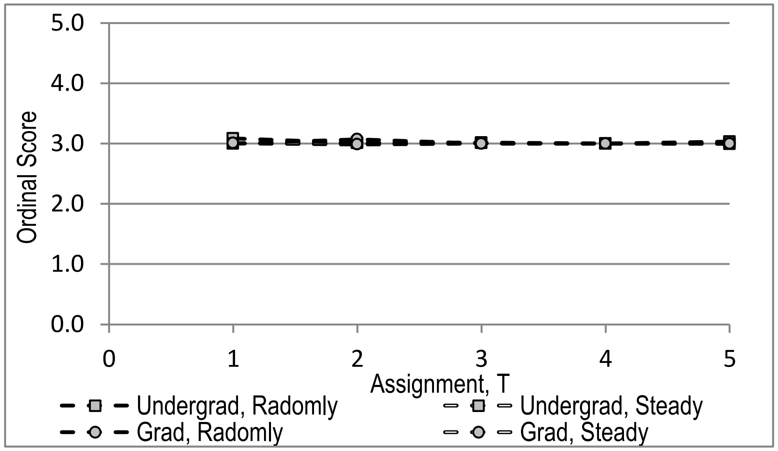


Figure 22. Average Critiques Attainment by Peer Evaluation on Ordinal Scale

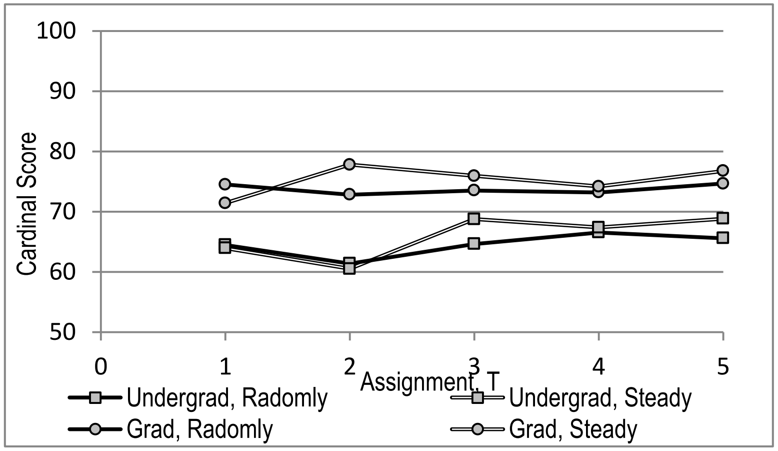


Figure 23. Average Submission Attainment by Peer Evaluation on Cardinal Scale

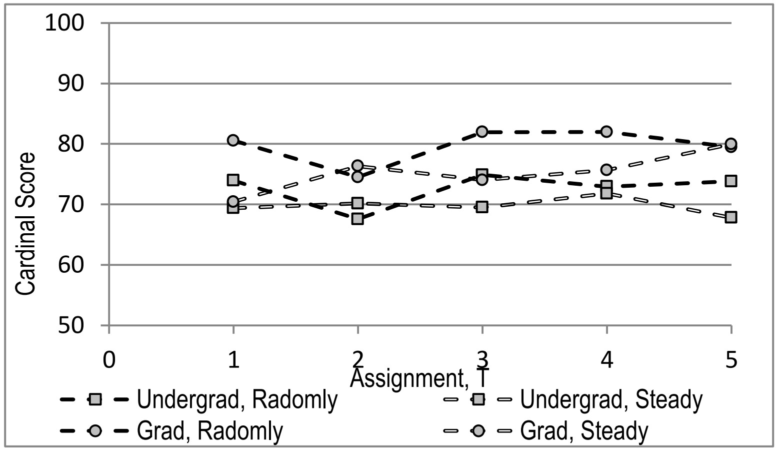


Figure 24. Average Critiques Attainment by Peer Evaluation on Cardinal Scale

Analysis of Submission attainment by peer evaluation on the ordinal scale showed that time had no statistically significant effect on the change in attainment, *t*(458) = -0.486, *p* = .627; the experimental conditions had no statistically significant effect on the initial level of miscalibration, *t*(458) = 1.174, *p* = .241; and the student level control variable had no statistically significant effect on the initial level of miscalibration as well, *t*(458) = –1.483, *p* = .139. These variables were respectively removed from the hierarchical model. The final model showed that Submission attainment by peer evaluation measured on the ordinal scale splits the population into two latent classes – the *lower performance* class, i.e., the participants, whose Submissions on average receive ranking-based score of 2.20, *t*(461) = 37.001, *p* = .000 (33% of the sample), and the *higher performance* class, i.e., the participant, whose Submissions on average receive ranking-based score of 3.36, *t*(461) = 15.316, *p* = .000 (67% of the sample) (Table 19).

Analysis of Critiques attainment by peer evaluation on the ordinal scale produced very similar results (Table 20), with the *lower performance* class accounting for 42% of the sample and the *higher performance* class – for the remaining 58%.

The cross-tabulation of the calibrating and the overconfident latent classes of submission and critiques miscalibration measured on the cardinal scale is presented in Table 21. This distribution shows that, judging by peer evaluation, among stronger Submission creators, there are as twice as many stronger reviewers than weaker reviewers. At the same time, among weaker creators, there are as twice as many weaker reviewers than stronger reviewers. Plots of Submission and Critiques attainment by peer evaluation on the ordinal scale are presented in Figure 25 and Figure 26 respectively.

Table 19. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Peer Evaluation on Ordinal Scale

```
Level-1 Model
ScrkStuAr_Attm = Π0 + E
Level-2 Model
Π0 = B00 + B01*(ScrkStuAr_Attm_Class)
Fixed Effect           Coefficient   St Error       T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            2.190      0.059        37.001        461    0.000
    ...Class, B01            1.168      0.076        15.316        461    0.000
```

Table 20. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Peer Evaluation on Ordinal Scale

```
Level-1 Model
ScrkStuCr_Attm = Π0 + E
Level-2 Model
Π0 = B00 + B01*(ScrkStuCr_Attm_Class)
Fixed Effect           Coefficient   St Error       T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            2.355      0.054        43.234        446    0.000
    ...Class, B01            1.088      0.075        14.461        446    0.000
```

Table 21. Distribution of Participants across Latent Classes of Submission and Critiques Attainment by Peer Evaluation (on Ordinal Scale)

Performance		Critiques		Total
		Lower	Higher	
Submission	Lower	20%	12%	33%
	Higher	21%	46%	67%
Total		42%	58%	100%

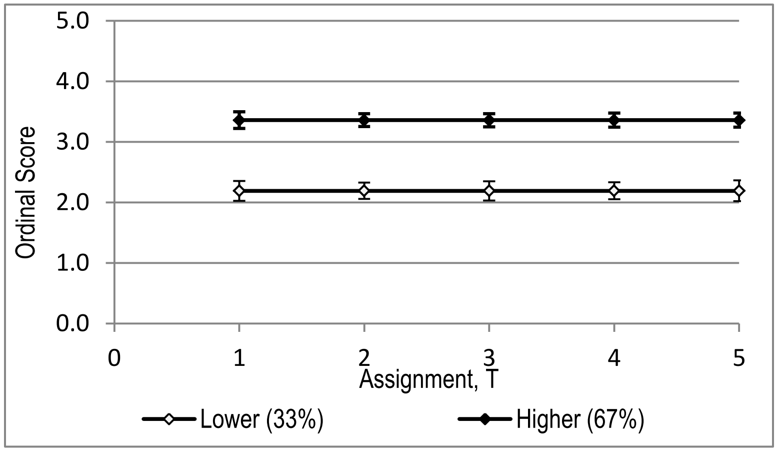


Figure 25. Latent Growth Classes of Submission Attainment by Peer Evaluation on Ordinal Scale (with Indicated Standard Errors)

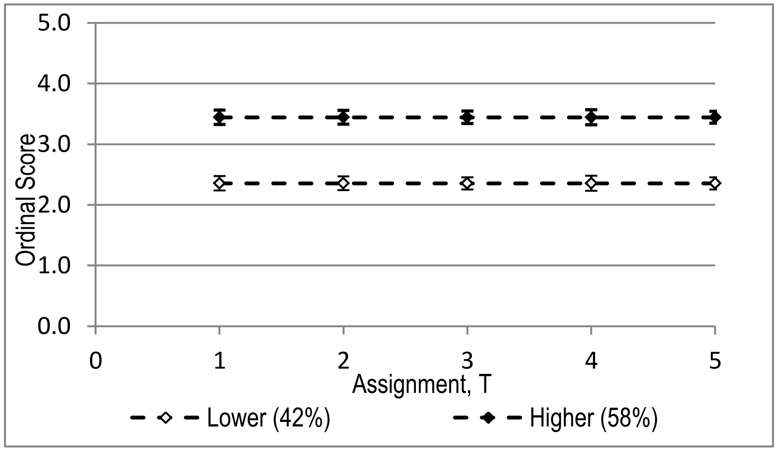


Figure 26. Latent Growth Classes of Critiques Attainment by Peer Evaluation on Ordinal Scale (with Indicated Standard Errors)

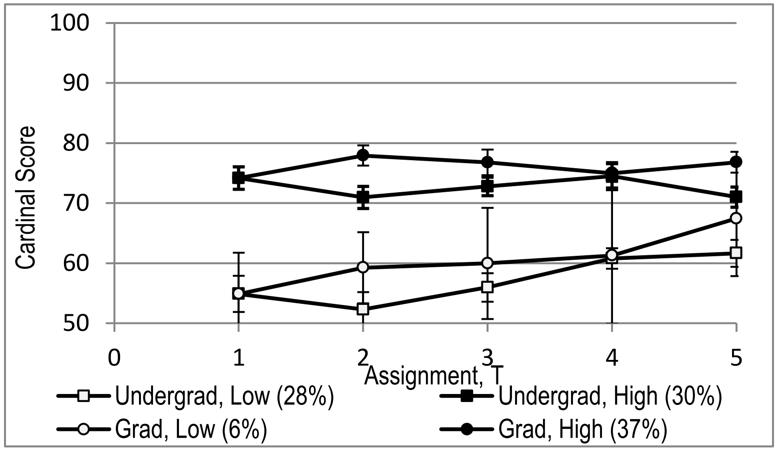


Figure 27. Latent Growth Classes of Submission Attainment by Peer Evaluation on Cardinal Scale (with Indicated Standard Errors)

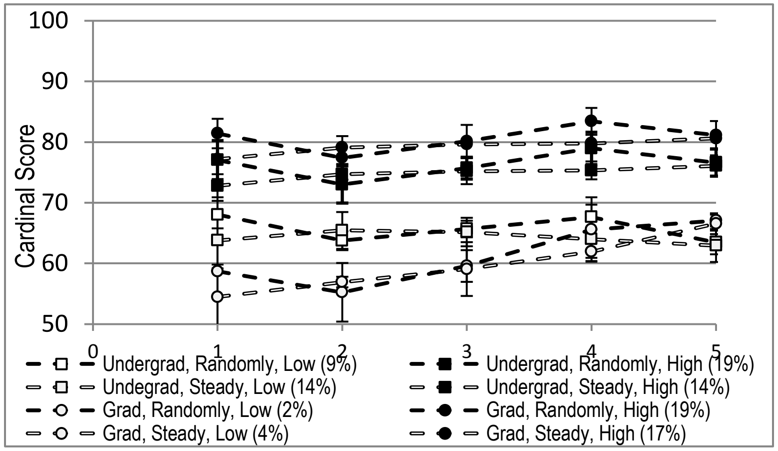


Figure 28. Latent Growth Classes of Critiques Attainment by Peer Evaluation on Cardinal Scale (with Indicated Standard Errors)

Analysis of Submission attainment by peer evaluation on the cardinal scale showed more complex dynamics. The best fitting model had two latent classes (the *lower performance* class and the *higher performance* class) and a significant cubic trend (Table 24). In addition, while the experimental condition dummy variable *RC* had no significant effect on the parameters of the level-1 model, and was removed from the final model, the student level variable *GRAD* had a significant effect on all slope coefficients but not on the intercept (both undergraduate and graduate students had on average the same scores in the first assignment).

Submission attainment scores appeared to fluctuate slightly above 75 points for the *higher performance* class of graduate students, and slightly below 75 points for the *higher performance* class of undergraduate students, with no definitive overall upward or downward trend. In *lower performance* class, both undergraduate and graduate students showed an upward trend, with scores of graduate students averaging slightly above undergraduate students' scores (Figure 27).

Interestingly, the cross-classification of participants across the lower and higher performance latent classes of submission attainment on ordinal and cardinal scales shows that about 19% of all participants are classified differently on different scales (Table 22).

Analysis of Critiques attainment by peer evaluation on the cardinal scale revealed even more complex dynamics. The best fitting model had two latent classes (the *lower performance* class and the *higher performance* class) and a significant cubic trend (Table 24). In addition, the experimental condition dummy variable *RC* had a significant effect on the intercept and all slopes of the level-1 model. The student level variable *GRAD* had a significant effect on the intercept, as well as on the slope of time squared; moreover, the interaction of *GRAD* with the latent class dummy variable had a significant effect on the intercept. This result lends no easy or intuitive interpretation (Figure 28). Noticeably, the spread between the *higher performance* class and the *lower performance class* trajectories is much larger among graduate than undergraduate students.

Interestingly, the cross-classification of participants across lower and higher performance latent classes of submission attainment on ordinal and cardinal scales shows that about 22% of all participants are classified differently depending on the scale (Table 23 and Figure 15).

Table 22. Cross-classification of Participants across Performance Latent Classes of Submission Attainment on Ordinal and Cardinal Scales

Performance		Cardinal		Total
		Lower	Higher	
Ordinal	Lower	23%	9%	33%
	Higher	10%	57%	67%
Total		34%	66%	100%

Table 23. Cross-classification of Participants across Performance Latent Classes of Critiques Attainment on Ordinal and Cardinal Scales

Performance		Cardinal		Total
		Lower	Higher	
Ordinal	Lower	24%	17%	42%
	Higher	5%	53%	58%
Total		30%	70%	100%

Table 24. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Peer Evaluation on Cardinal Scale

```
Level-1 Model
RatgStuAr_AvR = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*RatgStuAr_AvR_Class
Π1 = B10 + B11*GRAD
Π2 = B20 + B21*GRAD + B22*RatgStuAr_AvR_Class
Π3 = B30 + B31*GRAD
Fixed Effect           Coefficient   St Error       T-ratio          DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           54.879      1.965        27.931         454    0.000
    ...Class, B01           19.293      1.863        10.355         454    0.000
T slope, Π1
    INTRCPT2, B10           -7.444      3.239        -2.298         454    0.022
        GRAD, B11           15.050      4.848         3.105         454    0.002
T2 slope, Π2
    INTRCPT2, B20            5.702      2.092         2.726         454    0.007
        GRAD, B21           -9.645      3.515        -2.744         454    0.007
    ...Class, B22           -0.620      0.234        -2.652         454    0.009
T3 slope, Π3
    INTRCPT2, B30           -0.854      0.341        -2.509         454    0.013
        GRAD, B31            1.561      0.595         2.622         454    0.009
```

Table 25. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Peer Evaluation on Cardinal Scale

```
Level-1 Model
RatgStuCr_AvR = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*GRAD + B02*RC + B03*RatgStuCr_AvR_Class + B04*(GRAD*..._Class)
Π1 = B10 + B11*RC
Π2 = B20 + B21*GRAD + B22*RC + B23*RatgStuCr_AvR_Class + B24*(GRAD*..._Class)
Π3 = B30 + B31*RC
Fixed Effect           Coefficient   St Error       T-ratio          DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           68.030      1.602        42.472         434    0.000
        GRAD, B01           -9.329      1.809        -5.156         434    0.000
          RC, B02           -4.238      1.961        -2.161         434    0.031
    ...Class, B03            8.994      1.619         5.555         434    0.000
GRAD*..Class, B04           13.718      2.388         5.746         434    0.000
T slope, Π1
    INTRCPT2, B10           -9.444      3.680        -2.566         434    0.011
          RC, B11           12.390      4.660         2.659         434    0.009
T2 slope, Π2
    INTRCPT2, B20            6.194      2.398         2.583         434    0.010
        GRAD, B21            0.808      0.229         3.537         434    0.001
          RC, B22           -7.664      2.901        -2.642         434    0.009
    ...Class, B23            0.259      0.166         1.564         434    0.118
GRAD*..Class, B24           -0.801      0.279        -2.867         434    0.005
T3 slope, Π3
    INTRCPT2, B30           -1.030      0.389        -2.644         434    0.009
          RC, B31            1.200      0.469         2.560         434    0.011
```

Expert Evaluation

Similarly to peer evaluation, the plot of average attainment by expert evaluation on the ordinal scale is not informative in terms of the level or the rate of change of attainment because the average of reversed ranking scores is always equal to the median of the scale, in this case, to 3 (Figure 29 and Figure 30). It is reported here for consistency and completeness.

Visual examination of average attainment by expert evaluation on the cardinal scale suggests several interesting observations (Figure 31 and Figure 32). Firstly, on average, peer evaluations of Submission in the graduate course are slightly higher than those in the undergraduate course. Secondly, expert evaluations of Submissions and Critiques, on average, occupy approximately the same range of the cardinal scale as self-evaluations and higher than peer evaluations. Further, average attainment by expert evaluation of Submissions appears to be higher in the steady groups than in randomly assigned groups; and this tendency is especially more pronounced in the graduate course. In contrast, average attainment by expert evaluation of Critiques in graduate course appears to be lower in the steady groups than in randomly assigned groups, although practically these two profiles overlap. In the undergraduate course, average attainment by expert evaluation of Critiques in steady and randomly assigned groups show no clear dominance. Overall, average attainment by expert evaluation of both Submissions and Critiques indicates slight downward trend over time, in contrast with self- and peer evaluation.

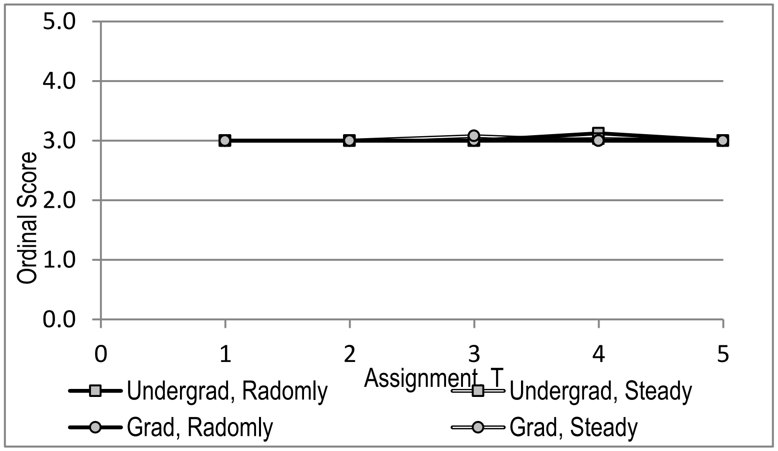


Figure 29. Average Submission Attainment by Expert Evaluation on Ordinal Scale

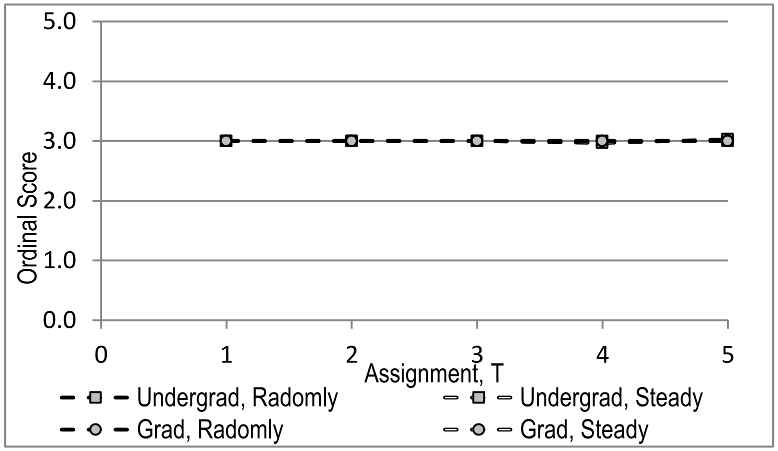


Figure 30. Average Critiques Attainment by Expert Evaluation on Ordinal Scale

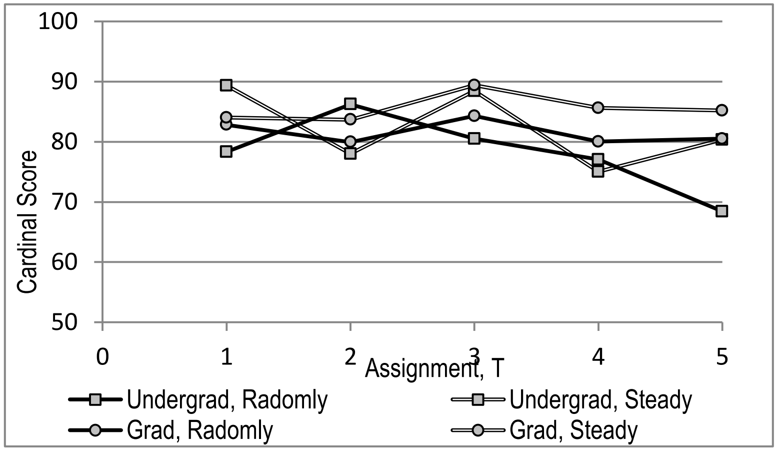


Figure 31. Average Submission Attainment by Expert Evaluation on Cardinal Scale

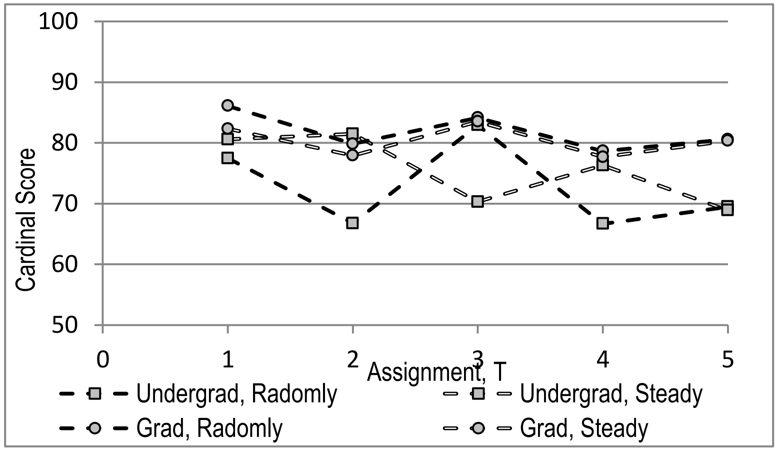


Figure 32. Average Critiques Attainment by Expert Evaluation on Cardinal Scale

Analysis of Submission attainment by expert evaluation on the ordinal scale showed that time, the experimental condition variable *RC* and the control variable *GRAD* had no statistically significant effects on the change in attainment. These variables were respectively removed from the hierarchical model. The final model showed that Submission attainment by expert evaluation on the ordinal scale splits the population into two latent classes – the *lower performance* class (with the average received ranking-based score of 2.5, 64% of the sample, and the *higher performance* class (with the average received ranking-based score of 3.8, 36% of the sample) (Table 26 and Figure 33). The cross-tabulation of the latent classes of Submission attainment by peer and expert evaluation showed that while none of the Submissions that were classified by peer evaluation as *lower* were classified by combined expert evaluation as *higher*, almost half of the Submissions that were classified by peer evaluation as *higher* were classified by expert evaluation as *lower* (Table 28). This result, however, should be interpreted with caution, given the issue with the expert evaluation reliability described in chapter III.

Analysis of Critiques attainment by expert evaluation on the ordinal scale produces result very similar to those of Submission: time, the experimental condition variable *RC* and the control variable *GRAD* had no statistically significant effects on the change in attainment. These variables were respectively removed from the hierarchical model. The final model showed that Critiques attainment by expert evaluation on the ordinal scale splits the population into two latent classes – the *lower performance* class (with the average received ranking-based score of 2.1, 54% of the sample, and the *higher performance* class (with the average received ranking-based score of 3.9, 46% of the sample) (Table 27 and Figure 34). The cross-tabulation of the latent classes of Critiques attainment by peer and expert evaluation shows that 20% of all Critiques were classified differently by peer evaluation and expert evaluation (Table 29).

Analysis of Submissions attainment by expert evaluation on the cardinal scale produces the best fitting model with two latent classes: the *lower performance* class comprising around 18% of participants, and the *higher performance* class comprising 82% of participants (Table 30 and Figure 35). The latent class variable had a significant effect on the intercept but not on the slope of the level-1 model; the spread of the trajectories of the latent classes was on average about 18 points. Time had a statistically significant negative effect on the dependent variable, i.e., on average, expert evaluation declined by 1.5 points from one assignment to the next. This is in sharp contrast with the behavior of self- and peer evaluations on cardinal scale showing overall steady, time-indifferent, pattern. The experimental conditions dummy variable *RC* had a statistically significant effect on the level-1 model's intercept, with attainment in steady groups being 3 points higher on average than in randomly assigned groups. The control dummy variable *GRAD* had no statistically significant effect on either intercept or slope and was removed from the final model.

Analysis of Critiques attainment by expert evaluation on the cardinal scale produces the best fitting model with no distinct latent growth classes, despite the initial indication of two latent classes by the LGM test. This is a somewhat puzzling result. Time had a statistically significant negative effect on the dependent variable – on average, expert evaluation declined by 1.7 points from one assignment to the next.

Table 26. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Expert Evaluation on Ordinal Scale

```
Level-1 Model
ScrkInsAr_Exp = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*ScrkInsAr_Exp_Class
Fixed Effect            Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00             2.512      0.059       42.491        461    0.000
    ...Class, B01             1.329      0.089       14.930        461    0.000
```

Table 27. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Expert Evaluation on Ordinal Scale

```
Level-1 Model
ScrkInsCr_Exp = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*ScrkInsCr_Exp_Class
Fixed Effect            Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00             2.146      0.074       29.173        446    0.000
    ...Class, B01             1.724      0.109       15.842        446    0.000
```

Table 28. Distribution of Participants across Latent Classes of Submission Attainment by Peer and Expert Evaluation (on Ordinal Scale)

Performance		Expert evaluation		Total
		Lower	Higher	
Peer evaluation	Lower	33%	0%	33%
	Higher	32%	36%	67%
Total		64%	36%	100%

Table 29. Distribution of Participants across Latent Classes of Critiques Attainment by Peer and Expert Evaluation (on Ordinal Scale)

Performance		Expert evaluation		Total
		Lower	Higher	
Peer evaluation	Lower	38%	4%	42%
	Higher	16%	42%	58%
Total		54%	46%	100%

The experimental conditions dummy variable *RC* also had no statistically significant effect on the level-1 model's intercept or slope. The control dummy variable *GRAD* had a statistically significant effect on the intercept, indicating that Critiques attainment by combined expert evaluation was about seven points higher on average for graduate as opposed to undergraduate students.

Closer scrutiny of the expert evaluation on the cardinal scale revealed that one of the specific causes of insufficiently high inter-observer reliability of the instructors was the presence of several outliers in evaluations. This indicates that in some cases the instructor evaluators varied substantially in how rubrics can be applied to assign score to a participant's Submission or Critiques. In addition, the declining trend may be indicative of a number of biases and, possibly, autocorrelation effects in expert evaluations, for example, increasing expectations. Consequently, in the rest of this study, although the results related to expert evaluations are presented for completeness and consistency, they should be treated with caution.

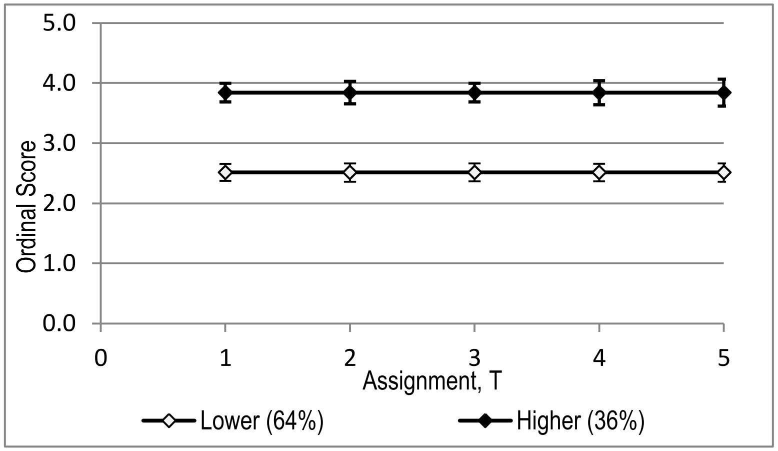


Figure 33. Latent Growth Classes of Submission Attainment by Expert Evaluation on Ordinal Scale (with Indicated Standard Errors)

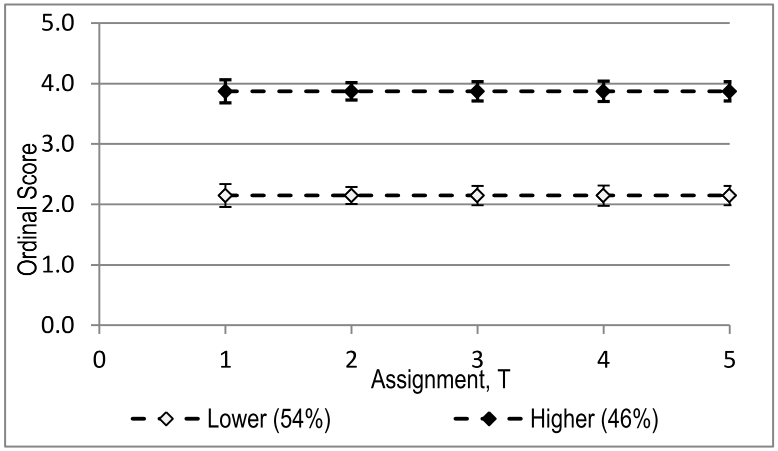


Figure 34. Latent Growth Classes of Critiques Attainment by Expert Evaluation on Ordinal Scale (with Indicated Standard Errors)

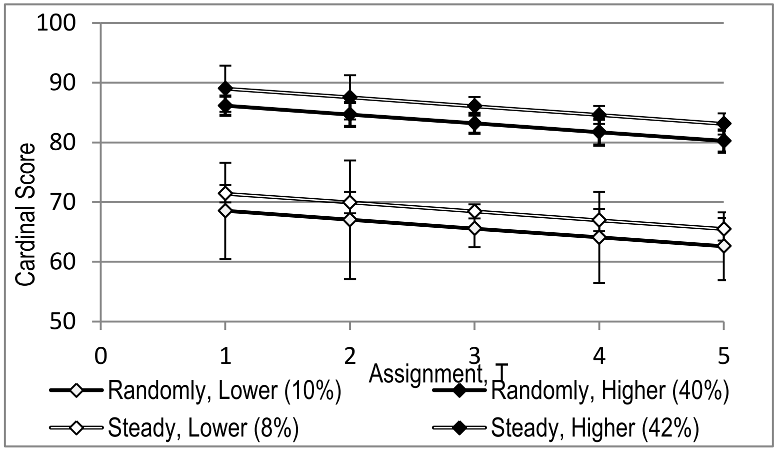


Figure 35. Latent Growth Classes of Submission Attainment by Expert Evaluation on Cardinal Scale (with Indicated Standard Errors)

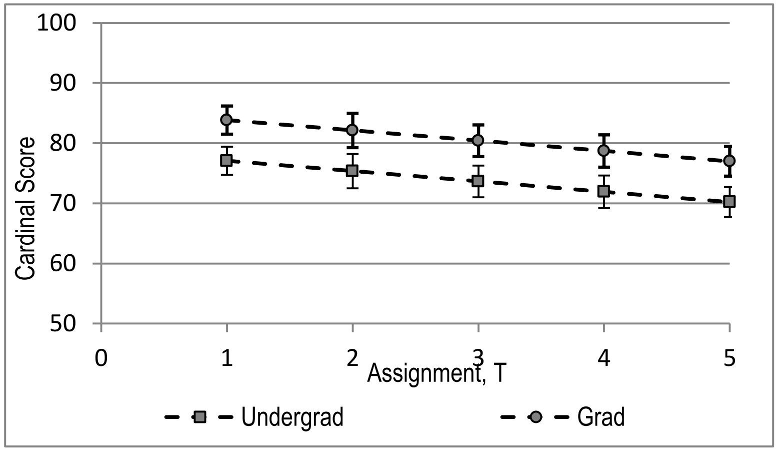


Figure 36. Latent Growth Classes of Critiques Attainment by Expert Evaluation on Cardinal Scale (with Indicated Standard Errors)

Table 30. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Attainment by Expert Evaluation on Cardinal Scale

```
Level-1 Model
RatgInsAr_Exp = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*RC + B02* RatgInsAr_Exp_Class
Π1 = B10
Fixed Effect           Coefficient   St Error      T-ratio        DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           68.531      1.808       37.914       459    0.000
          RC, B01            2.874      1.199        2.398       459    0.017
    ...Class, B02           17.608      1.593       11.057       459    0.000
T slope, Π1
    INTRCPT2, B10           -1.479      0.415       -3.566       459    0.001
```

Table 31. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Attainment by Expert Evaluation on Cardinal Scale

```
Level-1 Model
RatgInsCr_Exp = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*GRAD
Π1 = B10
Fixed Effect           Coefficient   St Error      T-ratio        DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           77.065      1.835       41.998        94    0.000
        GRAD, B01            6.768      2.755        2.457        94    0.016
T slope, Π1
    INTRCPT2, B10           -1.709      0.496       -3.448       445    0.001

Random Effect               St Dev   Var Comp           DF    Chi-sq  P-value
INTRCPT1, R0                12.067    145.607           94   387.136    0.000
  level-1, E                14.684    215.607
```

Dynamics of Miscalibration

In this subsection, dynamics of miscalibration of the original KA submissions and their critiques evaluated by their authors, peers and experts are considered. The aim is to examine longitudinal patterns of changes in miscalibration of Submissions and Critiques measured on the ordinal and cardinal scales.

Miscalibration with Respect to Peer Evaluation

Visual examination of plots of miscalibration of Submissions and Critiques with respect to peer evaluation showed non-trivial patterns for both scales (Figure 37 and Figure 38). On average overconfidence appeared to be prevailing form of miscalibration, although the magnitude of overconfidence was larger for Submissions than for Critiques on both scales. This implies a preliminary conclusion that self-evaluation tends to be more similar to peer evaluations for Critiques than for Submissions, likely because the task of critiquing seem to the participants less complex and multifaceted than the task of creating the original Submission. In other words, certain non-conflicting characteristics of Critiques' goodness may be more apparent to or preferred by participants than those of Submissions, despite the fact that detailed rubrics for evaluating Submissions were provided. Overall, the ordinal and cardinal scales show different dynamics of average miscalibration that are difficult to interpret.

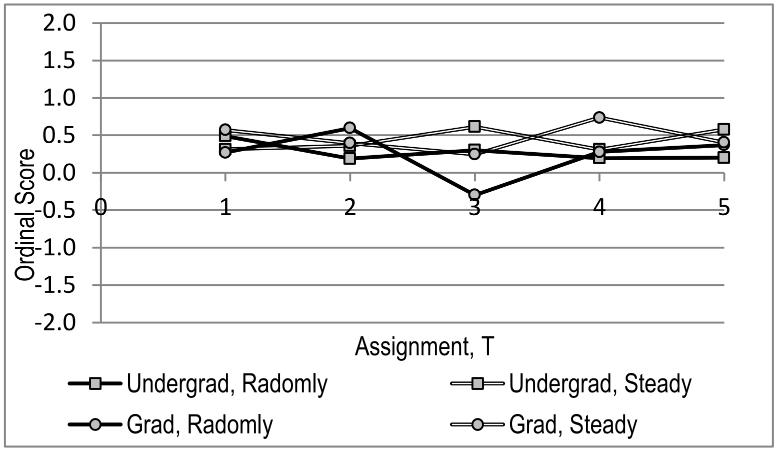


Figure 37. Average Submission Miscalibration with Respect to Peer Evaluation on Ordinal Scale

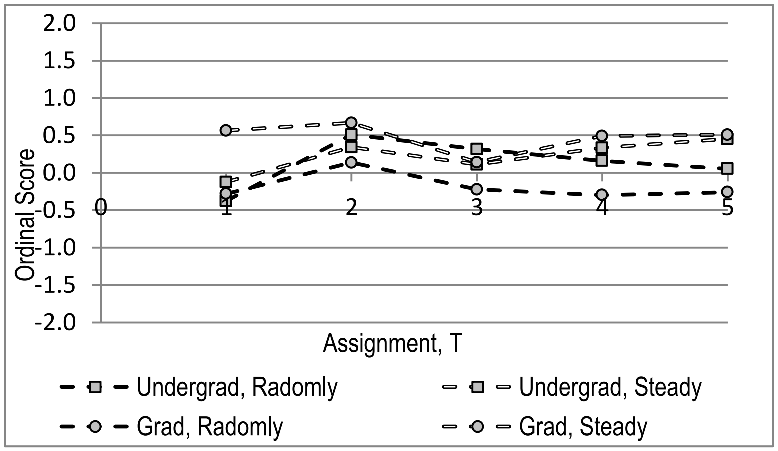


Figure 38. Average Critiques Miscalibration with Respect to Peer Evaluation on Ordinal Scale

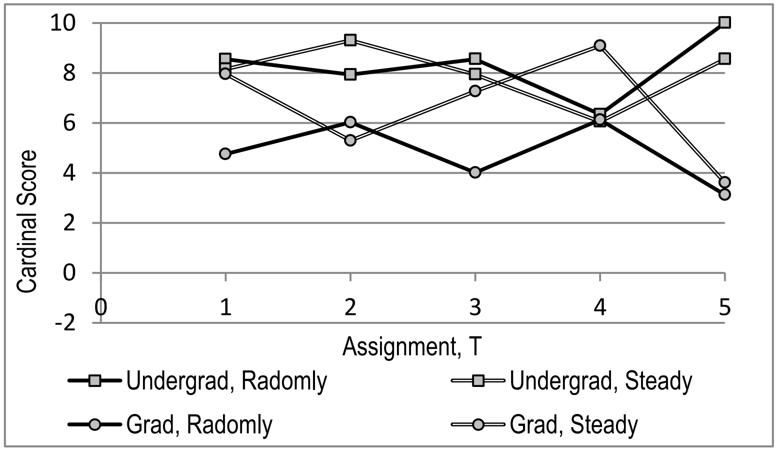


Figure 39. Average Submission Miscalibration with Respect to Peer Evaluation on Cardinal Scale

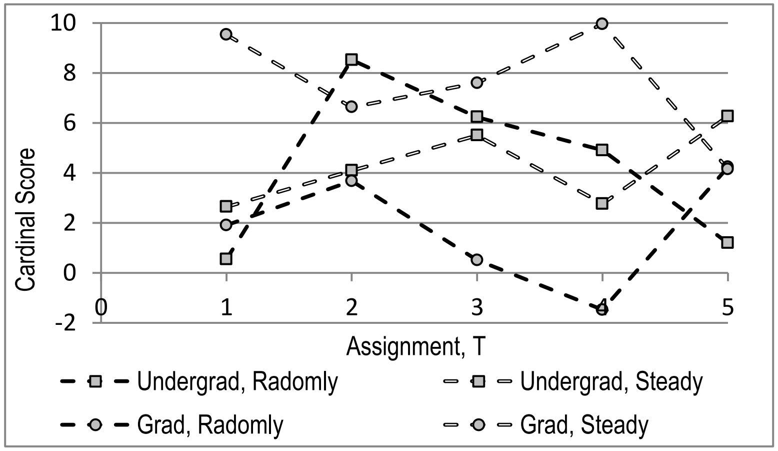


Figure 40. Average Critiques Miscalibration with Respect to Peer Evaluation on Cardinal Scale

Analysis of Submission miscalibration with respect to peer evaluation on the ordinal scale produced the best fitting model with two latent growth classes – the *underconfident* class, i.e., the participants, who on average rank their Submissions below the average score produced by peer ranking of their Submissions (32% of the sample) and the *overconfident* class, i.e., the participant, who on average rank their Submissions above the average score produced by peer ranking (68% of the sample) (Table 32 and Figure 41). Time had a statistically significant effect on miscalibration, $t(438) = –1.889$, $p = .059$; and the rate of change (slope) was significantly affected by the latent growth class, $t(438) = 2.507$, $p = .013$ (confidence of the underconfident declined faster than confidence of the overconfident increased). While the trajectory of the underconfident class is just below calibration (the zero line); the trajectory of the overconfident class begins 0.8 points higher, $t(438) = 5.176$, $p = .000$. While the control dummy variable *GRAD* had not statistically significant effect, $t(438) = –1.626$, $p = .104$, its interaction with the latent class dummy variable did have a statistically significant effect on the intercept of the level-1 model – graduate students in the overconfident class ranked themselves higher by additional 0.5 points, $t(438) = 2.284$, $p = .023$. The experimental conditions dummy variable *RC* had no statistically significant effect on either slope or intercept and was removed from the model. These results regarding the behavior of Submission miscalibration with respect to peer evaluation on the ordinal scale suggest divergence in the evaluation intersubjectivity: irrespective of the experimental conditions, on average, the overconfident were becoming more overconfident while the underconfident were becoming more underconfident.

The cross-tabulation of the latent classes of Submission miscalibration and Submission attainment by peer evaluation on the ordinal scale indicated that, after accounting for the disproportion between the higher and lower attainment classes, the fraction of the underconfident among the higher performers was much higher (40%) comparing to the fraction of the underconfident among lower performers (18%) (Table 34). At the same time, the fraction of the overconfident among the lower performers was much higher (81%) comparing to the fraction of the overconfident among higher performers (61%). This result supports the notion of the unskilled-and-unaware problem.

Analysis of Critiques miscalibration with respect to peer evaluation on the ordinal scale produced the best fitting model with two latent growth classes – the *underconfident* class, i.e., the participants, who on average rank their Critiques below the average score produced by peer ranking of their Submissions (42% of the sample) and the *overconfident* class, i.e., the participant, who on average rank their Submissions above the average score produced by peer ranking (64% of the sample) (Table 33 and Figure 42). Time had a statistically significant effect on miscalibration, *t*(426) = 2.269, *p* = .024; and the rate of change was significantly affected by the latent growth class, *t*(426) = –1.879, *p* = .060. Here, the trends opposite to those for Submissions can be observed: (confidence of the overconfident declined, and confidence of underconfident improved; the latter showed higher rate of change). While trajectory of the underconfident class was converging closer to the zero line, the trajectory of the overconfident remained above 0.5 points. The experimental conditions dummy variable *RC* had a statistically significant effect on the intercept, *t*(438) = 1.816, *p* = .070; participants in steady groups showed slightly higher confidence than those in the randomly allocated groups. The control dummy variable *GRAD* had not statistically significant effect on either intercept or slope. These results regarding the behavior of Critiques miscalibration with respect to peer evaluation on the ordinal scale suggests convergence in the evaluation intersubjectivity – irrespective of the experimental conditions, on average, the overconfident were becoming less overconfident while the underconfident were becoming more confident.

The cross-tabulation of the latent classes of Critiques miscalibration and Critiques attainment by peer evaluation on the ordinal scale indicated that, after accounting for the disproportion between the higher and lower attainment classes, the fraction of the underconfident among the higher performers was much higher (50%) comparing to the fraction of the underconfident among lower performers (36%) (Table 35). At the same time, the fraction of the overconfident among the lower performers was higher (64%) comparing to the fraction of the overconfident among higher performers (50%). This result is consistent with the finding in Submission evaluations and also supports the notion of the unskilled-and-unaware effect.

Table 32. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Miscalibration with Respect to Peer Evaluation on Ordinal Scale

Level-1 Model
ScrkStuAr_Misc = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*GRAD + B02* ScrkStuAr_Misc_Class + B03*(GRAD*...Class)
Π1 = B10 + B11*ScrkStuAr_Misc_Class

Fixed Effect	Coefficient	St Error	T-ratio	DF	P-value
INTRCPT1, Π0					
INTRCPT2, B00	-0.185	0.115	-1.611	438	0.108
GRAD, B01	-0.272	0.167	-1.626	438	0.104
...Class, B02	0.828	0.160	5.176	438	0.000
GRAD*..Class, B03	0.486	0.213	2.284	438	0.023
T slope, Π1					
INTRCPT2, B10	-0.091	0.048	-1.889	438	0.059
...Class, B11	0.164	0.065	2.507	438	0.013

Table 33. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Miscalibration with Respect to Peer Evaluation on Ordinal Scale

Level-1 Model
ScrkStuCr_Misc = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*RC + B02*ScrkStuCr_Misc_Class
Π1 = B10 + B11*ScrkStuCr_Misc_Class

Fixed Effect	Coefficient	St Error	T-ratio	DF	P-value
INTRCPT1, Π0					
INTRCPT2, B00	-0.874	0.162	-5.407	426	0.000
RC, B01	0.229	0.126	1.816	426	0.070
...Class, B02	1.830	0.212	8.636	426	0.000
T slope, Π1					
INTRCPT2, B10	0.128	0.056	2.269	426	0.024
...Class, B11	-0.177	0.094	-1.879	426	0.060

Table 34. Distribution of Participants across Latent Classes of Submission Miscalibration and Submission Attainment by Peer Evaluation (on Ordinal Scale)

		Submission attainment		Total
		Lower	Higher	
Submission miscalibration	Underconfident	6%	27%	33%
	Overconfident	27%	41%	67%
Total		33%	67%	100%

Table 35. Distribution of Participants across Latent Classes of Critiques Miscalibration and Critiques Attainment by Peer Evaluation (on Ordinal Scale)

		Critiques attainment		Total
		Lower	Higher	
Critiques miscalibration	Underconfident	15%	29%	44%
	Overconfident	27%	30%	56%
Total		42%	58%	100%

Table 36. Distribution of Participants across Underconfident and Overconfident Latent Classes of Submission and Critiques Miscalibration (on Ordinal Scale)

		Critiques miscalibration		Total
		Underconfident	Overconfident	
Submission miscalibration	Underconfident	23%	9%	33%
	Overconfident	20%	47%	67%
Total		44%	56%	100%

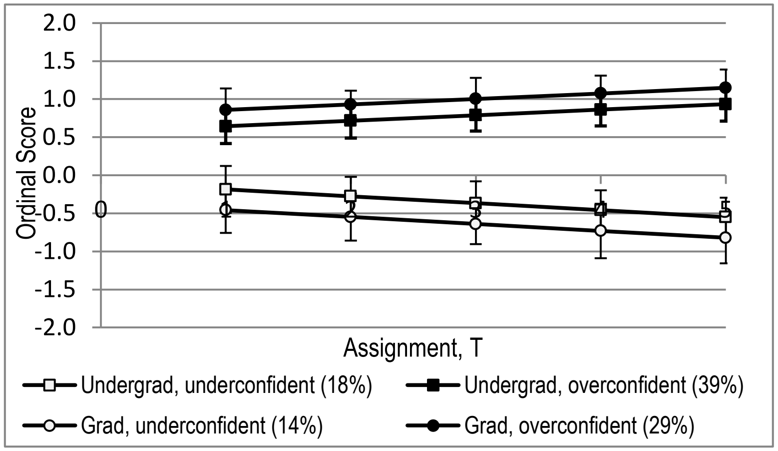


Figure 41. Latent Growth Classes of Submission Miscalibration with Respect to Peer Evaluation on Ordinal Scale (with Indicated St. Errors)

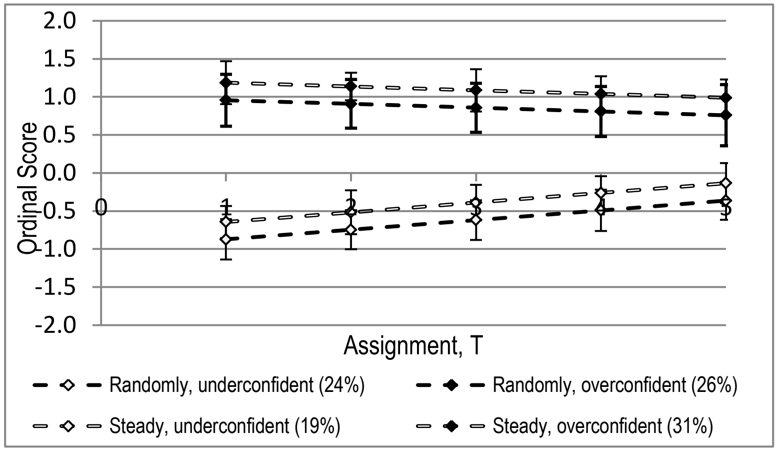


Figure 42. Latent Growth Classes of Critiques Miscalibration with Respect to Peer Evaluation on Ordinal Scale (with Indicated St. Errors)

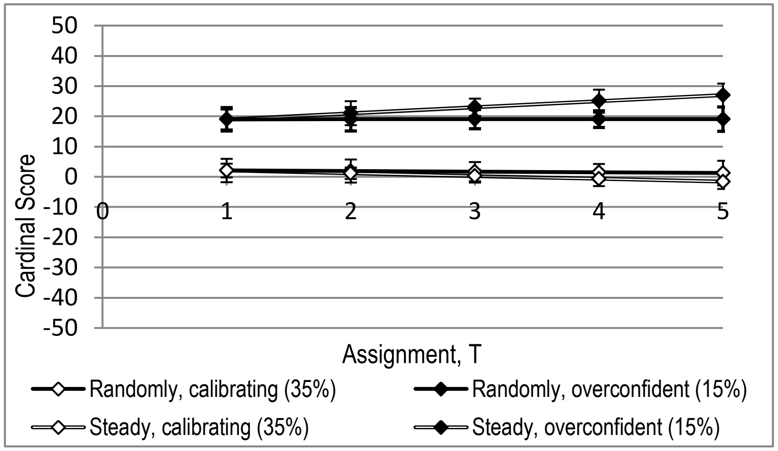


Figure 43. Latent Growth Classes of Submission Miscalibration with Respect to Peer Evaluation on Cardinal Scale (with Indicated St. Errors)

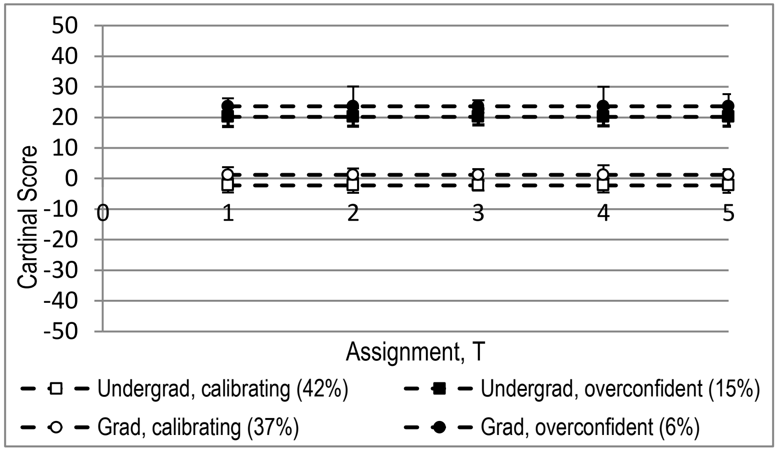


Figure 44. Latent Growth Classes of Critiques Miscalibration with Respect to Peer Evaluation on Cardinal Scale (with Indicated St. Errors)

Analysis of Submission miscalibration with respect to peer evaluation on the cardinal scale produced the best fitting model with two latent growth classes – the *calibrating* class, i.e., the participants, who on average rate their Submissions close to the average peer rating of their Submissions (70% of the sample, approximately equally split between undergraduate and graduate students) and the *overconfident* class, i.e., the participant, who on average rate their Submissions about 17 points higher than the average peer rating of their Submissions (30% of the sample, with approximately twice as many undergraduates than graduate students) (Table 37 and Figure 43). However, there was a statistically significant random effect indicating substantial variation in individual miscalibration trajectories among students within latent growth classes. This means that even within the calibrating class, there were participants that consistently showed some overconfidence or underconfidence over time. Time had no statistically significant effect on miscalibration. However, there was some evidence that the interaction of the latent class and experimental conditions dummy variables affected the rate of change (time slope), $t(438) = 2.044$, $p = .041$. This result suggested that overconfident participants in the steady peer groups may show a positive increase in their overconfidence. The experimental conditions dummy variable *RC* and the control dummy variable *GRAD* had no statistically significant effect on the intercept.

Analysis of Critiques miscalibration with respect to peer evaluation on the cardinal scale showed that after controlling for the student participant level time had no statistically significant effect on the change in miscalibration, $t(424) = 0.018$, $p = .985$; the experimental conditions had no statistically significant effect on the initial state of miscalibration, *t*(91) = 0.265, *p* = .792 and on the change in miscalibration over time, *t*(424) = 0.283, *p* = .778). The control dummy variable *GRAD* no significant effect on the rate of change. The final model showed that in terms of Critiques miscalibration on the cardinal scale, the population is also split into two latent classes – the *calibrating* class, i.e., participant, who on average rate their Critiques close to the average peer rating of their Critiques, with the slight, 2.3 points, average underconfidence (79% of the sample, approximately equally split between undergraduate and graduate students) and the *overconfident* class, i.e., participant, who on average rate their Submissions 22 points higher than the average peer rating of their Submissions (21% of the sample, with approximately 2.5 times as many undergraduates than graduate students) (Table 38). There is also a significant random effect indicating substantial variation in individual miscalibration trajectories among students within latent classes, similar to the effect observed in Submission miscalibration.

The cross-tabulation of the calibrating and the overconfident latent classes of submission and critiques miscalibration measured on the cardinal scale is presented in Table 39. This distribution, showing that the majority of the participants accurately calibrate their work and, moreover, calibrate accurately both Submissions and Critiques (when measured on the cardinal scale), is in sharp contrast with the distribution of miscalibration on the ordinal scale (Table 34 and Table 35), where the majority of participants show overconfidence in self-evaluating either Submission or Critiques or both. This observation indicates that the interpreting miscalibration is largely dictated by what scales is used to measure it.

Table 37. HLM Least-squares Estimates of Fixed and Random Effects (with Robust Standard Errors) of Submission Miscalibration with Respect to Peer Evaluation on Cardinal Scale

```
Level-1 Model
RatgStuAr_Misc = Π0 + Π1*T + E
Level-2 Model
Π0 = B00 + B01*RatgStuAr_Misc_Class + R0
Π1 = B10 + B11*RC + B12*RatgStuAr_Misc_Class + B13*(RC*...Class)
Fixed Effect           Coefficient   St Error      T-ratio         DF   P-value
INTRCPT1, Π0
    INTRCPT2, B00            2.085      1.274        1.636         93     0.105
    ...Class, B01           16.929      2.383        7.103         93     0.000
T slope, Π1
    INTRCPT2, B10           -0.206      0.588       -0.350        438     0.726
          RC, B11           -0.712      0.695       -1.024        438     0.307
    ...Class, B12            0.221      1.207        0.183        438     0.855
 RC*...Class, B13            2.699      1.320        2.044        438     0.041

Random Effect               St Dev   Var Comp           DF     Chi-sq   P-value
INTRCPT1, R0                 5.457     29.781           93    193.411     0.000
  level-1, E                11.208    125.615
```

Table 38. HLM Least-squares Estimates of Fixed and Random Effects (with Robust Standard Errors) for Critiques Miscalibration with Respect to Peer Evaluation on Cardinal Scale

```
Level-1 Model
RatgStuCr_Misc = Π0 + E
Level-2 Model
Π0 = B00 + B01*GRAD + B02* RatgStuCr_Misc_Class + R0
Fixed Effect           Coefficient   St Error      T-ratio         DF   P-value
INTRCPT1, Π0
    INTRCPT2, B00           -2.251      1.307       -1.722         92     0.088
        GRAD, B01            3.415      1.647        2.073         92     0.041
    ...Class, B02           22.424      1.769       12.678         92     0.000

Random Effect               St Dev   Var Comp           DF     Chi-sq   P-value
INTRCPT1, R0                 5.605     31.421           92    188.199     0.000
 level-1, E                 11.637    135.411
```

Table 39. Distribution of Participants across Calibrating and Overconfident Latent Classes of Submission and Critiques Miscalibration (on Cardinal Scale)

Miscalibration		Critiques		Total
		Calibrating	Overconfident	
Submission	Calibrating	60%	9%	69%
	Overconfident	18%	12%	31%
Total		79%	21%	100%

Miscalibration with Respect to Expert Evaluation

The profiles of the average miscalibration with respect to peer evaluation and with respect to expert evaluation on the ordinal scale closely mimic each other (Figure 45 and Figure 46). Also note the high Pearson correlation between these two miscalibration variables of 0.64 for Submissions and of 0.75 for Critiques (Table 11). The similarity of peer and expert evaluations of Submissions and Critiques in the ordinal scale is also supported by correlations between attainment by peer and expert evaluations (0.54 for Submissions and 0.60 for Critiques; Table 11).

The profiles of average miscalibration with respect to peer evaluation and with respect to expert evaluation on the cardinal scale appear dissimilar (Figure 47 and Figure 48). While average miscalibration with respect to peer evaluation shows prevailing overconfidence, average miscalibration with respect to expert evaluation shows a wider variation. In addition, while average miscalibration with respect to peer evaluation shows greater magnitude for Submission, average miscalibration with respect to expert evaluation shows greater magnitude for Critiques. This is indicative of the differences by participants and experts in subjective importance placed on evaluations of Submissions and Critiques. In other words, experts seemed to have placed more weight on evaluating goodness of Critiques than student participants, despite the fact that equal weights were explicitly places on evaluations of both Submissions and Critiques and communicated to participants.

In the graduate course, average cardinal-scale miscalibration with respect to expert evaluation fluctuated around zero, suggesting that on average self-evaluations of Submissions and Critiques were close to expert evaluations. However, low Pearson correlations between cardinal-scale self- and expert evaluations of Submissions and Critiques of 0.26 and 0.13 respectively suggest the lack of strong linear relationships between self- and expert evaluations. This puzzling result can be possibly explained by the presence of the outliers in expert evaluation and motivates further investigation.

HLM analysis of Submission and Critiques miscalibration with respect to expert evaluation on the ordinal and the cardinal scales was also conducted. Due to the issues with expert inter-observer reliability discussed above, the results cannot be considered trustworthy and, therefore, are not reported here.

Dynamics of Controversy

Visual examination of plots of controversy of Submissions and Critiques computed as DFM (Figure 49 and Figure 50) and as DFC (Figure 51 and Figure 52) showed non-trivial patterns for both approaches. On average, controversy dynamics in randomly recombined and steady groups did not appear to be strikingly different. The only observable visual hint is a more stable increase in Submission controversy in steady peer groups. This instigated the need for further analysis.

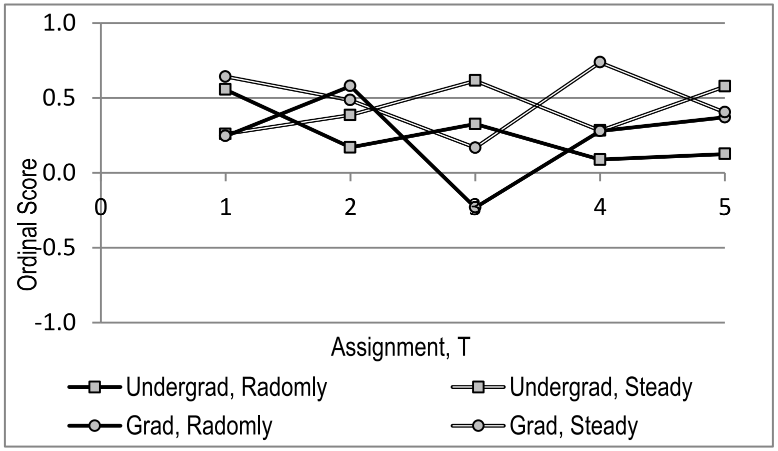


Figure 45. Average Submission Miscalibration with Respect to Expert Evaluation on Ordinal Scale

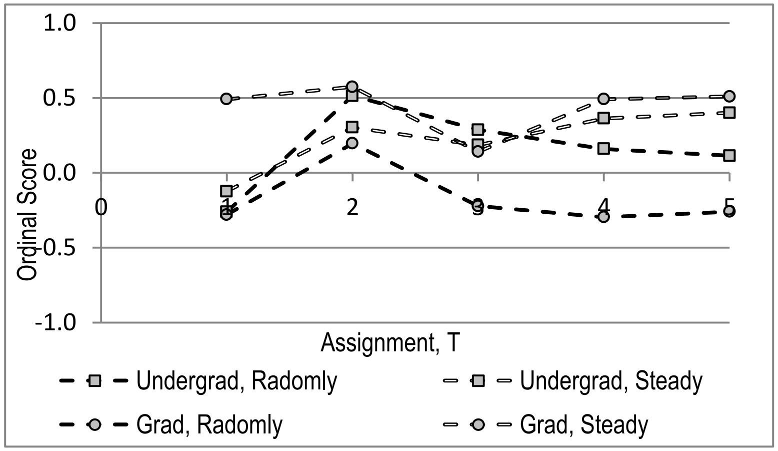


Figure 46. Average Critiques Miscalibration with Respect to Expert Evaluation on Ordinal Scale

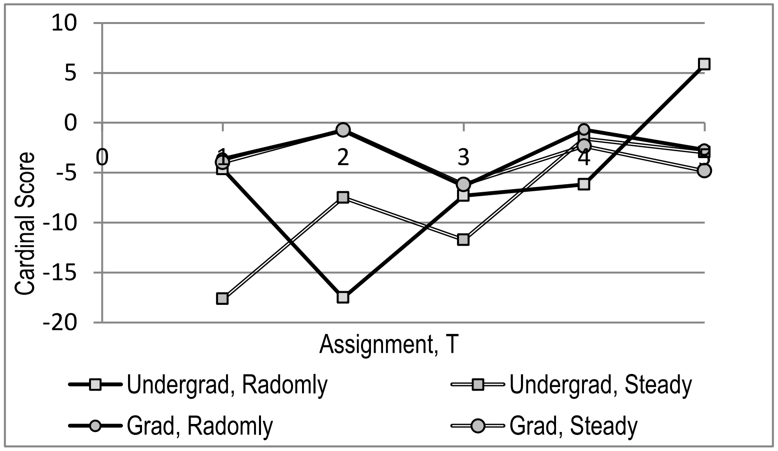


Figure 47. Average Submission Miscalibration with Respect to Expert Evaluation on Cardinal Scale

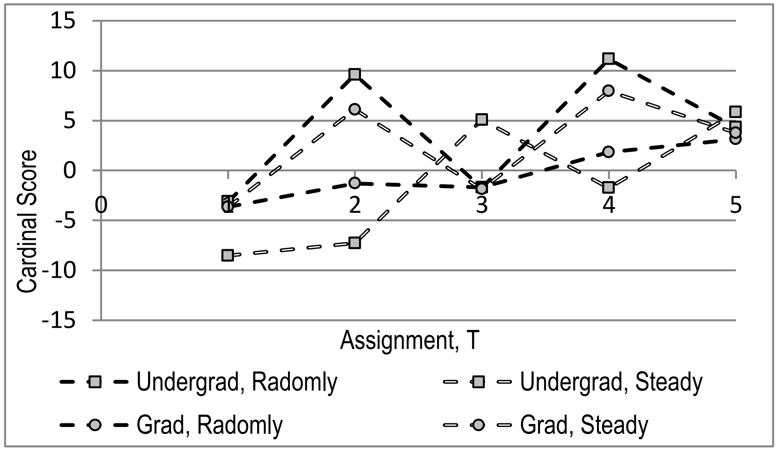


Figure 48. Average Critiques Miscalibration with Respect to Expert Evaluation on Cardinal Scale

The results of LGM analysis of controversy, presented in Table 14, suggested no evidence that latent classes of actors with characteristic controversy trajectories exist. This result held for both Submissions and Critiques, as well as for controversy computed as DFM and DFC based on the ordinal scale data (analysis of controversy computed based on the cardinal scale is outside the scope of this dissertation.) Further analysis revealed that, overall, Submission controversy (both DFM and DFC) followed statistically significant cubic trend (Table 40, Figure 53; Table 41, Figure 55). At the same time, Critique controversy (both DFM and DFC) followed statistically significant quadratic non-monotonic trend. The experimental conditions variable *RC* had a statistically significant effect (at 5% significance level) on the rate of change of controversy over time, with the exception of the case of Critiques controversy DFC where the effect of *RC* was significant only at the 10% level. The control variable *GRAD* had a statistically significant effect on Submission controversy; for Critiques controversy DFM, *GRAD* had a statistically significant effect only in interaction with *RC*; no statistically significant effect of *RC* on Critiques controversy DFC was found. After controlling for the student level *GRAD*, steady peer groups demonstrated more stable increasing dynamics of controversy than randomly assigned groups. Specifically, Submissions controversy DFM in steady groups gradually increased from below 0.5 to near 0.6 for both undergraduate and graduate courses (Table 40 and Figure 53). Similar dynamics in steady groups was observed for controversy computed as DFC (Table 41 and Figure 55). At the same time, Submissions controversy in randomly recombined groups was more unstable, showing a somewhat increasing cubic trend for undergraduate students, and decreasing for graduate students. Interestingly, in the first assignment, the overall level of Submission controversy was higher in the graduate course than in the undergraduate, and remained practically stable at 0.5, whereas Submission controversy in the undergraduate course increased from just below 0.4 to over 0.6. Thus, overall, there was mixed evidence of the controversy decrease over multiple iterations. However, controversy tended to remain higher in the steady groups than in randomly recombined groups, especially on Critiques. This was an important finding because it suggested that to adverse intersubjectivity effects, such as the ranking confusion and the expectations perplexity effects, were dominated by the favorable effects, such as social norming and the cross-pollination effects, in the randomly recombined peer groups.

A very important caveat in this analysis is that the differences in controversy as DFM or DFC were larger than the longitudinal changes. This has serious implications for the design choice of the computation approach to controversy in evaluation systems.

When miscalibration with respect to peer evaluations of respective KAs was included in the model, it did not pass the significance test; that is, the dynamics of miscalibration did not appear to have strong association with the dynamics of controversy. This finding suggests that creation competencies related to clarity and audience awareness are not strongly linked to creator's propensity to misjudge his own KA. The evidence of association between miscalibration and bias is discussed in the next subsection.

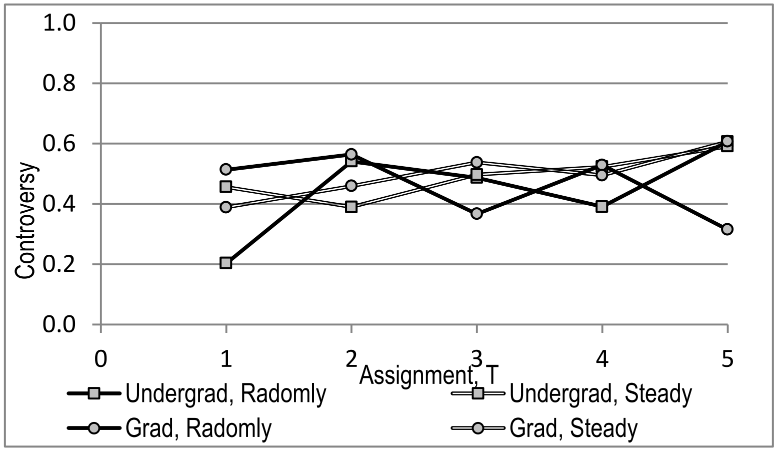


Figure 49. Average Submission Controversy as Deviation from Mean

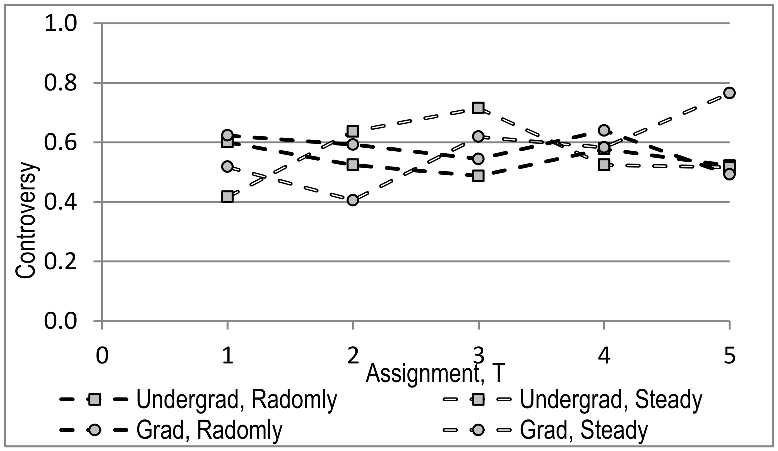


Figure 50. Average Critiques Controversy as Deviation from Mean

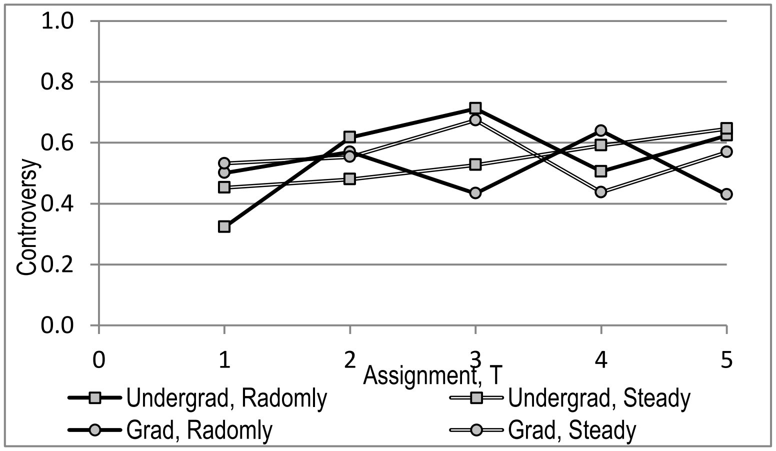


Figure 51. Average Submission Controversy as Deviation from Co-evaluators

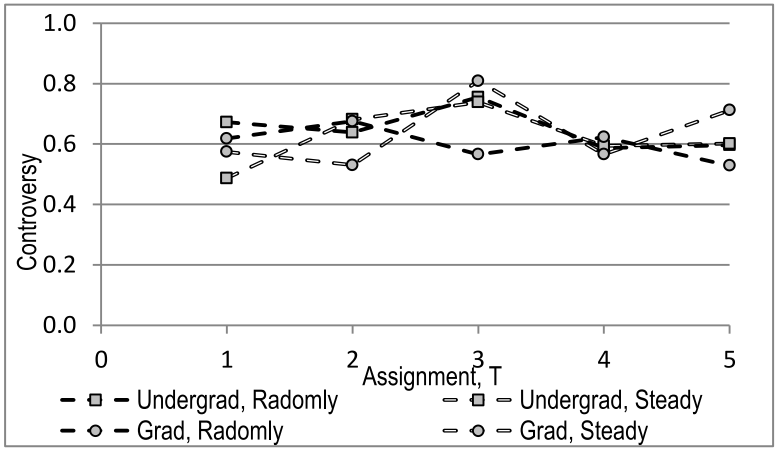


Figure 52. Average Critiques Controversy as Deviation from Co-evaluators

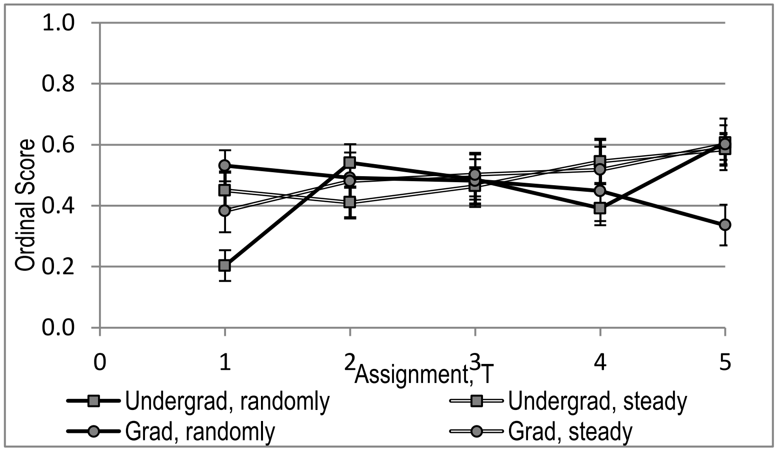


Figure 53. Submission Controversy Deviation from Mean

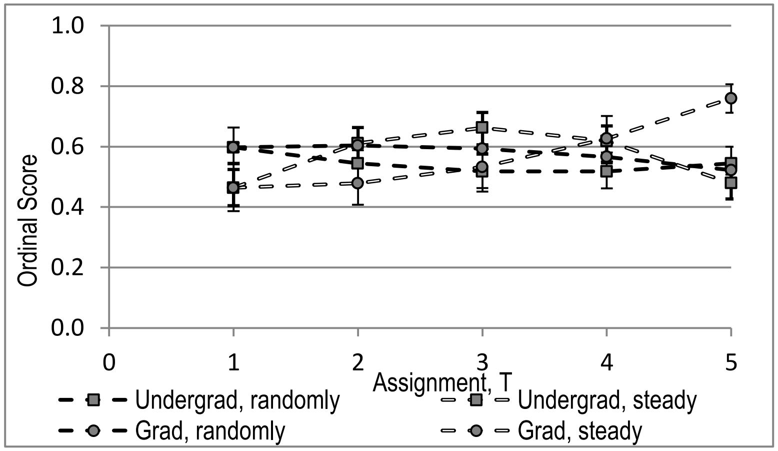


Figure 54. Critiques Controversy Deviation from Mean

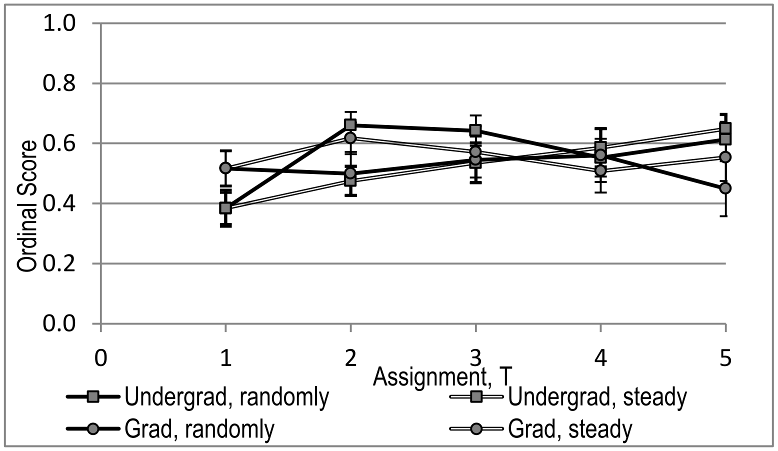


Figure 55. Submission Controversy as Deviation from Co-evaluators

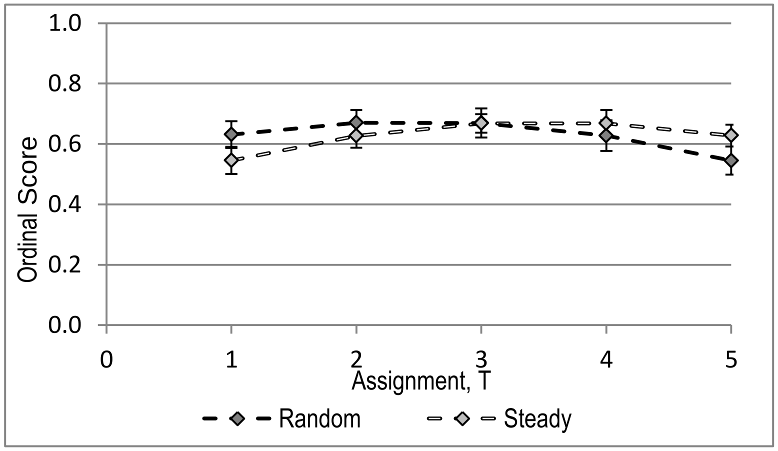


Figure 56. Critiques Controversy as Deviation from Co-evaluators

Table 40. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Controversy as Deviation from Mean

```
Level-1 Model
ScrkStuAr_ContDFM = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*GRAD + B02*RC + B03*(GRAD*RC) + R0
Π1 = B10 + B11*GRAD + B12*RC + B13*(GRAD*RC)
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Π3 = B30 + B31*GRAD + B32*RC + B33*(GRAD*RC)
Fixed Effect          Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           0.203      0.048        4.227         94    0.000
        GRAD, B01           0.328      0.070        4.678         94    0.000
          RC, B02           0.247      0.075        3.290         94    0.002
     GRAD*RC, B03          -0.396      0.112       -3.533         94    0.001
T slope, Π1
    INTRCPT2, B10           0.649      0.119        5.437        447    0.000
        GRAD, B11          -0.722      0.241       -2.993        447    0.003
          RC, B12          -0.759      0.189       -4.017        447    0.000
     GRAD*RC, B13           0.993      0.331        2.995        447    0.003
T2 slope, Π2
    INTRCPT2, B20          -0.371      0.089       -4.166        447    0.000
        GRAD, B21           0.412      0.170        2.429        447    0.016
          RC, B22           0.451      0.131        3.451        447    0.001
     GRAD*RC, B23          -0.568      0.228       -2.496        447    0.013
T3 slope, Π3
    INTRCPT2, B30           0.058      0.016        3.706        447    0.000
        GRAD, B31          -0.067      0.029       -2.353        447    0.019
          RC, B32          -0.070      0.022       -3.105        447    0.002
     GRAD*RC, B33           0.091      0.038        2.381        447    0.018

Random Effect              St Dev   Var Comp           DF    Chi-sq  P-value
INTRCPT1, R0                0.078      0.006           94   120.270    0.035
  level-1, E                0.311      0.097
```

Table 41. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Controversy as Deviation from Co-evaluators

```
Level-1 Model
ScrkStuAr_ContDFC = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*GRAD
Π1 = B10 + B11*GRAD + B12*RC + B13*(GRAD*RC)
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Π3 = B30 + B31*GRAD + B32*RC + B33*(GRAD*RC)
Fixed Effect           Coefficient   St Error      T-ratio        DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            0.384      0.040        9.610       449    0.000
        GRAD, B01            0.133      0.057        2.330       449    0.020
T slope, Π1
    INTRCPT2, B10            0.498      0.118        4.221       449    0.000
        GRAD, B11           -0.579      0.202       -2.866       449    0.005
          RC, B12           -0.384      0.139       -2.762       449    0.006
     GRAD*RC, B13            0.680      0.243        2.795       449    0.006
T2 slope, Π2
    INTRCPT2, B20           -0.259      0.086       -3.015       449    0.003
        GRAD, B21            0.338      0.143        2.355       449    0.019
          RC, B22            0.232      0.106        2.185       449    0.029
     GRAD*RC, B23           -0.448      0.181       -2.470       449    0.014
T3 slope, Π3
    INTRCPT2, B30            0.037      0.015        2.439       449    0.015
        GRAD, B31           -0.053      0.025       -2.136       449    0.033
          RC, B32           -0.034      0.019       -1.763       449    0.078
     GRAD*RC, B33            0.071      0.031        2.242       449    0.025
```

Table 42. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Controversy as Deviation from Mean

```
Level-1 Model
ScrkStuCr_ContDFM = Π0 + Π1*T + Π2*T2+ E
Level-2 Model
Π0 = B00 + B01*RC
Π1 = B10 + B11*GRAD + B12*RC + B13*(GRAD*RC)
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            0.597      0.044       13.500        438    0.000
          RC, B01           -0.134      0.064       -2.099        438    0.036
T slope, Π1
    INTRCPT2, B10           -0.066      0.047       -1.415        438    0.158
        GRAD, B11            0.081      0.075        1.082        438    0.280
          RC, B12            0.261      0.075        3.498        438    0.001
     GRAD*RC, B13           -0.280      0.094       -2.982        438    0.003
T2 slope, Π2
    INTRCPT2, B20            0.013      0.012        1.096        438    0.274
        GRAD, B21           -0.022      0.023       -0.951        438    0.342
          RC, B22           -0.061      0.018       -3.369        438    0.001
     GRAD*RC, B23            0.089      0.028        3.197        438    0.002
```

Table 43. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Controversy as Deviation from Co-evaluators

```
Level-1 Model
ScrkStuCr_ContDFC = Π0 + Π1*T + Π2*T2+ E
Level-2 Model
Π0 = B00 + B01*RC + R0
Π1 = B10 + B11*RC
Π2 = B20
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            0.631      0.038       16.610         94    0.000
          RC, B01           -0.086      0.050       -1.707         94    0.091
T slope, Π1
    INTRCPT2, B10            0.060      0.031        1.908        443    0.057
          RC, B11            0.042      0.020        2.151        443    0.032
T2 slope, Π2
    INTRCPT2, B20           -0.020      0.007       -2.726        443    0.007

Random Effect               St Dev   Var Comp           DF     Chi-sq  P-value
INTRCPT1, R0                 0.074      0.006           94    122.941    0.024
  level-1, E                 0.280      0.079
```

Dynamics of Bias

Visual examination of plots of bias in evaluations of Submissions and Critiques computed as DFM (Figure 57, Figure 58) and as DFC (Figure 59, Figure 60) showed non-trivial patterns for both approaches. Noticeably average controversy and bias showed the same patterns because these variables constitute aggregations of the same basic differences between evaluations by actors. On average, bias dynamics in steady and random groups do not appear to be strikingly different. The only observable visual hint is a more stable increase in controversy of Submission in steady peer groups. This instigated the need for further analysis.

The results of LGM analysis of controversy, presented in Table 14, suggested no evidence that latent classes of actors with characteristic bias trajectories exist. This result held for both Submissions and Critiques, as well as for bias computed as DFM and DFC based on the ordinal scale data (analysis of controversy computed based on the cardinal scale is outside the scope of this dissertation.) Further analysis revealed that, overall, Submission evaluation bias (both DFM and DFC) followed statistically significant cubic trend, with the exception of Critiques Evaluation bias DFM, for which only quadratic trend was significant (Tables 44 through 47; Figures 63 through 66). The experimental conditions variable *RC* had a statistically significant effect (at 5% significance level) on the temporal change rate of bias. The control variable *GRAD* had a statistically significant effect on bias in all cases but Critiques evaluation Bias DFC (for which the effect of *GRAD* on controversy was also nonsignificant). After controlling for the student level

GRAD, steady peer groups demonstrated more noticeable increasing bias than randomly recombined groups. At the same time, Submissions evaluation bias in randomly recombined groups was more unstable, showing a somewhat increasing cubic trend for undergraduate students, and decreasing for graduate students. Thus, overall, there was mixed evidence of the controversy decrease over multiple iterations. However, bias tended to remain higher in the steady groups than in randomly recombined groups, especially on Critiques, similarly to the effects found for controversy.

When miscalibration with respect to peer evaluations of respective KAs was included in the model, it did not pass the significance test; that is, the dynamics of miscalibration did not appear to have strong association with the dynamics of bias. This finding suggests that evaluation competency related to accuracy of evaluating KA created by other actors is not strongly linked to actor's propensity to evaluate his own KA. The reasons for the lack of evidence for association between miscalibration and bias are discussed in chapter V.

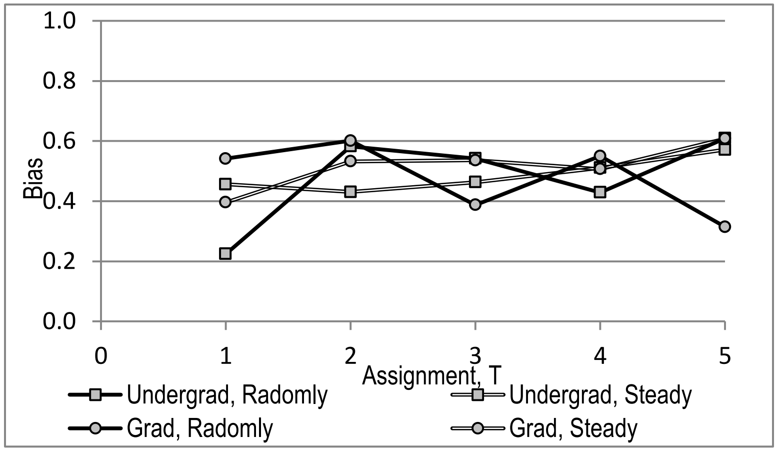


Figure 57. Average Submission Evaluation Bias as Deviation from Mean

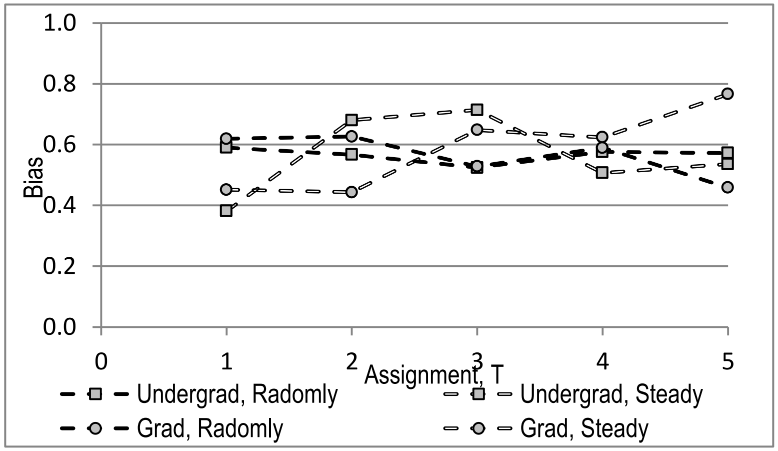


Figure 58. Average Critiques Evaluation Bias as Deviation from Mean

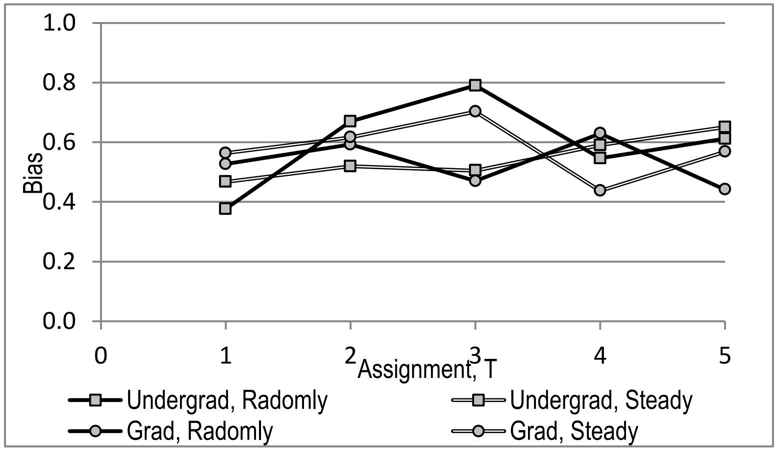


Figure 59. Average Submission Evaluation Bias as Deviation from Co-evaluators

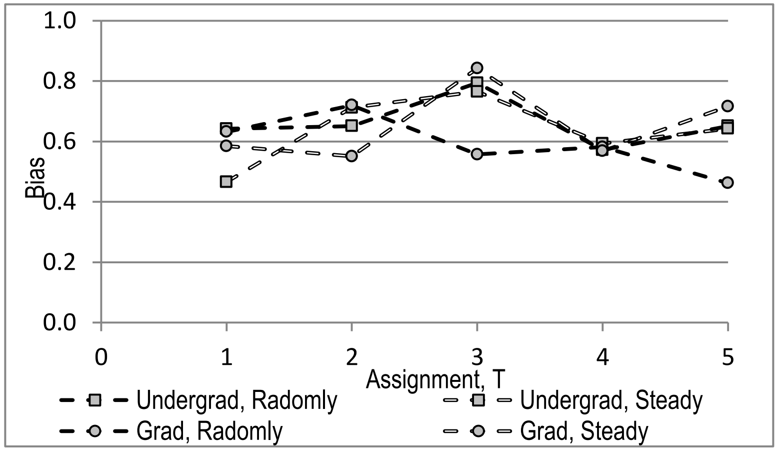


Figure 60. Average Critiques Evaluation Bias as Deviation from Co-evaluators

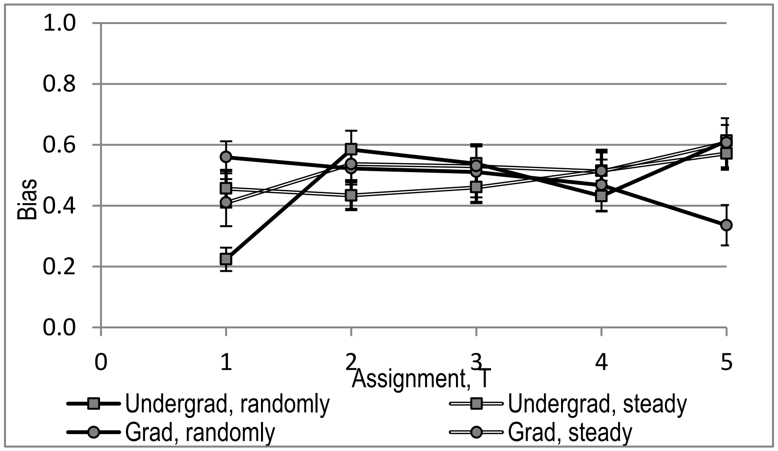


Figure 61. Submission Evaluation Bias as Deviation from Mean (with Indicated Standard Errors)

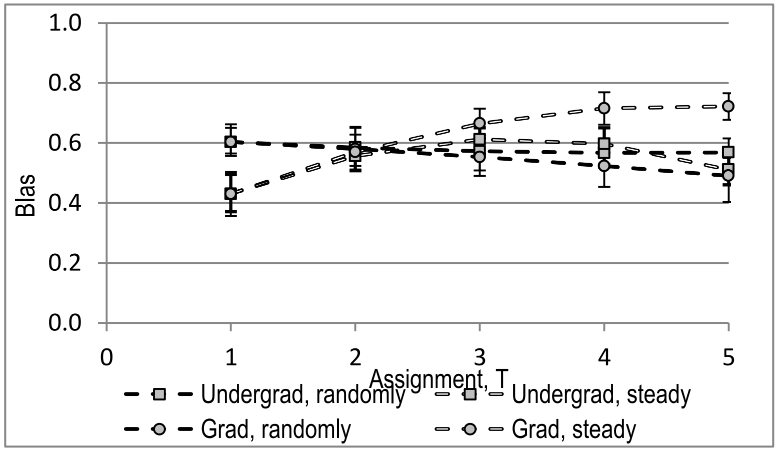


Figure 62. Critiques Evaluation Bias as Deviation from Mean (with Indicated Standard Errors)

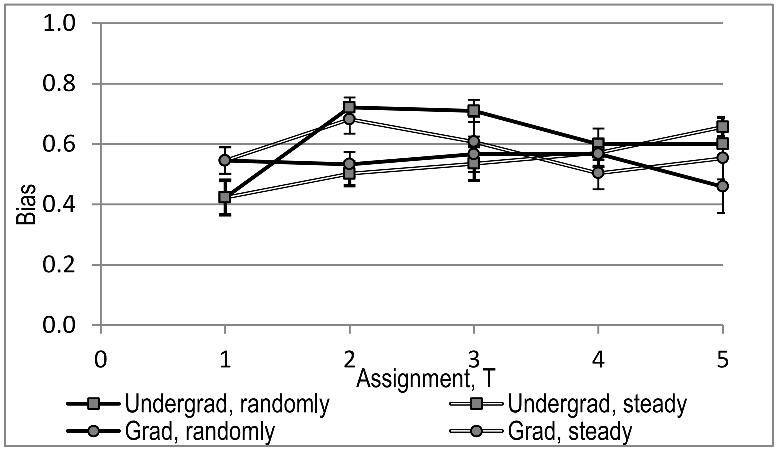


Figure 63. Submission Evaluation Bias as Deviation from Co-evaluators (with Indicated Standard Errors)

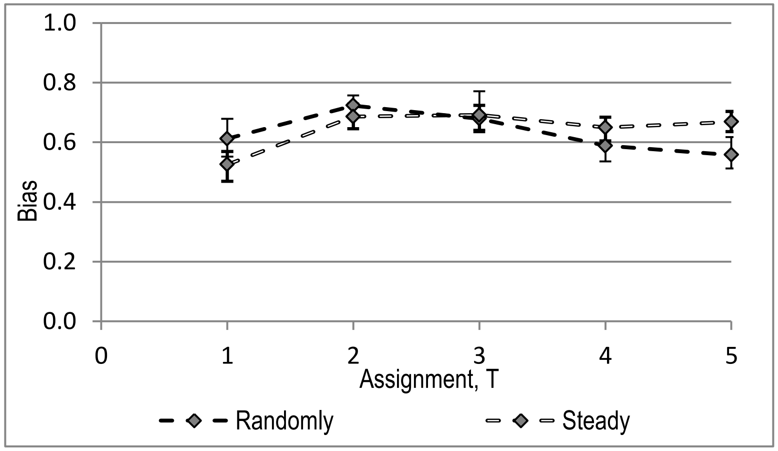


Figure 64. Critiques Evaluation Bias as Deviation from Co-evaluators (with Indicated Standard Errors)

Table 44. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Evaluation Bias as Deviation from Mean

```
Level-1 Model
ScrkStuAr_BiasDFM = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*GRAD + B02*RC + B03*(GRAD*RC)
Π1 = B10 + B11*GRAD + B12*RC + B13*(GRAD*RC)
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Π3 = B30 + B31*GRAD + B32*RC + B33*(GRAD*RC)
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            0.224      0.043        5.244        428    0.000
        GRAD, B01            0.335      0.068        4.932        428    0.000
          RC, B02            0.232      0.073        3.191        428    0.002
     GRAD*RC, B03           -0.382      0.118       -3.224        428    0.002
T slope, Π1
    INTRCPT2, B10            0.680      0.128        5.310        428    0.000
        GRAD, B11           -0.749      0.196       -3.823        428    0.000
          RC, B12           -0.736      0.186       -3.948        428    0.000
     GRAD*RC, B13            1.038      0.300        3.463        428    0.001
T2 slope, Π2
    INTRCPT2, B20           -0.377      0.088       -4.290        428    0.000
        GRAD, B21            0.419      0.129        3.238        428    0.002
          RC, B22            0.414      0.123        3.372        428    0.001
     GRAD*RC, B23           -0.583      0.189       -3.079        428    0.003
T3 slope, Π3
    INTRCPT2, B30            0.058      0.015        3.823        428    0.000
        GRAD, B31           -0.067      0.022       -3.115        428    0.002
          RC, B32           -0.062      0.020       -3.034        428    0.003
     GRAD*RC, B33            0.092      0.031        2.980        428    0.004
```

Table 45. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Submission Evaluation Bias as Deviation from Co-evaluators

```
Level-1 Model
ScrkStuAr_BiasDFC = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*GRAD + R0
Π1 = B10 + B11*GRAD + B12*RC + B13*(GRAD*RC)
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Π3 = B30 + B31*GRAD + B32*RC + B33*(GRAD*RC)
Fixed Effect           Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00            0.421      0.041       10.366         94    0.000
        GRAD, B01            0.124      0.052        2.384         94    0.019
T slope, Π1
    INTRCPT2, B10            0.536      0.089        6.001        434    0.000
        GRAD, B11           -0.595      0.141       -4.214        434    0.000
          RC, B12           -0.424      0.102       -4.171        434    0.000
     GRAD*RC, B13            0.779      0.179        4.358        434    0.000
T2 slope, Π2
    INTRCPT2, B20           -0.267      0.057       -4.674        434    0.000
        GRAD, B21            0.328      0.090        3.639        434    0.001
          RC, B22            0.223      0.070        3.191        434    0.002
     GRAD*RC, B23           -0.477      0.126       -3.797        434    0.000
T3 slope, Π3
    INTRCPT2, B30            0.036      0.009        3.802        434    0.000
        GRAD, B31           -0.049      0.015       -3.207        434    0.002
          RC, B32           -0.028      0.011       -2.437        434    0.015
     GRAD*RC, B33            0.071      0.021        3.317        434    0.001

Random Effect               St Dev   Var Comp           DF     Chi-sq  P-value
INTRCPT1, R0                 0.075      0.006           94    134.381    0.004
  level-1, E                 0.237      0.056
```

Table 46. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Evaluation Bias as Deviation from Mean

```
Level-1 Model
ScrkStuCr_BiasDFM = Π0 + Π1*T + Π2*T2 + E
Level-2 Model
Π0 = B00 + B01*RC + R0
Π1 = B10 + B11*RC
Π2 = B20 + B21*GRAD + B22*RC + B23*(GRAD*RC)
Fixed Effect          Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           0.603      0.041       14.800         95    0.000
          RC, B01          -0.173      0.061       -2.830         95    0.006
T slope, Π1
    INTRCPT2, B10          -0.022      0.049       -0.458        433    0.646
          RC, B11           0.184      0.069        2.656        433    0.009
T2 slope, Π2
    INTRCPT2, B20           0.003      0.012        0.292        433    0.770
        GRAD, B21          -0.005      0.006       -0.867        433    0.387
          RC, B22          -0.039      0.016       -2.438        433    0.015
     GRAD*RC, B23           0.018      0.007        2.733        433    0.007

Random Effect              St Dev   Var Comp           DF    Chi-sq  P-value
INTRCPT1, R0                0.330      0.109           93   125.768    0.013
  level-1, E                1.151      1.326
```

Table 47. HLM Least-squares Estimates of Fixed Effects (with Robust Standard Errors) of Critiques Evaluation Bias as Deviation from Co-evaluators

```
Level-1 Model
ScrkStuCr_BiasDFC = Π0 + Π1*T + Π2*T2+ Π3*T3 + E
Level-2 Model
Π0 = B00 + B01*RC
Π1 = B10 + B11*RC
Π2 = B20
Π3 = B30
Fixed Effect          Coefficient   St Error      T-ratio         DF  P-value
INTRCPT1, Π0
    INTRCPT2, B00           0.612      0.029       21.400        435    0.000
          RC, B01          -0.086      0.038       -2.293        435    0.022
T slope, Π1
    INTRCPT2, B10           0.225      0.067        3.363        435    0.001
          RC, B11           0.049      0.017        2.815        435    0.006
T2 slope, Π2
    INTRCPT2, B20          -0.131      0.042       -3.147        435    0.002
T3 slope, Π3
    INTRCPT2, B30           0.018      0.007        2.667        435    0.008
```

CHAPTER V

DISCUSSION AND CONCLUSION

This chapter summarizes and discusses research findings, theoretical and practical contributions, limitations of the study, and directions for future research.

Research Objectives and Contributions

The primary objective of this dissertation is to evaluate the analytical method and information system design for investigating the change and interactions of actors' creation and evaluation competencies in peer-based knowledge creation and refinements social systems. Prior theoretical and empirical research suggests that actors in such communities may acquire and enhance their competencies as creators and evaluators of knowledge artifacts through social learning as they engage in the interactions of mutual review and evaluations. Moreover, through such interactions, actors develop shared understanding of the topical domains and competencies at the social system level. As actors develop their competencies, they become more discriminating evaluators of KAs developed within the social system. To explore how this understanding transpires through peer evaluation interactions and how it reflects goodness of newly created KAs, this dissertation introduced the notions of social knowledge artifacts (SKA) and social knowledge creation communities (SKCC) to model such social interactions.

The research herein investigated the evaluation intersubjectivity in peer-based SKCC working on complex open-ended problems. The intricacy of solving complex open-ended problems and the subjectivity associated with individuals' understanding of solutions may obscure insights to be gained from peer feedback and impede the development of KA creation and evaluation competencies by actors. Moreover, this understanding may be affected by whether new knowledge is added to the community or lost when members join or leave. Consequently, trajectories of competency development, and, hence, of the value of actors' contributions to SKAs may not be the same across actors in SKCC. This dissertation contributes to the literature by investigating latent development patterns of creation and evaluation competencies, their interactions and the impact on SKAs, as well as the effect of openness of SKCC. Specifically, it answers the research questions:

RQ 1: *How creation and evaluation competencies change and interact over multiple creator-evaluator interactions in peer-based SKCC?*

RQ 2: *How competency dynamics impact SKAs?*

These questions are addressed by examining longitudinal dynamics of attainment, miscalibration and controversy of KAs such as the original solutions to the open-ended problems and peer critiques of these solutions, as well as biases in evaluations of these KAs. This research found some interesting longitudinal patterns in changes of KA creation and evaluation competencies.

This study contributes to the literature by bridging a well-studied domain of peer assessment in education to sparsely researched and scantily understood domain of evaluating solutions to complex open-ended problem KAs in peer-based knowledge creation and refinement communities. The dissertation proposed a design of the DLMA analytical method for operationalizing and investigating the evaluation intersubjectivity in peer-review-based SKCC. In particular, the design includes operationalization of the attainment, miscalibration, controversy and bias constructs in the system of double-looped mutual peer evaluations on ordinal and cardinal scales, and a method for correcting systematic non-linear distortions in the ordinal-scale measurements. On the basis of this analytical method, a peer review system was instantiated, where all participants act as creators and reviewers; reviewers provide evaluations and critiques to KA creators, and creators reciprocally evaluate critiques provided by reviewers in peer groups where each participant in a given assignment evaluates the same set of KA as his peers. The advantages of the process are threefold: (a) it permits collecting data on both the original KA submissions and their critiques evaluated by the same group of participants; (b) it holds accountable not only creators for quality of their submissions but also reviewers for quality of their critiques in response to submissions, creating an extrinsic incentive to provide better critiques; (c) thanks to evaluation reciprocity within the same peer group, it is competencies and biases of evaluators that the variance of ranking of a given KA comes from and not the comparison of the KA within varying sets of peers' KAs.

This study builds on existing social cognitive and educational psychology theories to develop hypotheses about creation and evaluation competency development in SKCC. It hypothesizes that attainment of the KA submissions and critiques as the reflection of the KA goodness, and, hence, of actors' creation and evaluation competencies, improves over multiple iterations. However, latent growth classes of attainment by self-, peer, and expert valuation, with varying attainment trajectories, exist. It further hypothesizes that over multiple double-looped creator-evaluator interactions, intersubjective understanding of various aspects of goodness of KA emerges and the longitudinal trajectories of miscalibration with respect to peer and expert evaluation converge. In addition, artifact controversy and evaluation bias are hypothesized to diminish over time as intersubjective understanding strengthens. The openness of SKCC, i.e., whether its membership remains steady or changes over iterations, is hypothesized to have an effect on the pace of change in attainment, miscalibration, controversy and bias. To answer the posed research questions, these hypotheses were tested using data collected in a controlled repeated-measures experiment conducted with university students.

This study provided a systematic comparison of different measurement and analytical approaches to evaluating development of knowledge artifacts in SKCC. Specifically, a system design proposed in this research permitted collecting evaluations data based on two different scales – ordinal and cardinal – using a single evaluator input GUI control. A typical existing peer review system uses only one of these scales on the basis of design decisions, and there is no consensus about which scale provides stronger peer-evaluation evidence of KA goodness. This study compared results obtained using the two evaluations scales to inform future design of peer evaluation systems. Further, this study adopted the concepts of controversy and bias from the broader research of evaluation systems to the domain of KA peer evaluations. Applying these notions to peer evaluations has the advantages of: (a) detecting anomalies in peer evaluations, such as ranking inconsistencies introduced by specific evaluators and overall variance in evaluation of a specific KA; (b) holding reviewers and creators accountable for evaluations they give to, respectively, submissions and critiques, creating an explicit incentive for them to be as accurate in their evaluations as they can. Finally, the major methodological advance of this study is to apply the LGM method, rarely used in IS research, to analyze actors' behavior.

Summary of Findings and Discussion

This study produced several important findings that inform understanding of the phenomenon of the evaluation intersubjectivity in peer evaluation systems. The LGM analyses results consistently indicate the existence of two latent classes in self-, peer and expert evaluations, as well as in miscalibration, but not in controversy and bias. This result holds for both submissions and critiques, regardless of the scale used. The latent growth classes of attainment essentially differentiate participants with consistently *higher performance* and *lower performance*. In other words, participants in, for example, the higher performance class, have much higher probability of receiving evaluations above average over the entire sequence of iterations. Thus, it can be concluded that their competencies, as assessed by peers or by experts are superior to those of the participants in the lower performance class. The existence of latent attainment classes suggests that the distribution of competency and actors' self-perception of competency is better characterized by latent growth trajectories than by the ‘bell curve’ (Herrnstein & Murray, 1994).

This finding of distinct trajectories of competencies is consistent with the results of the pilot study, which, however, also suggested that in bigger samples with more longitudinal observations the discovery of a larger number of latent classes is possible. Although only two latent classes of attainment and miscalibration with linear patterns were discovered in the experimental sample (with about 100 participants and five repeated measures), the pilot study conducted prior to the experiment with 400 participants and 10 repeated measures suggested that larger samples may reveal more latent classes with more complex, non-linear patterns.

Consistent with the theoretical predictions and the results of the pilot study, the current results support the existence of miscalibration in creators’ and evaluators’ self-evaluation with respect to peer evaluation, as well as expert evaluation. Moreover, as was expected, the ordinal scale reveals the larger *overconfident* class and the smaller *underconfident* class. Surprisingly, however, the cardinal scale reveals a different story – the larger *calibrating* class and the smaller *overconfident* class. Thus, the LGM analysis of miscalibration with respect to peer evaluations reveals two different phenomena depending in the measurement scale.

The existence of two latent classes in self- and peer evaluation, and miscalibration leads to interesting insights about the relationship between miscalibration and performance. Average self-evaluation on the ordinal scale indicates the *above-average effec*t (also known as the *Lake Wobegon effect*), which is directly linked to the Dunnig-

Kruger (the unskilled-and-unaware) effect, pointing out the existence of a large fraction of overconfident participants in the population. While a noticeably large proportion of the overconfident participants comes from the higher performance class, a significant fraction of the overconfident participants comes from the lower performers (referred to as the "unskilled and unaware"). This fraction is between a quarter and a third of the given sample (based on the ordinal scale). Not only these actors over-evaluate their competency but they also fail to realize it and to improve their performance even after receiving peer feedback over multiple iterations. This may be due to the following two reasons – the "unskilled and unaware" participants either were deaf or indisposed towards peers' critiques and suggestions to improve their KAs or peer feedback did not provide useful guidance to lead them in improving their competency and adjusting their self-perceptions. Finding which of these two reasons caused the "unskilled-and-unaware curse" requires more in-depth analyses of the content of peer critiques; it was not performed in this study and presents an interesting opportunity for future research. The important implications of identifying these subjects are, nevertheless, that leading them to competency improvement requires extra intervention, such as expert feedback, and that their critiques and evaluations of other participants KAs may not be as useful and reliable.

On the basis of the social-learning and social norming theories, miscalibration is expected to attenuate over multiple peer interactions over time, as actors continuously receive peer feedback, adjust their expectations and become more self-aware. This prediction for submissions was not supported by the pilot study – miscalibration of both types increased over time, i.e., the overconfident were showing stronger overconfidence, while the underconfident were becoming even more underconfident. In the experimental study, however, the result is even more intricate: just as in the pilot study, both overconfidence and underconfidence with respect to the ordinal-scale peer evaluations of submissions increased; however, critiques miscalibration of both types decreased. (Note that, at the same time, the cardinal scale gained no evidence of either convergence or divergence of miscalibration, for submissions as well as for critiques). This means that while creators grow more persistent in their over- or underconfidence with regard to their solutions to the problem (and the spread between the overconfident and the underconfident widens), intersubjective expectations and evaluations about critiques converge. The implication of this is that while the DLMA provides the basis for diagnostics of actors' performance and their perceptions of performance, for the low performing and highly overconfident participants such peer creator-evaluator interactions may be insufficient to guarantee quick results of social leaning. Longer sequence of peer interactions and/or expert intervention may be needed.

Contradicting miscalibration results between submissions and critiques also suggest that these KAs evaluations differ – either evaluation criteria are different or relative importance placed by actors on them is different. Both critiques and original KA submissions are essentially solutions to complex open-ended problems; however, offering critiques may be perceived by participants as an easier or less important task. At the same time, competency of giving useful critiques in the given population may be underdeveloped.

The posed research questions also tap into another important aspect of intersubjectivity, namely, whether the actors with higher miscalibration of their own KA, are also more biased evaluators, and whether their KAs tend to be more controversial. Conducted correlation and HLM analyses revealed the lack of evidence of the association between miscalibration and bias, as well as between miscalibration and controversy. Neither there was any evidence of the existence of the latent growth classes of bias and controversy. These findings are consistent across the pilot and the experimental studies. Although this study theoretically reasoned the existence of these possible associations and was particularly focused on finding them empirically, investigating the reasons for the lack of such associations is left to future studies.

The study found no significant evidence of the effect of the experimental conditions (randomly changed or steady peer group membership) on self- and peer evaluation, as well as on miscalibration. At the same time, the experimental conditions had a statistically significant effect on bias and controversy. Specifically, bias and controversy appeared to increase over time in steady peer groups, whereas there was a weak evidence that randomly recombined peer groups are better in reducing controversy and bias than steady groups. This effect was more noticeable in critiques than in submissions. Longitudinal changes in bias and controversy typically followed quadratic or cubic trends that complicates interpretation. If such dynamics are confirmed in future studies with sufficient test power, the implication of this result is that peer groups with open membership produce stronger intersubjective consensus over time, i.e., the *cross-pollination* effect and the *ranking confusion* effects dominate over the *social norming* and the *expectations perplexity* effects. In other words, this means that convergence of evaluations is stronger in peer groups with random membership, or, generalizing it to the notion of SKCC, open communities may reach stronger convergence in evaluations than closed communities. This result should be interpreted with caution because it was explored only using the more restrictive ordinal scale.

As discussed above, the choice of evaluation measurement scale influences what conclusions can be made about creation and evaluation competency dynamics. The cardinal scale (rating) shows that more participants are accurate than overconfident, while the ordinal scale (ranking) shows that more participants are overconfident than underconfident. Despite the advantages of ordinal evaluation, its critical disadvantage is that average ranking of the entire sample of participants is always equal to the median of the ordinal scale. Therefore, even if KA goodness changes over time, at the aggregate level, it cannot be observed on the basis of peer or expert rankings. The cardinal scale allows anchoring specific absolute values to specific levels of performance, which may help identify overall changes of competency over time. However, the cardinal scale is susceptible to biases due to evaluators' overall leniency or stringency, initial expectations regarding other actors' performance and changes in expectations over time or while observing a greater number of KAs, and evaluators' subjective interpretation of the rubric criteria. This leads to random and systematic errors in the cardinal evaluations that distort observed patterns of actual performance. The results of this study indicate that while average cardinal peer evaluation tended to increase over time, average cardinal expert evaluations tended to decline. Thus, it can be concluded that the interpretation of peer and expert evaluation results in a SKCC largely depends on the choice of the evaluation scale. This has important implications for peer review system design. Although most peer review systems use only one of the two scales, using both of them simultaneously may provide more data for additional insights. The challenges of practically implementing such solutions are twofold – designing a simple, intuitive and usable evaluation input GUI control and designing a comprehensible representation of the information output to users. The results of this study also show that, although rating data contains more information that can be useful, it also contains more noise in evaluation behaviors and is more sensitive to extreme cases and outliers.

To summarize, these findings reveal some interesting and counter-theoretic results and implications for researchers, educators and decision makers in knowledge management. In concordance with theory, miscalibration regarding KA does exist among many actors in social systems of social knowledge creation. However, contrary to our theory-based expectations, miscalibration of the original KA over multiple iterations does not attenuate, whereas miscalibration of critiques does reduce. The openness of SKCC may not have an effect on miscalibration but it appears to have an effect on bias and controversy. Latent growth classes of attainment and miscalibration exist, whereas such classes of bias and controversy do not exist.

Implications

Understanding dynamics and interactions of competency development through peer evaluations is important and relevant to several audiences. In education, the move toward large-scale, online and hybrid forms teaching and learning, Massive Open Online Courses (MOOCs) being an extreme example, calls for novel instructional methods that engage learners in developing creativity, critical thinking, communication, and collaboration competencies, through complex open-ended assignments and, at the same time, provide meaningful evaluation of these assignments, which are not predisposed to automated assessment. In other words, there is a growing need for developing peer assessment systems as a scalable alternative to the traditional instructor assessments in the face of declining feasibility of the centralized expert evaluation. In learning environment striving to developing high-level cognitive skills and delivering high-level learning objectives in situation when the instructor-to-student ratio is disproportionately low, automated testing is insufficient and, at the same time, instructor assessment is infeasible (Degree of Freedom, 2013; Raman & Joachims, 2014; Shah et al., 2013).

Although design proposed in this dissertation was implemented in an educational peer review system for individual submissions, it can also be adopted and extended in other domains and applications, such as team projects, conference publishing, non-academic large online courses. Conceptualizations of SKA evaluation can also be applied to develop and test models of intersubjectivity in SKCC in open innovation and knowledge crowdsourcing. This dissertation offers interesting insights to inform developments in these and other areas related to technology-enabled peer evaluation systems. For example, in the business organizations, effective knowledge and expertise management is a critical factor for sustained competitive advantage. Understanding how organizational knowledge emerges from individual expertise and is evaluated and validated in peer-based communities is necessary for improving existing and designing new knowledge management mechanisms. For organizations facilitating user-created content and open social knowledge creation in social media (e.g., eHow.com, wikiHow.com, about.com, Pinterest, Yelp, etc.), the ability to capture actionable reliability, validity, and utility metrics for artifacts, creators, and evaluators will improve the efficacy of these platforms.

Limitations and Future Research Directions

The inferences that can be drawn from this study are inevitably constrained by the data collection and analysis methods. At the same time, these limitations lend promising opportunities for future research. First of all, this study used a sample of students from one university taking one course. Therefore, any generalizations to other populations and domains should be made with caution. In the future, this study may be replicated with samples from other populations in academic, as well as in non-academic settings. In addition, the sample size of about 100 participants, although generally considered reasonable for statistical analysis, may be relatively small to ensure statistical power of such methods as LGM and HLM applied to the RCB experiment design. Specifically, the attempted analyses of the expert evaluation turned out to be very sensitive to the outlier cases. Further, the number of longitudinal observations in this study (five), although sufficient for longitudinal research, is not sufficient for studying more complex patterns (Zheng, Pavlou, & Gu, 2014). The impact of the smaller number of participants and longitudinal observations is evidenced by the fact that the pilot study with 450 participants and nine longitudinal observations gained evidence of the larger number of latent classes and non-linear patterns. The undoubted advantage of the experiment design in the current study was a greater control over the sources of variance in comparison with the pilot study.

One of the objectives of this study was to benchmark attainment by self- and peer evaluations against attainment by expert evaluation to answer the question of how goodness of KAs changes over multiple iterations. Unfortunately, the two researchers' expert evaluations showed modest inter-observer reliability. Given the sample size, the analysis of miscalibration with respect to expert evaluation turned out to be very sensitive to several outlier cases and did not produce generalizable results. On the one hand, this was expected because complex assignment evaluations are highly subjective and sensitive to evaluator psychological and social biases, as was shown by past studies, as well as anecdotal evidence (Wagorn, 2008). On the other hand, it leaves the question open for further research of whether the overall competency of actors in the SKCC consistently improves over multiple iterations of creator-evaluator interactions and whether it can be reliably inferred from peer evaluations.

Creation and evaluation competencies of dealing with complex open-ended problems in this study were evaluated holistically, i.e., without differentiating specific skills or abilities. Although participants were provided problem-specific rubrics for peer evaluations, summative evaluation data was collected on the overall goodness. The spirit of this study is that creative problem solving resists standardization and mechanistic decomposition. The argument about the use of different types of rubrics (Goldin, 2011) and the appropriateness of rubrics for evaluating creative tasks (Wilson, 2007) will need to be explored by further research.

While the current version of the DLMA method design permits comprehensive analysis of the intersubjectivity constructs with the ordinal data, the analyses of cardinal evaluation data was covered only partially. Attainment and miscalibration were analyzed, but rating-based bias and controversy variables are yet to be developed. This study demonstrated that although ordinal or cardinal measures of performance are closely correlated, the use of only one of them may lead to incomplete interpretation of actors' intersubjective behaviors. Future studies should provide deeper comparisons of controversy and bias captured using these alternative scales.

The most notable conundrum of this study is the lack of evidence of the relationship between miscalibration and bias. This issue occurred in both the pilot and the experimental studies. The most viable explanation to this is that the miscalibration variable is a vector, i.e., it is described by magnitude (size) and direction (sign), whereas bias and controversy are measured only by absolute value of deviation, i.e. are scalar. Thus, to answer the question of whether miscalibration is associated with bias and controversy, alternative ways of capturing bias and controversy may be necessary. While the measure of miscalibration is very straightforward, and the phenomenon of miscalibration has been well explored (Kruger & Dunning, 1999; Ryvkin et al., 2012), bias and controversy may be captured in a number of ways, e.g., using ordinal or cardinal scales, as deviation from mean or deviation from co-evaluators, on the basis of the naïve or the evidence-based approaches (Kulkarni et al., 2013; Lauw et al., 2008; Piech et al., 2013). The relationship of these constructs with miscalibration and attainment remains largely underexplored, thus, further refinement of these methods and their empirical testing offers a promising direction for future research.

Although past studies indicated the limitations of the naïve approach to capturing bias and controversy in other contexts, in this study it was used intentionally to establish the base line for evaluating performance of the DFM and DFC computational approaches. Further exploration of these mutual dependent constructs requires the use of more sophisticated evidence-based models. The emerging literature suggests studying phenomena of controversy and bias as anomalies in bipartite graphs with mutual dependencies (Dai et al., 2012) and applying Bayesian network analysis (Waters et al., 2015). These approaches offer promising IS research and design opportunities for a variety of domains, including KA evaluation systems such as wikis and other social media applications (Cusinato et al., 2009; Mizzaro, 2003).

The effects of bias and controversy information feedback to participants have not been studied. Two questions are of particular interest in this respect. Firstly, are participants able to understand this performance information and how they correct their creation and evaluation behaviors? In the present study, this information was assumed to be understood and used at least at the very basic level, but no formal testing was intended and made. Secondly, how does behavior change depending on whether this information was provided or not? These research questions present interesting opportunities for further experiments.

This study did not focus on the analysis of formative feedback beyond creators' reciprocal evaluations of evaluators' critiques. As was pointed out above, the content of critiques may possibly play critical role in creators' development of their competencies. Therefore, techniques such as automated content analysis and natural language processing of critiques offer exciting opportunities for extending this research.

Finally, an interesting and important application related to analyses of the evaluation intersubjectivity in SKCC and performance feedback information is the use of data visualization techniques. Visualizing peer review data promises more opportunities for discovering interesting patterns (Xiong, Litman, Wang, & Schunn, 2012). Mobius SLIP offers a basic tool for visualizing mutual evaluation data for both KA submission and critiques. Developing more engaging and comprehensible representations for individual and aggregated data and empirically examining their effects on intersubjectivity dynamics and social learning interactions is an interesting design problem for future research.

REFERENCES

Akerlof, G. A. (1970). The Market for “Lemons”: Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500.

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. http://doi.org/10.1086/268949

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., … Wittrock, M. C. (2001). A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom’s Taxonomy of Educational Objectives. Longman.

Appelbaum, M. I., & Cramer, E. M. (1974). Some Problems in the Nonorthogonal Analysis of Variance. *Psychological Bulletin Psychological Bulletin*, *81*(6), 335–343.

Babik, D., Singh, R., Zhao, X., & Ford, E. (n.d.). What You Think and What I Think? Studying Intersubjectivity in Evaluation of Knowledge Artifacts. *Information Systems Frontiers*, *Under review*.

Bailey, R. A. (2008). *Design of Comparative Experiments* (1st ed.). Cambridge University Press.

Bamberger, P. A. (2005). Peer Assessment, Individual Performance, and Contribution to Group Processes: The Impact of Rater Anonymity. *Group & Organization Management*, *30*(4), 344–377. http://doi.org/10.1177/1059601104267619

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. http://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory* (1st ed.). Prentice Hall.

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Haven [Conn.]: Yale University Press.

Berkowitz, A. D. (2004). The Social Norms Approach: Theory, Research and Annotated Bibliography. *Higher Education Center for Alcohol and Other Drug Abuse and Violence Prevention. US Department of Education*. Retrieved from http://www.alanberkowitz.com/articles/social_norms.pdf

Bigge, M. L., & Shermis, S. S. (2003). *Learning Theories for Teachers* (6 edition). Boston: Pearson.

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. http://doi.org/10.1080/0969595980050102

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). *Taxonomy of Educational Objectives. The Classification of Educational Goals*. 1st ed.].

Boyd, R., & Richerson, P. J. (2002). Group Beneficial Norms Can Spread Rapidly in a Structured Population. *Journal of Theoretical Biology*, *215*(3), 287–296. http://doi.org/10.1006/jtbi.2001.2515

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, *18*(1), 32–42. http://doi.org/10.3102/0013189X018001032

Brown, J. S., & Duguid, P. (2001). Knowledge and Organization: A Social-Practice Perspective. *Organization Science*, *12*(2), 198–213. http://doi.org/10.1287/orsc.12.2.198.10116

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.

Campbell, D. J. (1988). Task Complexity: A Review and Analysis. *Academy of Management Review*, *13*(1), 40–52.

Chan, D. (2002). Latent Growth Modeling. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 302–349). San Francisco, CA, US: Jossey-Bass.

Chesbrough, H. W., Vanhaverbeke, W., & West, J. (2006). *Open Innovation: Researching a New Paradigm*. Oxford: Oxford University Press.

Cho, K., Chung, T. R., King, W. R., & Schunn, C. D. (2008). Peer-based Computer-supported Knowledge Refinement: An Empirical Investigation. *Communications of ACM*, *51*(3), 83–88. http://doi.org/10.1145/1325555.1325571

Cho, K., & Kim, B. (2007). Suppressing Competition in a Computer-Supported Collaborative Learning System. In J. A. Jacko (Ed.), *Human-Computer Interaction. HCI Applications and Services* (pp. 208–214). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-73111-5_24

Cho, K., & Schunn, C. D. (2007). Scaffolded Writing and Rewriting in the Discipline: A Web-based Reciprocal Peer Review System. *Computers & Education*, *48*(3), 409–426. http://doi.org/10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. http://doi.org/10.1037/0022-0663.98.4.891

Competence. (n.d.). *Oxford Advanced American Dictionary*. Retrieved from http://oaadonline.oxfordlearnersdictionaries.com/dictionary/competence

Cramer, E. M., & Appelbaum, M. I. (1980). Nonorthogonal Analysis of Variance - Once Again. *Psychological Bulletin Psychological Bulletin*, *87*(1), 51–57.

Crooks, T. (2001). The Validity of Formative Assessments. University of Leeds.

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. http://doi.org/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).

Degree of Freedom. (2013, April 8). MOOCs and Assessment - Educational Testing. Retrieved from http://degreeoffreedom.org/between-two-worlds-moocs-and-assessment/

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), 331–50.

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.

Ford, E., & Babik, D. (2013, November 21). Methods and Systems for Educational On-Line Methods.

Gagne, R. M. (1985). The Conditions of Learning and Theory of Instruction. Holt Rinehart & Winston.

Gehringer, E. F. (2001). Electronic Peer Review and Peer Grading in Computer-Science Courses. *SIGCSE Bulletin*, *33*, 139–143.

Gillespie, A., & Cornish, F. (2010). Intersubjectivity: Towards a Dialogical Analysis. *Journal for the Theory of Social Behaviour, 40*(1), 19–46. http://doi.org/10.1111/j.1468-5914.2009.00419.x

Goldin, I. (2011). *A Focus on Content: The Use of Rubrics in Peer Review to Guide Students and Instructors*. Retrieved from http://d-scholarship.pitt.edu/8375/1/goldin%2Ddissertation%2D20110805.pdf

Haaga, D. A. F. (1993). Peer Review of Term Papers in Graduate Psychology Courses. *Teaching of Psychology*, *20*(1), 28–32.

Hamer, J., Ma, K. T. K., & Kwong, H. H. F. (2005). A Method of Automatic Grade Calibration in Peer Assessment. In *Proceedings of the 7th Australasian Conference on Computing Education* (Vol. 42, pp. 67–72). Darlinghurst, Australia: Australian Computer Society, Inc. Retrieved from http://dl.acm.org/citation.cfm?id=1082424.1082433

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.

Herrnstein, R. J., & Murray, C. A. (1994). *The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Holsapple, C. W., & Joshi, K. D. (2001). Organizational Knowledge Resources. *Decision Support Systems Decision Support Systems*, *31*(1), 39–54.

Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to Promote Peer Feedback: Pre-Service Teachers' Comments in Asynchronous Computer-Mediated Communication. *Journal of Educational Computing Research*, *43*(1), 89–112. http://doi.org/10.2190/EC.43.1.f

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

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. http://doi.org/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

Jung, T., & Wickrama, K. A. S. (2008). An Introduction to Latent Class Growth Analysis and Growth Mixture Modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317. http://doi.org/10.1111/j.1751-9004.2007.00054.x

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. http://doi.org/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. http://doi.org/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.

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. http://doi.org/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. http://doi.org/10.1037/0022-3514.82.2.189

Kruglanski, A. W. (1989). The Psychology of Being "Right": The Problem of Accuracy in Social Perception and Cognition. *Psychological Bulletin*, *106*(3).

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., … Klemmer, S. R. (2013). Peer and Self Assessment in Massive Online Classes. *ACM Transactions on Computer-Human Interaction*, *20*(6), 1–31. http://doi.org/10.1145/2505057

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.

Lee, C. J., & Schunn, C. D. (2011). Social Biases and Solutions for Procedural Objectivity. *Hypatia*, *26*(2), 352–373.

Li, L., Liu, X., & Zhou, Y. (2012). Give and Take: A Re-analysis of Assessor and Assessee's Roles in Technology-Facilitated Peer Assessment. *British Journal of Educational Technology*, *43*(3), 376–384. http://doi.org/10.1111/j.1467-8535.2011.01180.x

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. http://doi.org/10.1046/j.0266-4909.2001.00198.x

Luhmann, N. (1995). *Social Systems*. Stanford University Press.

Lu, R., & Bol, L. (2007). A Comparison of Anonymous Versus Identifiable E-peer Review on College Student Writing Performance and the Extent of Critical Feedback. *Journal of Interactive Online Learning*, *6*(2), 100–115.

Lynch, C. F., Ashley, K. D., Alven, V., & Pinkwart, N. (2006). Defining "Ill-Defined" Domains: A Literature Survey. *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, 8th International Conference on Intelligent Tutoring Systems*, *80*.

Matusov, E. (1996). Intersubjectivity Without Agreement. *Mind, Culture, and Activity*, *3*(1), 25–45. http://doi.org/10.1207/s15327884mca0301_4

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.

Miller, N. E., & Dollard, J. (1941). *Social Learning and Imitation*. New Haven; London: Institute of Human Relations by Yale University Press; H. Milford, Oxford University Press.

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.

Nonaka, I. (1994). A Dynamic Theory of Organizational Knowledge Creation. *Organization Science*, *5*(1), 14–37.

Nonaka, I., & Toyama, R. (2003). The Knowledge-creating Theory Revisited: Knowledge Creation as a Synthesizing Process. *Knowledge Management Research & Practice*, *1*(1), 2–10. http://doi.org/10.1057/palgrave.kmrp.8500001

Nonaka, I., & Von Krogh, G. (2009). Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory. *Organization Science*, *20*(3), 635–652.

Norman, D. A. (1992). Design Principles for Cognitive Artifacts. *Research in Engineering Design Research in Engineering Design*, *4*(1), 43–50.

Parsons, T. (1991). *The Social System*. Psychology Press.

Perkins, H. W., & Berkowitz, A. D. (1986). Perceiving the Community Norms of Alcohol Use among Students: Some Research Implications for Campus Alcohol Education Programming. *Substance Use & Misuse*, *21*(9-10), 961–976. http://doi.org/10.3109/10826088609077249

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.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned Models of Peer Assessment in MOOCs. In *The 6th International Conference on Educational Data Mining (EDM 2013)*. Retrieved from http://www.stanford.edu/~cpiech/bio/papers/tuningPeerGrading.pdf

Piramuthu, S., Kapoor, G., Zhou, W., & Mauw, S. (2012). Input Online Review Data and Related Bias in Recommender Systems. *Decision Support Systems*, *53*(3), 418–424. http://doi.org/10.1016/j.dss.2012.02.006

Polanyi, M. (2009). *The Tacit Dimension*. University of Chicago Press.

Prins, F. (2006). A Conceptual Framework for Integrating Peer Assessment in Teacher Education. *Studies in Educational Evaluation*, *32*(1), 6–22. http://doi.org/10.1016/j.stueduc.2006.01.005

Raman, K., & Joachims, T. (2014). Methods for Ordinal Peer Grading. *arXiv:1404.3656 [cs]*. Retrieved from http://arxiv.org/abs/1404.3656

Reigeluth, C. M. (1983). *Instructional-design theories and models / Edited by Charles M. Reigeluth*. Hillsdale, N.J. : Lawrence Erlbaum Associates, 1983-.

Reily, K., Finnerty, P. L., & Terveen, L. (2009). Two Peers Are Better Than One: Aggregating Peer Reviews for Computing Assignments is Surprisingly Accurate. In *Proceedings of the ACM 2009 International Conference on Supporting Group Work* (pp. 115–124). New York, NY, USA: ACM. http://doi.org/10.1145/1531674.1531692

Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a General Theory of Planning. *Policy Sciences*, *4*(2), 155–169. http://doi.org/10.1007/BF01405730

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the Unskilled Doomed to Remain Unaware? *Journal of Economic Psychology*, *33*(5), 1012–1031. http://doi.org/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.

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. http://doi.org/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.

Sawyer, R. K. (2008). Optimising Learning Implications of Learning Sciences Research. In *Innovating to Learn, Learning to Innovate* (p. 45).

Schegloff, E. (1982). Discourse as an Interactional Achievement: Some Use of "Uh-Huh" and Other Things That Come Between Sentences. *Georgetown University Round Table on Languages and Linguistics, Analyzing Discourse: Text and Talk*, 71–93.

Schleicher, D. J., Bull, R. A., & Green, S. G. (2008). Rater Reactions to Forced Distribution Rating Systems. *Journal of Management*, *35*(4), 899–927. http://doi.org/10.1177/0149206307312514

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464. http://doi.org/10.1214/aos/1176344136

Searle, J. R. (1992). *The Rediscovery of the Mind*. Cambridge, Mass.: MIT Press.

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

Shepard, L. A. (2007). Formative Assessment: Caveat Emptor. Erlbaum.

Simon, H. A. (1969). *The Sciences of the Artificial*. The MIT Press.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford; New York: Oxford University Press.

Sitthiworachart, J., & Joy, M. (2004). Effective Peer Assessment for Learning Computer Programming. *SIGCSE BULLETIN*, *36*, 122–126.

Slavin, R. E. (1992). When and why does cooperative learning increase academic achievement? Theoretical and empirical perspectives. Cambridge University Press.

Slovic, P. (1995). The Construction of Preference. *American Psychologist*, *50*(5), 364–371. http://doi.org/10.1037/0003-066X.50.5.364

Sluijsmans, D., & Moerkerke, G. (1999). Student Involvement in Performance Assessment: A Research Project. *European Journal of Open, Distance and E-Learning*. Retrieved from http://www.eurodl.org/?p=archives&year=1999&article=38

Spetzler, C. S., & Stael Von Holstein, C.-A. S. (1975). Probability Encoding in Decision Analysis. *Management Science*, *22*(3), 340–358.

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.

Tetreault, M. K. T. (2012). Encyclopedia of Diversity in Education. In *Positionality and Knowledge Construction* (Vols. 1–4, pp. 1676–1677). Thousand Oaks, CA: SAGE Publications, Inc. Retrieved from http://knowledge.sagepub.com/view/diversityineducation/n542.xml

The Definition of Competence. (n.d.). Retrieved June 1, 2014, from http://dictionary.reference.com/browse/competence

Tinapple, D., Olson, L., & Sadauskas, J. (2013). CritViz: Web-Based Software Supporting Peer Critique in Large Creative Classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology*, *15*(1), 29.

Topping, K. J. (1998). Peer Assessment Between Students in Colleges and Universities. *Review of Educational Research*, *68*(3), 249 –276. http://doi.org/10.3102/00346543068003249

Topping, K. J. (2005). Trends in Peer Learning. *Educational Psychology*, *25*(6), 631–645. http://doi.org/10.1080/01443410500345172

Topping, K. J. (2009). Peer Assessment. *Theory into Practice*, *48*(1), 20–27.

Uebersax, J. S. (1988). Validity Inferences from Interobserver Agreement. *Psychological Bulletin*, *104*(3).

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. http://doi.org/10.1016/j.edurev.2008.11.002

Viskovatoff, A. (1999). Foundations of Niklas Luhmann's Theory of Social Systems. *Philosophy of the Social Sciences*, *29*(4), 481–516. http://doi.org/10.1177/004839319902900402

Vista, A., Care, E., & Griffin, P. (2015). A New Approach Towards Marking Large-scale Complex Assessments: Developing a Distributed Marking System That Uses an Automatically Scaffolding and Rubric-targeted Interface for Guided Peer-review. *Assessing Writing*, *24*, 1–15.

Voss, J. F. (2005). Toulmin's Model and the Solving of Ill-Structured Problems. *Argumentation*, *19*(3), 321–329. http://doi.org/10.1007/s10503-005-4419-6

Wagorn, P. (2008, October 15). A Story About a Physics Exam … [Blog]. Retrieved March 1, 2015, from http://www.ideaconnection.com/blog/2008/10/a-story-about-a-physics-exam/

Walsham, G. (2006). Doing Interpretive Research. *European Journal of Information Systems*, *15*(3), 320–330.

Wang, H., Dash, D., & Druzdzel, M. J. (2002). A Method for Evaluating Elicitation Schemes for Probabilistic Models. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society*, *32*(1), 38–43.

Waters, A., Tinapple, D., & Baraniuk, R. (2015). BayesRank: A Bayesian Approach to Ranked Peer Grading. In *ACM Conference on Learning at Scale*. Vancouver.

Wilson, M. (2007). Why I Won't Be Using Rubrics to Respond to Students' Writing. *English Journal*, 62–66.

Wittrock, M. C. (1978). Cognitive Movement in Instruction. *Educational Psychologist*, *13*, 15–29. http://doi.org/10.1080/00461527809529192

Xiong, W., Litman, D., Wang, J., & Schunn, C. D. (2012). An interactive analytic tool for peer-review exploration. In *Proceedings of the Seventh Workshop on Building*

Educational Applications Using NLP (pp. 174–179). Retrieved from http://dl.acm.org/citation.cfm?id=2390405

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. (Eric), Pavlou, P. A., & Gu, B. (2014). Latent Growth Modeling for Information Systems: Theoretical Extensions and Practical Applications. *Information Systems Research*, *25*(3), 547–568. http://doi.org/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.

APPENDIX A

DOUBLE-LOOP MUTUAL ASSESSMENT METHOD

The following set of models gives formal algebraic representation of the Double-Loop Mutual Assessment (DLMA) method. Although each model illustrates a very simple routine, they are described in this sequence to simplify understanding of a more complex model.

Single Group, Single Assignment Model (Model 1)

Ordinal-scale-based Attainment Scores

Suppose, there are *N* subjects in a peer group that are indexed *i* = {1, 2, …, *N*}. Each subject *i* rank-orders other (*N*-1) peers' Submissions so that the "best" is ranked 1 and the "worst" is ranked (*N*-1), that is, each subject *i* does not rank-order his own Submission among other peers' Submissions.

The matrix of ranks of Submissions produced by the group is

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

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's Submission (or, symmetrically, received by a subject j's Submission from a subject i). Note that $\boldsymbol{a}_{i,\ 1\times N} = [a_{i1}\ a_{i2}\ \dots\ a_{ij}\ \dots\ a_{iN}]$ is a row vector of ranks given by a subject i to all his peers' Submissions such that

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

where E_j is the indicator function such that

$$E_j = \begin{cases} 1 \text{ if the Submission was turned in by a subject } j \\ 0 \text{ if the Submission was not turned in by a subject } j. \end{cases}$$

Note that the row vector $\boldsymbol{e}_{1\times N} = [E_1\ E_2\ \dots\ E_N]$ 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 Submission, without exceptions, needs to be rank-ordered. The third condition means that rank-ordering is enforced, that is, no two peers' Submissions may have the same rank. The forth condition means that a missing Submission is given N points.

Similarly, matrix of ranks of Critiques produced by the group is (ranks given are in rows; ranks received are in columns)

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

subject to the data integrity constraints

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

where F_j is the indicator function such that

$$F_j = \begin{cases} 1 \text{ if the Critique was turned in by a subject } j \\ 0 \text{ if the Critique was not turned in by a subject } j. \end{cases}$$

Note that the row vector $\boldsymbol{f}_{1\times N} = [F_1 \; F_2 \; \dots \; F_N]$ is the vector of the indicator function's values.

Suppose now that C is the maximum possible attainment score for a Submission; i.e., the attainment score of C is given to a Submission that received the rank of 1, and the attainment score of 1 is given to a Submission that received the rank of (N-1). A failure to turn in a Submission results in the attainment score of 0. Then the following rule transforms the rank a_{ij} given by a subject i to subject j's Submission into the attainment score c_{ji} received by the subject j's Submission from the subject i:

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

or

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

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 Submission attainment scores for the entire group is

$$\boldsymbol{C}_{N\times N} = \boldsymbol{A}' \frac{(1-C)}{\mathbf{1}\boldsymbol{e}' - 2} + \mathbf{1}'\mathbf{1} \; \frac{C(\mathbf{1}\boldsymbol{e}' - 1) - 1}{\mathbf{1}\boldsymbol{e}' - 2}$$

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

$\mathbf{1}_{1\times N}$ = [*1 1 … 1*] is the row vector of ones, c_{ji} = 0 for all a_{ji} = *N*, and

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

Similarly, if *D* is the maximum possible attainment score for Critique; i.e., the attainment score of *D* is given to a Critique that received the rank of 1; the attainment score of 1 is given to a Critique that received the rank of (*N*-1); a failure to submit a Critique results in the attainment score of 0. Then the transformation rule for the rank b_{ij} given by a subject *i* to a subject *j*'s Critique into the attainment score d_{ij} received by the subject *j*'s Critique from the subject *i* is:

$$d_{ji} = \begin{cases} 1 + (\sum_{h=1}^{N} F_h - 1 - b_{ij}) \dfrac{D-1}{\sum_{h=1}^{N} F_h - 2} \ \text{ if } b_{ij} \neq N \\ 0 \text{ if } b_{ij} = N, \end{cases}$$

or

$$d_{ji} = \begin{cases} b_{ij} \dfrac{(1-C)}{\sum_{h=1}^{N} E_h - 2} + \dfrac{C(\sum_{h=1}^{N} E_h - 1) - 1}{\sum_{h=1}^{N} E_h - 2} \ \text{ if } b_{ij} \neq N \\ \quad 0 \text{ if } b_{ij} = N \end{cases}$$

The matrix of the individual Critique attainment scores for the entire group is

$$\boldsymbol{D}_{N\times N} = \boldsymbol{B}' \frac{(1-D)}{\boldsymbol{1f}' - 2} + \boldsymbol{1}'\boldsymbol{1} \ \frac{D(\boldsymbol{1f}' - 1) - 1}{\boldsymbol{1f}' - 2}$$

where scores received are in rows, scores given are in columns, $d_{ji} = 0$ for all $b_{ji} = N$, and

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

Note that values of *C* and *D* reflect relative weights given to the attainment scores for Submissions and Critiques in the total attainment score for the assignment.

A subject *i*'s Submission attainment score is the average attainment score received from all his peers, who turned in their Critiques and Submission Evaluations, ideally (*N-1*). Hence, the column vector of attainment scores for Submissions is

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

Hence, the subject *i*'s Submission attainment score is

$$\bar{c}_i = \frac{\boldsymbol{c}_i\,\mathbf{1}'}{\mathbf{1}\boldsymbol{f}' - 1}$$

where ${\boldsymbol{c_i}}_{1\times N}$ is the row vector of Submission attainment scores received by a subject *i*.

Similarly, the column vector of Critique attainment scores is

$$\bar{\boldsymbol{d}} = \frac{\boldsymbol{D}\,\mathbf{1}'}{\mathbf{1}\boldsymbol{g}' - 1}$$

where the row vector $\boldsymbol{g}_{1\times N}$ = [G_1 G_2 ... G_N] is the vector of the values of the indicator function G_j such that

$$G_j = \begin{cases} 1 \text{ if Critique Evaluation was turned in by a subject } j \\ 0 \text{ if Critique Evaluation was not turned in by a subject } j \\ 0 \text{ if } E_j = 0 \text{ (if Submission was not turned in by a subject } j\text{)} \end{cases}$$

(the last condition means that if a subject did not turn in a Submission, he cannot turn in Critique Evaluation), and

$$\mathbf{1}\boldsymbol{g}' = \sum_{j=1}^{N} G_j \ \leq N\,.$$

Hence, the subject *i*'s Critique attainment score is

$$\bar{d}_i = \frac{\boldsymbol{d}_i\ \mathbf{1}'}{\mathbf{1}\boldsymbol{g}' - 1}$$

where $\boldsymbol{d}_{i,\ 1\times N}$ is the row vector of Critique attainment scores received by a subject *i*.

Cardinal-scale-based Attainment Score

Relaxing the data integrity assumptions $a_{i1} \neq a_{i2} \neq \cdots \neq a_{ij} \neq \cdots \neq a_{iN}$ and $b_{i1} \neq b_{i2} \neq \cdots \neq b_{ij} \neq \cdots \neq b_{iN}$ allows each Submission and Critique to be assessed on the cardinal scale (rating) rather than ordinal scale (ranking). The rest of the computations remain intact.

Computing Assessor Error as Deviation from Mean

In general, the *assessor error* (ER) is a measure of divergence of evaluations produced by a given subject from evaluations produced by the rest of the peer group in assessing all artifacts produced in the group (Submission or Critique), except for given subject's own artifacts. The assessor error may be computed as either deviation from mean or as deviation from co-evaluators (see Lauw, Lim, Wang, 2006, 2008).

The subject *i*’s *assessor error as deviation from mean* (ERM) on Submission is defined as

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

where | . | denotes the absolute value operator, and ø denotes an undefined (missing) value (assessor error cannot be defined if Submission Evaluation was not turned in). The column vector of assessor errors for Submissions is

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

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 matrix, all elements for which F_i is zero are undefined (missing values).

Similarly, subject *i*’s *assessor error as deviation from mean* (ERM) on Critique is defined as

$$\delta_i^G = \begin{cases} \sum_{j=1}^{N} \left|\bar{d}_j - d_{ji}\right| \;\; \forall\, i \neq j \;\text{ if } G_i = 1 \\ \emptyset \;\text{ if } G_i = 0. \end{cases}$$

The column vector of assessor errors for Critique is

$$\boldsymbol{\delta}^{G} = \left|\left(\bar{\boldsymbol{d}}\mathbf{1}\right)' - \boldsymbol{D}'\right|\mathbf{1}'.$$

In this matrix, all elements for which G_i is zero are undefined (missing values).

Computing Assessor Error as Deviation from Co-evaluators

The subject *i*'s *assessor error as deviation from co-evaluators* (ERC) on Submission is defined as

$$\delta_i^F = \begin{cases} \frac{1}{2}\sum_{h=1}^{N}\sum_{j=1}^{N}\left|c_{hj} - c_{hi}\right| \ \ \forall\, i \neq j, i \neq h \ \text{ if } F_i = 1 \\ \emptyset \ \text{ if } F_i = 0. \end{cases}$$

Similarly, subject *i*'s *assessor error as deviation from co-evaluators* (ERC) on Critiques is defined as

$$\delta_i^G = \begin{cases} \frac{1}{2}\sum_{h=1}^{N}\sum_{j=1}^{N}\left|d_{hj} - d_{hi}\right| \ \ \forall\, i \neq j, i \neq h \ \text{ if } G_i = 1 \\ \emptyset \ \text{ if } G_i = 0. \end{cases}$$

Computing Assessee Error as Deviation from Mean

In general, the *assessee error* (EE) is a measure of divergence among peers on the assessment of a given subject's artifact (Submission or Critique), excluding the subject's self-evaluation. Similarly to the assessor error (ER), the assessee error may be computed as either deviation from mean or as deviation from co-evaluators.

Subject *i*'s *assessee error as deviation from mean* (EEM) on Submissions is defined as

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

The column vector of assessee errors for Submissions is

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

where $\boldsymbol{Q}_{\mathrm{N \times N}}$ 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). Similarly, subject *i*'s *assessee error as deviation from mean* (EEM) on Critiques is defined as

$$\gamma_i^F = \begin{cases} \sum_{j=1}^{N} \left|\bar{d}_i - d_{ij}\right| \ \ \forall\, i \ \neq j \ \text{ if } F_i = 1 \\ \emptyset \ \text{ if } F_i = 0. \end{cases}$$

The column vector of assessor errors for Submissions is

$$\boldsymbol{\gamma}^F = (\left|\bar{\boldsymbol{d}}\mathbf{1} - \boldsymbol{D}\right| \circ \boldsymbol{Q}_{N \times N})\ \mathbf{1}'.$$

Note that in the matrices $\boldsymbol{\gamma}^E$ and $\boldsymbol{\gamma}^F$ the values corresponding to, respectively, *E* and *F* equal 0 are undefined (i.e., missing) values.

Computing Assessee Error as Deviation from Co-evaluators

The subject *i*'s *assessee error as deviation from co-evaluators* (EEC) on Submission is defined as

$$\gamma_i^E = \begin{cases} \frac{1}{2}\sum_{j=1}^{N}\sum_{h=1}^{N}\left|c_{ih} - c_{ij}\right| \;\; \forall\, i \neq j, j \neq h \;\; \text{if } E_i = 1 \\ \emptyset \;\; \text{if } E_i = 0. \end{cases}$$

Similarly, subject *i*'s *assessee error as deviation from co-evaluators* (EEC) on Critiques is defined as

$$\gamma_i^F = \begin{cases} \frac{1}{2}\sum_{h=1}^{N}\sum_{j=1}^{N}\left|d_{hj} - d_{hi}\right| \;\; \forall\, i \neq j, i \neq h \;\; \text{if } F_i = 1 \\ \emptyset \;\; \text{if } F_i = 0. \end{cases}$$

Computing Average Group Errors, Intra-group Inter-observer Reliability, and Normalized Errors

Unlike the use of cardinal scale (rating), the use of ordinal scale (ranking) introduces systematic non-linear distortions in attainment score, assessor error and assesse error calculations, which will be explained below. In order to correct these distortions in the case of the use of ordinal scale, the following computations of "normalized" errors are necessary. In the case of the use of cardinal scale, they can be omitted.

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

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

the *average group assessor error* (AGER) for Critiques is defined as

$$\bar{\delta}^G = \frac{\mathbf{1}\boldsymbol{\delta}^{\boldsymbol{G}}}{\mathbf{1}\boldsymbol{g}'} = \frac{\sum_{i=1}^{N-1}\delta_i^G}{\sum_{i=1}^{N-1}G_i} \ .$$

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

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

the *average group assessee error* (AGEE) for Critiques is defined as

$$\bar{\gamma}^F = \frac{\mathbf{1}\boldsymbol{\gamma}^{\mathbf{F}}}{\mathbf{1}\boldsymbol{f}'} = \frac{\sum_{i=1}^{N-1}\gamma_i^F}{\sum_{i=1}^{N-1}F_i} \ .$$

It can be shown that corresponding AGER and AGEE are equal. That is, $\bar{\delta}^F = \bar{\gamma}^E$ and $\bar{\delta}^G = \bar{\gamma}^F$.

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 $\bar{\delta}$ is the AGER of the given peer group,

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

$\bar{\delta}_{div}$ is the AGER of the peer 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 in evaluating each other's artifacts. 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 IGIOR can be calculated in the same manner for both Submissions and Critiques. Also note that since it can be shown that corresponding AGER and AGEE are equal, it does not matter whether AGER or AGEE are used to compute IGIOR. Computations of $\bar{\delta}_{con}$ and $\bar{\delta}_{div}$ are explained in appendix B. Separate IGIORs are computed for Submissions (y^F) and Critiques (y^G).

Bias (*normalized ER)* is the ER adjusted for the chosen values of *C* or *D* so that it ranges between zero and one. Bias can be interpreted as a measure of a given subject's divergence from evaluations of the rest of the peer group in assessment of all peers' artifact irrespective of the overall rank of the subject’s artifact and the maximum possible attainment score for an artifact.

A given subject *i*’s *bias in Submission Evaluation* is defined as

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

where r_i^F is the rank of the subject *i*’s Submission among the rest of the Submissions of the peer group based on the Submission’s attainment score (in other words, r_i^F corresponding to the larges value in the vector $\bar{\boldsymbol{c}}$ is equal 1 and r_i^F corresponding to the smallest value in the vector $\bar{\boldsymbol{c}}$ is equal *N*);

$\delta_{con}(r_i^F)$ is the ER corresponding to the rank r_i^F 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 $\delta_{con}(r_i^F)$ and δ_{div} are explained in appendix B.

Similarly, *bias in Critique Evaluation* for a given subject *i* is defined as

$$\hat{\delta}_i^G(\delta_i^G, r_i^G) = \frac{\delta_i^G - y^G \ \delta_{con}^G(r_i^G(\bar{d}_i^G))}{\delta_{div}^G}$$

Controversy (*normalized EE)* is the EE adjusted for the chosen values of *C* or *D* so that it ranges between zero and one. 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 a Submission produced by a subject *i* is defined as

$$\hat{\gamma}_i^E(\gamma_i^E, r_i^E\ ) = \frac{\gamma_i^E - y^E\ \ \gamma_{con}^E(r_i^E(\bar{c}_i^E))}{\gamma_{div}^F}\ ;$$

controversy of Critiques given by subject *i* is defined as

$$\hat{\gamma}_i^F(\gamma_i^F, r_i^F\ ) = \frac{\gamma_i^F - y^F\ \ \gamma_{con}^F(r_i^F(\bar{d}_i^F))}{\gamma_{div}^F}.$$

Bias and controversy are recorded in the following vectors:

	Bias	Controversy
Submission	$\hat{\boldsymbol{\delta}}^F$	$\hat{\boldsymbol{\gamma}}^E$
Critiques	$\hat{\boldsymbol{\delta}}^G$	$\hat{\boldsymbol{\gamma}}^F$

In the case of the use of the ordinal scale (raking), normalization is necessary for several reasons. Firstly, *C* may not be equal to *D*, but the measures of subjects' divergence in assessing both Submissions and Critiques need to be comparable. Since ER and EE depend on the chosen values of *C* and *D*, the measures of subjects' divergence need to be normalized. Secondly, the values of *C* and *D* may be different for different courses, yet the measures of subjects' divergences need to be comparable across courses. Finally, while ER and EE are the same for all subjects in the special extreme case of the perfect divergence (because no one is better in assessing others than the rest of the group), in cases with less than the perfect divergence, and in the extreme special case of the with the perfect convergence in particular, ER and EE have non-zero values and are non-linearly dependent on the rank r_i of the subject i in the peer group based on the attainment score. In other words, in a peer group with non-equal attainment scores for an artifact (that is, where at least some convergence exists among peers' assessment of the quality of each artifact, and the artifacts can be ranked according to the attainment score), each rank position is characterized by a systematic non-zero ER and EE just because of its relative ranking place among all artifacts in the peer group. This is due to the fact that subjects' own self-evaluations are not included in the computations of the attainment scores. A detailed explanation of this phenomenon is given in appendix B.

Computing Miscalibration with Respect to Peer Assessment

If, in addition to evaluating peers' artifacts, subjects self-assess their own artifact's, that is, evaluate their own Submissions or Critiques among those of their peers, *miscalibration*, or *self-assessment inaccuracy*, can be computed. Miscalibration is defined as the difference between attainment measures derived from self-assessment and an external assessment source such as peer assessment.

Suppose, in the matrix $\boldsymbol{A}_{N\times N}$, each diagonal element a_{ij} (for which $i = j$) is a rank given by a subject i to his own Submission, such that $a_{ij} \in \{1, 2, \ldots, N\}$. Before proceeding with computation of Attainment scores, let's perform the following operations with $\boldsymbol{A}_{N\times N}$:

(1) Extract the diagonal elements from $\boldsymbol{A}_{N\times N}$ into a separate vector $\boldsymbol{\alpha}_{1\times N} = \mathrm{diag}(\boldsymbol{A}_{N\times N})$;

(2) In each column, subtract one from each a_{ij} (for which $i \neq j$), which is larger than a_{ij} (for which $i = j$). In other words, in each column, each off-diagonal rank value, which is larger than the corresponding diagonal element in the column, should be reduced by one;

(3) Replace all diagonal elements with N (i.e., set a_{ij} (for which $i = j$) to N).

Then, self-assessment Submission Attainment score is defined as

$$\varepsilon_i^F = \begin{cases} (1 + (\sum_{h=1}^{N} E_h - \alpha_i)\dfrac{C-1}{\sum_{h=1}^{N} E_h - 1}) \text{ if } F_i = 1 \\ \emptyset \text{ if } F_i = 0, \end{cases}$$

and subject *i*’s Submission Evaluation miscalibration (self-assessment inaccuracy) is computed as the difference between the self-assessment Submission Attainment score and average peer assessment Submission Attainment score (i.e., as *deviation from mean*)

$$\Delta_i^F = \begin{cases} \dfrac{\varepsilon_i^F - \bar{c}_i}{C-1} \text{ if } F_i = 1 \\ \emptyset \text{ if } F_i = 0. \end{cases}$$

Note that miscalibration contains information not only on the magnitude (size) of inaccuracy but also on its direction (sign). If miscalibration is positive (i.e., self-assessment attainment exceeds peer assessment attainment of the artifact), the subject is said to show *overconfidence*. Likewise, if miscalibration is negative (i.e., self-assessment attainment is smaller than peer assessment attainment of the artifact), the subject is said to show *underconfidence*.

Similarly, if the same manipulations are performed with the matrix $\boldsymbol{B}_{N\times N}$, and β_i is a rank given by a subject i to his own Critiques, then self-assessment Critiques Attainment score is defined as

$$\varepsilon_i^G = \begin{cases} (1 + (\sum_{h=1}^{N} F_h - \beta_i)\frac{D-1}{\sum_{h=1}^{N} F_h - 1}) \text{ if } G_i = 1 \\ \emptyset \text{ if } G_i = 0, \end{cases}$$

and subject i's Critiques Evaluation miscalibration is computed as a difference between the self-assessment Critique Attainment and the peer assessment Critique attainment score, i.e.,

$$\Delta_i^G = \begin{cases} \frac{\varepsilon_i^G - \bar{d}_i}{D-1} \text{ if } G_i = 1 \\ \emptyset \text{ if } G_i = 0. \end{cases}$$

Similarly to assessor and assesse errors, miscalibration with respect to peer assessment can also be calculated as deviation from co-evaluators. It can be easily shown that miscalibration calculated as deviation from mean is identical to miscalibration calculated as deviation from co-evaluators:

$$\frac{1}{N-1}\sum_{j=1}^{N-1}(\varepsilon_i - c_{ij}) = \frac{(N-1)\varepsilon_i}{N-1} - \frac{1}{N-1}\sum_{j=1}^{N-1} c_{ij} = \varepsilon_i - \bar{c}_i$$

Note that if any peer evaluations (of Submissions or Critiques) were missing, *N* has to be substituted for $\sum_{j=1}^{N} E_j$ or $\sum_{j=1}^{N} F_j$ respectively.

Note that miscalibration computed as deviation from mean would not be equal to miscalibration computed as deviation from co-evaluators if absolute values of deviations were used; i.e., if we were interested only in magnitude of miscalibration and not its direction. That is,

$$\frac{1}{N-1}\sum_{j=1}^{N-1}\left|\varepsilon_i - c_{ij}\right| \neq \left|\varepsilon_i - \bar{c}_i\right|$$

Single Group, Multiple Assignments (Model 2)

Model 2 is an extension of Model 1 where instead of a single common assignment subjects are given several sequential assignments indexed by $k = \{1,\ 2,\ \ldots, K\}$. The calculations described in Model 1 repeat *K* times producing matrices $\boldsymbol{A}_{k_{N\times N}}, \boldsymbol{B}_{k_{N\times N}}, \boldsymbol{C}_{k_{N\times N}}, \boldsymbol{D}_{k_{N\times N}}$. The attainment scores, ERs, EEs, AGER, AGEE, IGIOR, bias and controversy are computed for each assignment in a manner identical to the one described for the model 1. The row vector of IGIORs for Submissions is

$$\boldsymbol{y}^F = [y_1^F \quad y_2^F \quad \cdots \quad y_K^F].$$

Similarly, the row vector of IGIORs for Critiques is

$$\boldsymbol{y}^G = [y_1^G \quad y_2^G \quad \cdots \quad y_K^G].$$

Bias and controversy values for Submissions and Critiques for each assignment are recorded in the following vectors:

	Bias	Controversy
Submissions	$\hat{\boldsymbol{\delta}}_k^F$	$\hat{\boldsymbol{\gamma}}_k^E$
Critiques	$\hat{\boldsymbol{\delta}}_k^G$	$\hat{\boldsymbol{\gamma}}_k^F$

Multiple Groups, Single Assignment (Model 3)

Model 3 is an extension of Model 1 where a larger number of subjects M (larger than a small number N) consists of several (L) groups of an approximately equal size N_l; groups are indexed by $l = \{1, 2, \ldots, L\}$. Selecting L such that N_l is close to 6 is recommended. Preferably, each subject is assigned to a group l at random, which requires a simplest random assignment mechanism based on the uniform distribution. The calculations described in Model 1 are performed for each of L groups (replacing N with N_l), producing matrices $\boldsymbol{A}_{l_{N\times N}}$, $\boldsymbol{B}_{l_{N\times N}}$, $\boldsymbol{C}_{l_{N\times N}}$, $\boldsymbol{D}_{l_{N\times N}}$. Column vectors of normalized ERs and EEs are produced similarly. The column vector of IGIORs for Submissions for all subjects is

$$\boldsymbol{y}^E = \begin{bmatrix} y_1^F \\ y_2^F \\ \vdots \\ y_L^F \end{bmatrix}$$

and the column vector of IGIORs for Critique for all subjects is

$$\boldsymbol{y}^F = \begin{bmatrix} y_1^G \\ y_2^G \\ \vdots \\ y_L^G \end{bmatrix}$$

Multiple Groups, Multiple Assignments (Model 4)

Model 4 is a hybrid of Model 2 and Model 3 in which

1. All subjects are given several sequential assignments indexed by $k = \{1, 2, \ldots, K\}$ (with all assumptions of Model 2);
2. For each assignment, all M subjects are divided into L groups of the size of N_l indexed by $l = \{1, 2, \ldots, L\}$ (with all assumptions of Model 3); $M = \sum_{l=1}^{L} N_l$;
3. In each assignment, subjects are divided into groups randomly, so that a subject is placed in a new group of random peers;
4. Specific task given to a group may be the same for all groups or unique to each group; in any case, subjects within each group independently work on the same group-specific task.

The Submission IGIOR matrix is

$$\boldsymbol{Y}_{L\times K}^F = [y_{lk}^F]_{L\times K} = \begin{bmatrix} y_{11}^F & y_{12}^F & \cdots & y_{1K}^F \\ y_{21}^F & y_{22}^F & & y_{2K}^F \\ \vdots & & \ddots & \vdots \\ y_{L1}^F & y_{L2}^F & \cdots & y_{LK}^F \end{bmatrix}$$

and the Critique IGIOR matrix is

$$\boldsymbol{Y}_{L\times K}^{G} = [y_{lk}^{G}]_{L\times K} = \begin{bmatrix} y_{11}^{G} & y_{12}^{G} & \cdots & y_{1K}^{G} \\ y_{21}^{G} & y_{22}^{G} & & y_{2K}^{G} \\ \vdots & & \ddots & \vdots \\ y_{L1}^{G} & y_{L2}^{G} & \cdots & y_{LK}^{G} \end{bmatrix}.$$

The row vector of the subject *i*'s bias values for Submissions for all assignments is

$$\widehat{\boldsymbol{\delta}}_{i}^{F} = \begin{bmatrix} \hat{\delta}_{1i}^{F} & \hat{\delta}_{2i}^{F} & \cdots & \hat{\delta}_{Ki}^{F} \end{bmatrix},$$

Then the subject *i*'s average bias for Submissions is equal

$$\hat{\delta}_{i}^{F} = \frac{1}{K}\,\widehat{\boldsymbol{\delta}}_{i}^{F}\,\mathbf{1}'_{1\times K} = \sum_{k=1}^{K} \hat{\delta}_{ki}$$

The subject *i*'s average bias for Critiques, as well as average controversy for Submissions and Critiques are defined similarly.

Multiple Groups, Multiple Assignments, Multiple Criteria (Model 5)

Model 5 is an extension of Model 4 where peers' Submissions are to be assessed based not on a single criterion of overall quality but on several more specific criteria indexed by $u = \{1, 2, \ldots, U\}$. Criteria are assumed to be the same for all assignments. Similarly, peers' Critiques are ranked based on several criteria indexed by $v = \{1, 2, \ldots, V\}$. To utilize multiple criteria for assessment, Models 1, 2 and 3 can be extended in a similar fashion.

Then, for an assignment *k*, for a given group *l* of the size N_l, the matrix of ranks of Submissions based on a criterion *u* is

$$\boldsymbol{A}_{ulk_{N\times N}} = [a_{ulkij}]_{N\times N} = \begin{bmatrix} N & a_{ulk12} & \dots & a_{ulk1N} \\ a_{ulk21} & N & & a_{ulk2N} \\ \vdots & & \ddots & \vdots \\ a_{ulkN1} & a_{ulkN2} & \cdots & N \end{bmatrix}$$

where a_{ulkij} is the rank given by a subject i to the Submission of a subject j in a group l on an assignment k based on an Submission criterion u. Matrix $\boldsymbol{B}_{vlk_{N\times N}}$ is defined similarly, with b_{vlkij} being the rank given by a subject i to the Critique of a subject j in a group l on an assignment k bases on a Critique criterion v. The matrices of attainment scores for each criterion u and v, $\boldsymbol{C}_{ulk_{N\times N}}$ and $\boldsymbol{D}_{vlk_{N\times N}}$ respectively, are defined as described in Model 1, assuming that the maximum possible attainment score is the same for all criteria. The matrices of attainment scores aggregating all criteria for a group l and assignment k are defined as weighted averages of the matrices of scores for individual criteria:

$$\boldsymbol{C}_{lk_{N\times N}} = \frac{\sum_{u=1}^{U} \boldsymbol{C}_{ulk_{N\times N}} w_u}{\sum_{u=1}^{U} w_u}$$

$$\boldsymbol{D}_{lk_{N\times N}} = \frac{\sum_{v=1}^{V} \boldsymbol{D}_{vlk_{N\times N}} z_v}{\sum_{v=1}^{V} w_u}$$

where w_u is the weight of a criterion u in the Submission score and z_v is the weight of criterion v in the Critique score.

Hence, the column vector of attainment scores for a group l for Submissions in an assignment k is

$$\bar{\boldsymbol{c}}_{lk} = \frac{\boldsymbol{c}_{lk_{N\times N}}\,\mathbf{1}'}{\sum_{i=1}^{N-1} E_i}$$

and the column vector of attainment scores for a group *l* for Critiques in an assignment *k* is

$$\bar{\boldsymbol{d}}_{lk} = \frac{\boldsymbol{d}_{lk_{N\times N}}\,\mathbf{1}'}{\sum_{i=1}^{N-1} F_i}$$

Other extensions of the basic model are also possible but not discussed here.

APPENDIX B

COMPUTING ORDINAL-SCALE AVERAGE GROUP ERRORS

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

Perfect Convergence

First, consider the case of the *perfect intra-group inter-observer convergence,* that is, the result of mutual summative peer assessment, for which IGIOR is equal one. Suppose the row vector $\boldsymbol{s_{1\times N}}$ = [*1, 2, …, N*] is the vector of latent ranks of potential attainment (*goodness*) of a given set of artifacts (e.g., Submissions). That is, it is assumed that each artifact is of such *goodness* and each subject has such evaluation skill that when asked to rank-order these artifacts, the subjects come to the perfect convergence on ranking of each artifact (it is also assumed that all subjects turn in their artifacts). Under these assumptions, the latent ranks should be equal to the ranks generated by the DLMA system, that is $s_i = r_i$ for all *i*. However, since each Artifact's attainment is computed using rankings received from peers (ranging in {*1, 2, …, 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 Submission 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

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

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:

$$\begin{vmatrix} 6 & 1 & 2 & 3 & 4 & 5 \\ 1 & 6 & 2 & 3 & 4 & 5 \\ 1 & 2 & 6 & 3 & 4 & 5 \\ 1 & 2 & 3 & 6 & 4 & 5 \\ 1 & 2 & 3 & 4 & 6 & 5 \\ 1 & 2 & 3 & 4 & 5 & 6 \end{vmatrix}$$

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

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

such that

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

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

$$\begin{vmatrix} 0.00 & 5.00 & 5.00 & 5.00 & 5.00 & 5.00 \\ 5.00 & 0.00 & 4.00 & 4.00 & 4.00 & 4.00 \\ 4.00 & 4.00 & 0.00 & 3.00 & 3.00 & 3.00 \\ 3.00 & 3.00 & 3.00 & 0.00 & 2.00 & 2.00 \\ 2.00 & 2.00 & 2.00 & 2.00 & 0.00 & 1.00 \\ 1.00 & 1.00 & 1.00 & 1.00 & 1.00 & 0.00 \end{vmatrix}$$

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

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

The column vector of assessor error computed as deviation from mean (ERM) is

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

where | . | denotes the matrix of absolute values of the element-wise differences of the two square matrices.

The column vector of assessee error computed as deviation from mean (EEM) is

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

where $\boldsymbol{Q}_{\mathrm{N\times N}}$ 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, ERM and EEM scores for the case of the peer group with the perfect convergence where N = 6 and C = 5:

s = r	$\bar{\boldsymbol{c}}_{con}$	$\boldsymbol{\delta}_{con}(\boldsymbol{r})$	$\boldsymbol{\gamma}_{con}(\boldsymbol{r})$
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

Figure 65 graphically illustrates attainment, ERM and EEM scores as deviations from mean for the case of the peer group with the perfect convergence where *N* = 6 and *C* = 5.

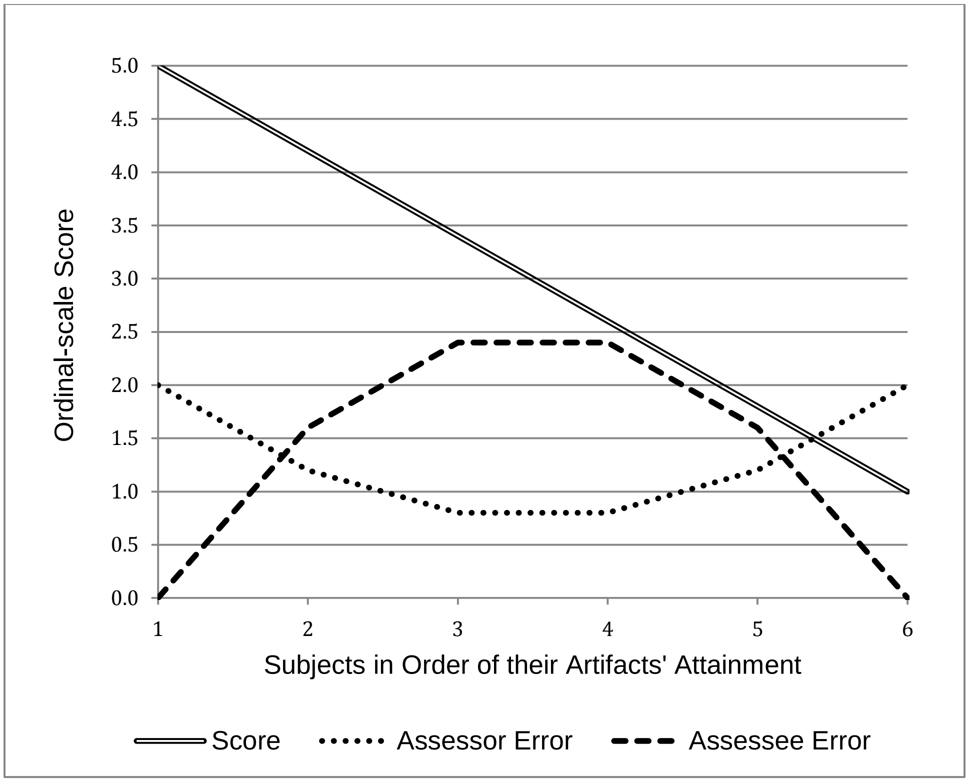


Figure 65. Non-linear Error Behavior in Perfect Convergence (Deviation from Mean)

Thus, ERM and EEM scores of for each relative rank in a peer group with the perfect convergence are obtained.

Similarly, the column vector of assessor error computed as deviation from co-evaluators (ERC) is

$$\boldsymbol{\delta}_{con} = \frac{1}{2}\sum_{h=1}^{N}\sum_{j=1}^{N}\left|c_{hj} - c_{hi}\right| \quad \forall\, i \neq j, i \neq h$$

where | . | denotes the matrix of absolute values of the element-wise differences of the two square matrices.

The column vector of assessee error computed as deviation from co-evaluators (EEC) is

$$\boldsymbol{\gamma}_{con} = \frac{1}{2}\sum_{j=1}^{N}\sum_{h=1}^{N}\left|c_{ih} - c_{ij}\right| \;\; \forall\, i \neq j, j \neq h$$

The following table summarizes the attainment, ERC and EEC scores for the special case of the peer group with the perfect convergence where *N* = 6 and *C* = 5:

$\boldsymbol{s = r}$	$\bar{\boldsymbol{c}}_{con}$	$\boldsymbol{\delta}_{con}(\boldsymbol{r})$	$\boldsymbol{\gamma}_{con}(\boldsymbol{r})$
1	5.00	10.00	0.00
2	4.20	6.00	4.00
3	3.40	4.00	6.00
4	2.60	4.00	6.00
5	1.80	6.00	4.00
6	1.00	10.00	0.00

Figure 66 graphically illustrates attainment, ERC and EEC scores as deviations from co-evaluators for the case of the peer group with the perfect convergence where *N* = 6 and *C* = 5.

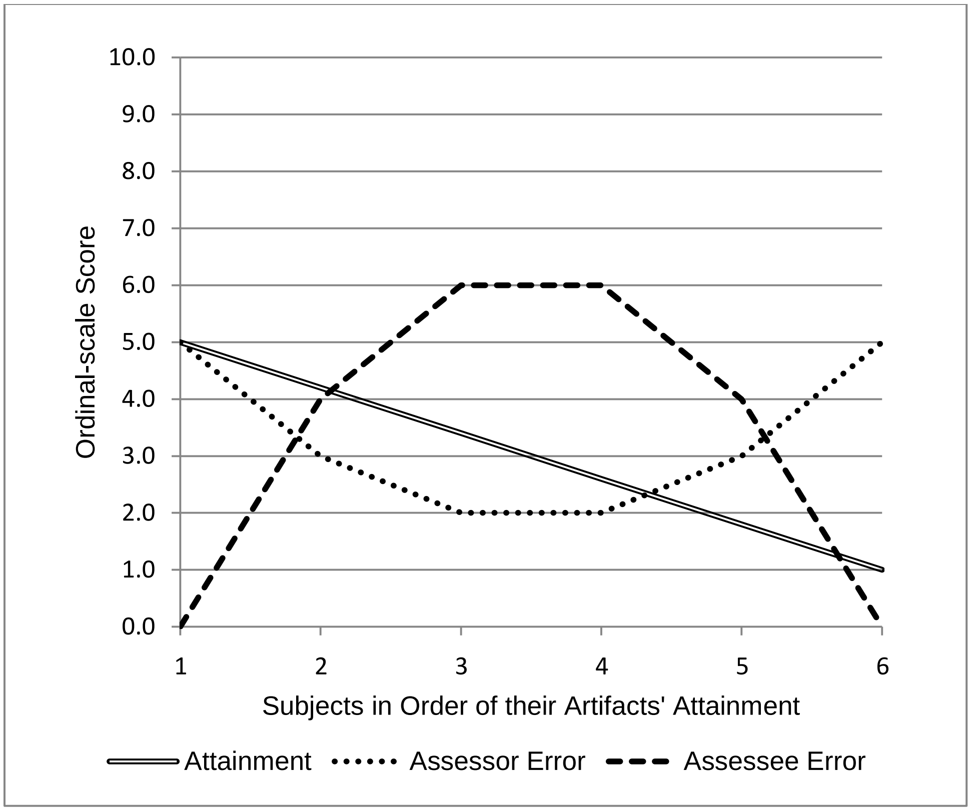


Figure 66. Non-linear Error Behavior in Perfect Convergence (Deviation from Co-evaluators)

Thus, ERC and EEC scores of for each relative rank in the peer group with the perfect convergence are obtained. The AGER for the peer group with the perfect convergence is

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

For the peer group with the perfect convergence where *N* = 6 and *C* = 5, the deviations from mean AGER $\bar{\delta}_{con}$ is equal 4/3; the deviations from co-evaluators AGER $\bar{\delta}_{con}$ is equal 10/3. The AGEE for the peer group 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 in the perfect convergence where *N* = 6 and *C* = 5, the deviations from mean AGEE $\bar{\gamma}_{con}$ is equal 4/3; the deviations from co-evaluators AGEE $\bar{\gamma}_{con}$ is equal 10/3.

Thus, values of AGER and AGEE for the extreme special case of the perfect convergence are obtained.

Perfect Divergence

Now, 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, ..., N*] does not reflect the latent ranks of *goodness* of a given set of artifacts. That is, it is assumed that each artifact is of such *goodness*, and/or each subject has such assessment skill that when asked to rank-order these artifacts, the subjects are not be able to converge in evaluations of the *goodness* of artifacts and their ranking, resulting in the perfect divergence on ranking of each artifact (it is also again assumed that all subjects turn in their artifacts). 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 peers the entire range of possible ranks, and no one subject is better in assessing his peers' artifacts 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*×*N* matrix filled with integers {*1, 2, ..., 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 ***s*** in reverse order (or $(N+1)\mathbf{1}-\ \boldsymbol{s}$) in the top row of ***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 ***A****div* looks as follows:

$$\begin{vmatrix} 6 & 5 & 4 & 3 & 2 & 1 \\ 1 & 6 & 5 & 4 & 3 & 2 \\ 2 & 1 & 6 & 5 & 4 & 3 \\ 3 & 2 & 1 & 6 & 5 & 4 \\ 4 & 3 & 2 & 1 & 6 & 5 \\ 5 & 4 & 3 & 2 & 1 & 6 \end{vmatrix}$$

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

Similarly to the extreme special case of the perfect convergence, using the rule for transforming ranks into scores the matrix of attainment scores ***C****div* is obtained as

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

such that

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

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

$$\begin{vmatrix} 0.00 & 5.00 & 4.00 & 3.00 & 2.00 & 1.00 \\ 1.00 & 0.00 & 5.00 & 4.00 & 3.00 & 2.00 \\ 2.00 & 1.00 & 0.00 & 5.00 & 4.00 & 3.00 \\ 3.00 & 2.00 & 1.00 & 0.00 & 5.00 & 4.00 \\ 4.00 & 3.00 & 2.00 & 1.00 & 0.00 & 5.00 \\ 5.00 & 4.00 & 3.00 & 2.00 & 1.00 & 0.00 \end{vmatrix}$$

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

$$\bar{\boldsymbol{c}}_{div} = \frac{\boldsymbol{C}_{div}\ \mathbf{1}'}{\mathbf{1}\boldsymbol{f}'}.$$

The column vector of assessor error computed as deviation from mean (ERM) is

$$\boldsymbol{\delta}_{div} = |(\bar{\boldsymbol{c}}_{div}\mathbf{1})' - \boldsymbol{C}_{div}'|\ \mathbf{1}'\ .$$

The column vector of assessee error computed as deviation from mean (EEM) is

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

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

$\boldsymbol{s = r}$	$\bar{\boldsymbol{c}}_{div}$	$\boldsymbol{\delta}_{div}(\boldsymbol{r})$	$\boldsymbol{\gamma}_{div}(\boldsymbol{r})$
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, values of ER and EE are obtained for each relative rank in the peer group with the perfect divergence. Note that all subjects' submissions in such group are characterized by equal attainment, ER and EE scores.

Similarly, the column vector of assessor error computed as deviation from co-evaluators (ERC) is

$$\boldsymbol{\delta}_{div} = \frac{1}{2}\sum_{h=1}^{N}\sum_{j=1}^{N}\left|c_{hj} - c_{hi}\right| \ \ \forall\, i \neq j, i \neq h$$

where | . | denotes the matrix of absolute values of the element-wise differences of the two square matrices.

The column vector of assessee error computed as deviation from co-evaluators (EEC) is

$$\boldsymbol{\gamma}_{div} = \frac{1}{2}\sum_{j=1}^{N}\sum_{h=1}^{N}\left|c_{ih} - c_{ij}\right| \ \ \forall\, i \neq j, j \neq h$$

The following table summarizes the attainment, ERC and EEC scores for the special case of the peer group with the perfect convergence where *N* = 6 and *C* = 5:

$\boldsymbol{s = r}$	$\bar{\boldsymbol{c}}_{con}$	$\boldsymbol{\delta}_{con}(\boldsymbol{r})$	$\boldsymbol{\gamma}_{con}(\boldsymbol{r})$
1	5.00	20.00	20.00
2	4.20	20.00	20.00
3	3.40	20.00	20.00
4	2.60	20.00	20.00
5	1.80	20.00	20.00
6	1.00	20.00	20.00

The AGER for the peer group in the perfect divergence is

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

For the peer group with the perfect divergence where *N* = 6 and *C* = 5, the deviations from mean AGER $\bar{\delta}_{div}$ is equal 6.00; the deviations from co-evaluators AGER $\bar{\delta}_{div}$ is equal 20.00.

The AGEE for the peer group with the perfect divergence is

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

For the peer group with the perfect divergence where *N* = 6 and *C* = 5, the deviations from mean AGEE $\bar{\delta}_{div}$ is equal 6.00; the deviation from co-evaluators AGEE $\bar{\delta}_{div}$ is equal 20.00.

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

APPENDIX C

SAMPLE ASSIGNMENT

Title: Assignment 2 - Use Case Diagram and Descriptions

BACKGROUND:

Consider the problem, business needs and functionality you worked on in Assignment 1. A serial entrepreneur who has this fantastic idea for a game to train employees has approached you. In this game, you are timed and have to answer multiple-choice questions (think Buffalo Wild Wings trivia game.) Based on how accurate and how fast you respond in this game, is how you are ranked in your management team. The questions are on a centralized system, and you have no control over the actual content that is being asked; you must build the system so that it functions independently of any content.

PURPOSE:

The purpose of this assignment is to help you begin to analyze and model the business needs of the system. In working on this assignment, please carefully consider feedback you received from your peers on assignment 1.

DELIVERABLES:

This assignment will require you to complete three tasks: Submission, Review, and Reaction.

For the Submission, turn in a single PDF document (of no more than three pages; do not put your name on the document) presenting a use case diagram that identifies the primary actors (roles played by users) and business processes for which a new system is requested.

The diagram should include:

- Clearly identified actors of the system;

- The list of goals the actors have in the use of the system.

In addition, provide fully developed descriptions of TWO use cases for most primary business processes in your system. Feel free to use attached use case description template but save all work (diagrams and description tables) in a single PDF document. See schedule for Submission deadline.

In Review, review Submissions of 4 - 5 of your peers and write short critiques to them, as well as a self-critique. In addition, compare their and your own Submissions to each other and provide holistic evaluation of them in the order of merit (based in the rubric below) using the colored SLIP Slider bar. See schedule for Review deadline.

In Reaction, compare Reviews from 4 - 5 of your peers and your own Reviews to each other and evaluate them in the order of insightfulness, helpfulness, and professionalism. In other words, does a peer’s review help you make your next submission better? See schedule for Reaction deadline.

You will have to turn in all three parts for each of your assignments by corresponding deadline to receive full credit. If you missed the Submission, you can still turn in your Review, but you will not be able to submit Reaction. If you missed Reaction but turned in your Submission, your score for Review will be reduced by 50%.

GUIDELINES:

To successfully complete this assignment, please use the rubric below for writing and evaluating Submissions.

RUBRICS

Present a use case diagram that identifies the primary actors and business processes for which a new system is requested. In addition, provide fully developed descriptions of TWO use cases for most primary business processes in your system.

	Excellent	Good	Fair	Poor	Very Poor
Use case diagram completeness Does use case diagram identify all required actors and processes?	The diagram clearly identifies all actors and goals/processes necessary to satisfy the system's requirements	The diagram identifies all actors and most goals/processes necessary to satisfy the system's requirements but one or two goals may be missing	The diagram identifies most actors and goals/processes necessary to satisfy the system's requirements but several may be missing	Only a few actors and/or goals/processes are presented but very many are missing	The diagram does not sufficiently present actors and/or goals/processes to satisfy the system's requirements
Use case diagram presentation Is use case professionally presented?	The diagram is very clearly and professionally presented according to the examples in the textbook	The diagram is clearly and professionally presented but has minor inconsistencies	The diagram is mostly well presented but has several flaws	The diagram has serious clarity and formatting issues	The diagram is very unclear and/or unprofessional
Use case description completeness Do use case descriptions contain all necessary components?	The descriptions contain all necessary components to clearly and fully describe two primary use cases	The descriptions contain all necessary components buy some are not clearly described	The descriptions contain most of necessary components buy some are missing or poorly described	The descriptions contain only some of the necessary components; most components very poorly described	Many of the necessary components are missing and/or very poorly described; use case descriptions are missing all together
Use case description presentation Are use case descriptions professionally presented?	The descriptions are very clearly and professionally presented according to the template	The descriptions are very clearly and professionally presented but have minor flaws/inconsistencies	The descriptions are mostly well presented but have several flaws	The descriptions have serious clarity and formatting issues	Very unprofessional or missing