Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 641–646
DOI: 10.15439/2025F8584
ISSN 2300-5963 ACSIS, Vol. 43

Cognitive-Aware Peer Assessment: Design Implications from a Classroom Deployment

Naama Bouskila, Lihi Dery
0009-0001-1996-3451
0000-0002-8710-3349
Ariel University
Ariel, Israel
Email: {naama.bouskila, lihid}@ariel.ac.il

***Abstract*—Peer assessment is widely used in higher education, yet the cognitive demands placed on student assessors, particularly under conditions of overload and repetition, remain poorly understood. We examine how two cognitive factors, information overload and what we term assessment fatigue, influence evaluation behavior and user experience. Assessment fatigue is defined as cognitive strain resulting from repeated evaluative tasks. The study draws on data from a university-level deployment of a structured peer evaluation system. We applied Structural Equation Modeling (SEM) to analyze how behavioral data and self-reported perceptions of overload and fatigue relate to overall system satisfaction. Results reveal a significant indirect pathway from information overload to system satisfaction, mediated by fatigue. Based on these findings, we propose design recommendations for cognitively-aware assessment systems that adapt to students' cognitive constraints, contributing to the development of AI-supported educational tools that are more robust and human-centered.**

I. Introduction

PEER ASSESSMENT is widely used in higher education to enhance learning and promote critical thinking [14, 19]. As educational platforms increasingly integrate automated or semi-automated assessment tools, peer evaluation systems are becoming candidates for AI-supported enhancement [7, 8]. While the design of aggregation methods and elicitation strategies has recently received some attention [18, 21], the *cognitive experience* of the assessors—specifically the effects of overload and fatigue—has been largely overlooked. The cognitive demands placed on student assessors when reviewing multiple peer projects—remain poorly understood. As students progress through a sequence of evaluations, they may experience information overload and fatigue, which can affect the consistency of their ratings and their overall satisfaction with the process. This paper addresses that gap by focusing on two cognitive factors that may degrade the quality of peer evaluations over time: information overload (IOL) and what we term assessment fatigue. There is no single definition of information overload (see e.g., the survey paper by Bawden and Robinson [2]). A commonly accepted definition is that information overload occurs when a user has insufficient cognitive capacity to handle the presented information. IOL has been linked to decreased decision quality and increased inconsistency [9, 11]. These cognitive factors do not only impact the user experience—they may also compromise the accuracy of peer evaluations. Under high cognitive load, students may fail to distinguish between projects, default to repeated ratings, or reduce the depth of evaluative effort, potentially affecting the fairness and validity of assessment outcomes.

We introduce the term assessment fatigue to describe cognitive depletion stemming from repeated evaluative tasks within peer review sessions. While information overload is well established in cognitive psychology and organizational behavior, its role in peer assessment systems remains underexplored. The concept of fatigue in this context has received even less scholarly attention.

We study these phenomena using data collected from the R2R peer assessment platform [8], deployed in a university course on data analysis. In this setting, students assessed peer projects using a structured interface that combined Likert-scale ratings with pairwise comparison queries. Figure 1 illustrates the R2R system flow. Specifically, we ask: (1) What is the impact of information overload on peer assessment? (2) What is the impact of assessment fatigue on peer assessment? (3) How satisfied are users with the R2R peer assessment system? Our goal is to model the relationships between these cognitive factors and evaluation patterns, and to consider how future systems might be designed to account for such effects.

In addition to collecting behavioral data such as score patterns and tie frequency, we administered a post-assessment questionnaire that measured perceived information overload (IOL), assessment fatigue, and satisfaction with the system. To model the relationships among these subjective constructs and evaluation behavior, we applied Structural Equation Modeling (SEM), treating each construct as a latent variable.

We use the results of our behavioral analysis to motivate several design recommendations for improving peer assessment systems. These include incorporating fatigue-sensitive prompts and weighting inputs based on behavioral reliability. The goal is to help future systems better support assessors under cognitive strain and improve the overall consistency and fairness of evaluations. By combining behavioral modeling with system-level implications, this work contributes to the development of human-centered, AI-supported educational

This work was supported by the Ministry of Innovation, Science & Technology, Israel

technologies that account for the cognitive limitations of student evaluators. Specifically, our contributions are threefold:

- We define and estimate assessment fatigue as a latent variable using behavioral and self-report data.
- We apply Structural Equation Modeling (SEM) to analyze the relationships between information overload, assessment fatigue, and peer evaluation behavior.
- We propose design recommendations for peer assessment systems that address cognitive strain during evaluation.

II. Related Work

Peer assessment has been widely used in educational settings to support evaluation and feedback among students. Many systems have been developed to structure this process, aiming to improve reliability, fairness, and ease of use [19, 20, 14]. However, persistent challenges include strategic grading [18], inconsistency among assessors [21], and score inflation [15]. Rating-based systems are vulnerable to calibration issues and leniency bias [8], while systems that rely only on rankings can impose high cognitive demands on students [16].

The R2R platform addresses some of these limitations by combining Likert-scale ratings with pairwise comparisons, triggered only when students assign the same score to multiple projects [8]. This approach collects informative input while keeping the evaluation task manageable. The present study builds on this framework to examine how assessors' cognitive states—specifically overload and fatigue—affect their rating behavior and system satisfaction.

Information overload has been studied in decision-making, education, and organizational behavior [1, 6, 9], but its relevance to peer assessment remains underexplored. We also introduce the concept of assessment fatigue, defined as cognitive strain from repeated evaluation tasks. Although well-studied in cognitive psychology, fatigue has not been explicitly modeled in the peer assessment context. This study examines both constructs and their relationship to behavioral patterns and subjective user experience.

Previous studies in large-scale peer review, such as in MOOCs, have documented variability in rater performance over time [13, 17]. However, such work typically treats inconsistency as noise, rather than exploring its cognitive origins. In contrast, our study models cognitive factors as latent constructs and uses Structural Equation Modeling (SEM) to analyze their connection to evaluation behavior.

Recent work in educational AI explores how systems can adapt to learners' cognitive and emotional states [7], including adaptive pacing [22], feedback support [5], and scaffolding [10]. However, most peer assessment platforms still treat all students as equally capable reviewers. We argue that behavioral indicators of cognitive strain—such as reduced variability, scoring decay, or rushed evaluations—can inform the design of more responsive peer assessment systems.

III. Methodology

The study was conducted in an undergraduate data analysis course at Ariel University. Students participated in a structured peer assessment exercise using the R2R platform [8]. Each student viewed and evaluated 9–12 peer presentations. Evaluations began with assigning 1–5 Likert-scale scores to each project based on predefined rubrics. When multiple projects received the same score, the system triggered pairwise comparison queries to elicit ordinal preferences among them. This hybrid approach—combining cardinal ratings with selective ordinal input—yields a partially ordered set of preferences per student.

Once all data were collected, the system aggregated the individual inputs into a full ranking over all projects. For each median score level, it computed an internal ranking among the tied projects using either Borda or Copeland voting rules. The final output was a global project ranking that preserved both score-based judgments and fine-grained pairwise distinctions [8].

To capture students' evaluation behavior over time, we computed a set of behavioral indicators derived from their interactions with the R2R system. These included the mean score each student assigned across all evaluated projects, as well as a decay coefficient—defined as the difference between the mean of the first third and the last third of scores in the session—to estimate whether students' ratings declined over time. We also recorded the proportion of ties (i.e., repeated scores across projects). Another variable tested is response latency—measured as the average time between rating actions—was also extracted as a behavioral indicator. However, it did not show a significant association with any of the cognitive variables (overload, engagement, or satisfaction), and was therefore excluded from the final model.

Upon completing their peer assessments, students filled out a questionnaire including validated items on information overload, perceived fatigue, and system usability. Responses were measured using a five-point Likert scale and aggregated into three latent constructs: Information Overload, Assessment Fatigue, and Satisfaction. The full questionnaire can be found in the Appendix VI.

A total of 231 students participated in the study. Some did not rate all projects in their session, or failed to complete the questionnaire. We were thus left with a total of 194 students.

We employed Structural Equation Modeling (SEM). The model was pre-specified based on the theoretical assumptions on the relationships between cognitive factors (fatigue and overload), behavioral patterns, and subjective experience (satisfaction with the system). The SEM was implemented in R using the `lavaan` package, with maximum likelihood estimation. Model fit was assessed using standard indices: CFI, RMSEA, and SRMR. The latent variables were modeled as predictors of observed behavioral features to understand how cognitive strain manifests in evaluation patterns.

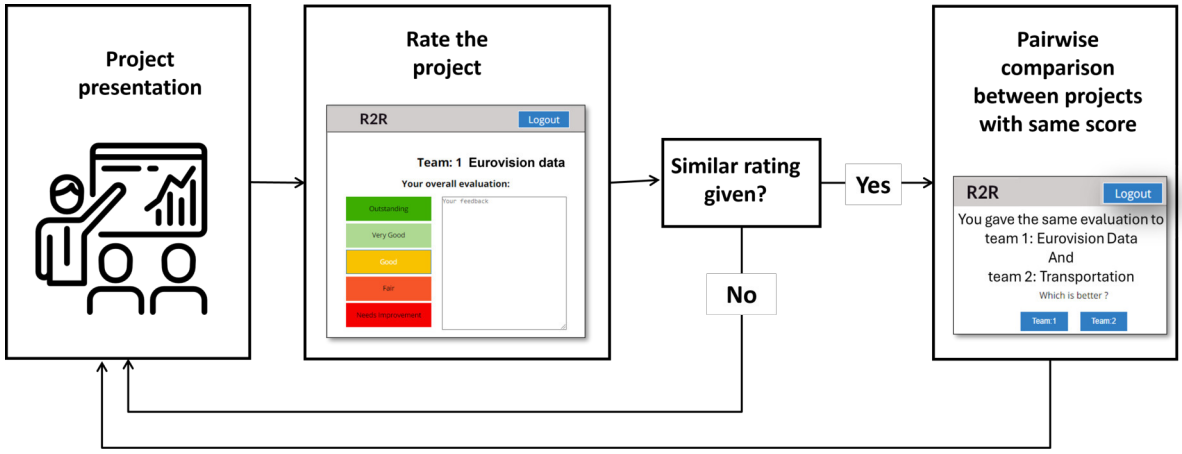

Fig. 1. R2R system flow. Students view project presentations and then rate the projects. If a student assigns the same rating to two or more projects, a series of pairwise comparison queries is executed.

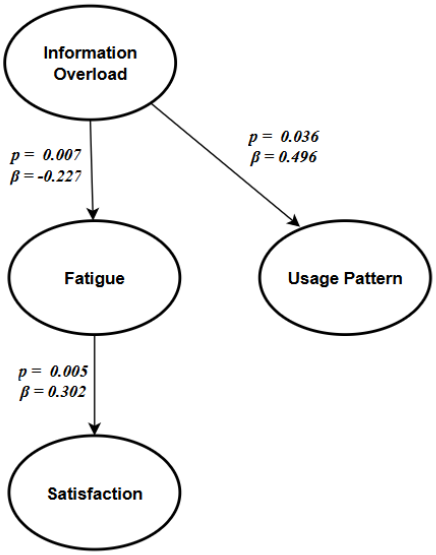

Fig. 2. Structural model results. Information overload predicts both assessment fatigue and usage patterns. Fatigue in turn predicts system satisfaction. A significant indirect effect from overload to satisfaction is mediated through fatigue.

IV. Results

The SEM demonstrated an acceptable to good fit with the data. The scaled chi-square test was marginally significant, $\chi^2(112) = 139.24$, $p = .041$. Additional fit indices supported the model: CFI = 0.934, TLI = 0.921, RMSEA = 0.035, and SRMR = 0.063.

Figure 2 presents a simplified diagram of the structural paths estimated in the model. The most robust finding was a positive and significant path from the latent fatigue construct to system satisfaction ($\beta = 0.302$, $p = .005$). Although labeled as fatigue, the items comprising this factor reflect cognitive engagement (e.g., active reflection and effort). Thus, higher scores indicate lower fatigue or greater evaluative involvement. Interpreted accordingly, the result suggests that students who were more cognitively engaged during the peer assessment process reported higher satisfaction with the system—a pattern consistent with prior findings on self-regulated learning and intrinsic motivation.

Information overload significantly predicted assessment fatigue ($\beta = -0.227$, $p = .007$), indicating that students who reported higher overload tended to also report greater fatigue. However, the direct path from information overload to system satisfaction was not statistically significant ($\beta = 0.012$, $p = .912$), suggesting that the relationship between overload and satisfaction operates indirectly, through its effect on fatigue.

The significance of the indirect path from information overload to satisfaction through fatigue ($ab_3 = -0.048$, $p = .019$) supports a partial mediation model. Fatigue acts as a conduit through which perceived overload affects downstream user experience. This finding aligns with cognitive theories that emphasize the cumulative toll of high information load and the importance of affective and physical strain in shaping task satisfaction [9, 1].

In the context of peer assessment, these mechanisms may help explain how students' evaluative consistency and system satisfaction decline as cognitive demands accumulate over the course of repeated assessment tasks.

In addition to the latent variables, the model incorporated behavioral measures. Information overload significantly predicted the mean rubric usage pattern ($ab_3 = 0.496$, $p = .036$), indicating that cognitive strain manifests not only in subjective reports but also in measurable changes in evaluation behavior. Students experiencing higher overload engaged differently with the interface, suggesting that behavior traces may offer real-time or asynchronous indicators of cognitive state.

Several additional paths were included in the model for completeness but were not statistically significant. These included the direct path from information overload to satisfaction

($ab_3 = 0.012$, $p = .912$), as well as paths from behavioral pattern indicators to satisfaction ($ab_3 = -0.053$, $p = .548$) and fatigue ($ab_3 = -0.070$, $p = .315$).

In summary, the model confirms that perceived information overload predicts fatigue, and that fatigue, in turn, positively predicts system satisfaction. The direct path from overload to satisfaction was not significant, indicating that fatigue mediates this relationship. These findings answer our core research questions and highlight the value of modeling cognitive constructs such as overload and fatigue as latent variables. The use of SEM allowed us to capture latent constructs that are not directly observable, such as perceived fatigue and overload, and to examine their relationships with both self-reported satisfaction and system-logged behavioral patterns. This modeling approach made it possible to detect indirect effects and account for measurement error, offering a more accurate view of how cognitive states shape peer assessment experience.

V. Design Recommendations

Assessment fatigue, i.e., cognitive fatigue during peer assessment, can be addressed through real-time detection mechanisms that rely on unobtrusive behavioral signals. Because the assessment environment already logs granular student actions—such as response times, ratings, and tie-breaking decisions—it is feasible to infer signs of strain or disengagement as students work through their evaluation tasks. Specifically, indicators such as repeated use of a single score (e.g., consistently assigning the highest rating), low variance across rubric dimensions, skipping optional comment fields, or evidence of scoring decay over time may suggest cognitive overload or reduced attentional engagement. The latter can be quantified using a decay coefficient, defined as the difference in mean scores between the first and last segments of a student's evaluation sequence. A consistent downward trend in scores, relative to how other assessors rated the same projects, may reflect increasing fatigue or diminished evaluative attention as the session progresses.

When such patterns are detected in real time, the system can scaffold the rating process to reduce cognitive load. Instead of allowing students to proceed with rapid, unreflective scoring, the interface can present brief prompts encouraging deeper engagement. For example, when repeated use of the same score is detected, the system might ask: "Consider pacing your next few ratings more slowly to ensure accuracy" or "Before submitting, consider whether this score reflects the project's clarity, originality, and completeness." Other prompts might invite rubric-based reflection, such as: "Did this project meet expectations for structure, argumentation, or use of data?". These light-weight nudges are designed to trigger evaluative reappraisal without interrupting workflow. Drawing on principles of adaptive scaffolding and self-regulated learning, such prompts can help students maintain attentional focus and calibrate their ratings more accurately, even under conditions of assessment fatigue.

In addition to in-the-moment interventions, cognitive strain can be addressed retrospectively through post-hoc adjustments to aggregation. After the evaluation session concludes, the system can compute a reliability score for each assessor based on fatigue-related behavioral signals. This score, scaled between 0.1 and 1.0, reflects the inferred trustworthiness of the student's inputs. These weights can then be used to adjust both rating-based and ranking-based aggregation methods. In the case of ratings, a student who assigns the same score to all projects or shows minimal rubric variance may have their evaluations down-weighted in the final average. For ranking-based aggregation, methods such as Borda or Copeland can be adapted to incorporate these weights directly. The adjusted Borda score for a project c_j would be computed as:

$$\text{AdjustedBorda}(c_j) = \sum_{i=1}^{n} w_i \cdot \text{BordaRank}_i(c_j)$$

where w_i is the reliability weight for student i, and $\text{BordaRank}_i(c_j)$ is the rank assigned by that student to project c_j. This post-hoc adjustment strategy requires no change to the frontend interface or assessment process.

VI. Discussion

Real-time scaffolding and reliability-based weighting support student wellbeing by recognizing and mitigating the cognitive toll of extended evaluation tasks. These adaptive mechanisms may also enhance fairness, by reducing the influence of assessors operating under strain. Thus, integrating fatigue-sensitive design into peer assessment systems carries not only cognitive and pedagogical value, but also reflects human-centered AI principles—namely, respect for cognitive limits, equitable treatment, and minimal intrusion.

Several directions remain open for extending this work. One promising avenue is the integration of large language models (LLMs) to enhance feedback analysis and detect signs of fatigue or disengagement through natural language patterns. While the present study focuses on rating behavior and assessment structure, written feedback can offer a complementary signal of cognitive strain. For example, LLMs can be used to analyze comment length, sentence count, lexical richness, or syntactic complexity. Qualitative shifts such as sentiment flattening ("nice job") or use of generic, non-specific phrases ("good project," "well done") may indicate reduced engagement or cognitive effort. Detecting such patterns at scale could enable a multi-modal fatigue detection pipeline that combines behavioral and textual indicators.

A second line of future work involves implementing the design recommendations proposed in this paper within a live peer assessment platform. This includes integrating fatigue-sensitive prompts, real-time scaffolding for ratings, and post-hoc reliability-weighted aggregation. Controlled experiments can then evaluate the effectiveness of these interventions in improving evaluation quality and fairness of outcomes.

ACKNOWLEDGMENT

This research was supported by the Ministry of Innovation, Science & Technology, Israel.

REFERENCES

[1] Roy F. Baumeister, Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, 74(5):1252–1265, 1998. https://doi.org/10.1037/0022-3514.74.5.1252.

[2] David Bawden and Lyn Robinson. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science*, 35(2):180–191, 2009. https://doi.org/10.1177/0165551508095781.

[3] Gee-Woo Bock, Mimrah Mahmood, Sanjeev Sharma, and Youn Jung Kang. The impact of information overload and contribution overload on continued usage of electronic knowledge repositories. *Journal of Organizational Computing and Electronic Commerce*, 20(3):257–278, 2010. https://doi.org/10.1080/10919392.2010.494530.

[4] Chao-hsiu Chen. The implementation and evaluation of a mobile self-and peer-assessment system. *Computers & Education*, 55(1):229–236, 2010. https://doi.org/10.1016/j.compedu.2010.01.008.

[5] Olga Chernikova, Daniel Sommerhoff, Matthias Stadler, Doris Holzberger, Michael Nickl, Tina Seidel, Enkelejda Kasneci, Stefan Küchemann, Jochen Kuhn, Frank Fischer, et al. Personalization through adaptivity or adaptability? a meta-analysis on simulation-based learning in higher education. *Educational Research Review*, page 100662, 2024. https://doi.org/10.1016/j.edurev.2024.100662.

[6] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011. https://doi.org/10.1073/pnas.1018033108.

[7] Amir Darvishi, Hadi Khosravi, Shazia Sadiq, and Dragan Gašević. Incorporating ai and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, 2022. https://doi.org/10.1111/bjet.13233.

[8] Lihi Dery. Interactive and iterative peer assessment. In *Proceedings of the European Conference on Artificial Intelligence (ECAI), https://doi.org/10.3233/FAIA240656*, 2024.

[9] Martin J. Eppler and Jeanne Mengis. The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society*, 20(5):325–344, 2004. https://doi.org/10.1080/01972240490507974.

[10] Frank Fischer, Elisabeth Bauer, Tina Seidel, Ralf Schmidmaier, Anika Radkowitsch, Birgit J Neuhaus, Sarah I Hofer, Daniel Sommerhoff, Stefan Ufer, Jochen Kuhn, et al. Representational scaffolding in digital simulations–learning professional practices in higher education. *Information and Learning Sciences*, 123(11/12):645–665, 2022. https://doi.org/10.1108/ILS-06-2022-0076.

[11] Janusz A Hołyst, Philipp Mayr, Michael Thelwall, Ingo Frommholz, Shlomo Havlin, Alon Sela, Yoed N Kenett, Denis Helic, Aljoša Rehar, Sebastijan R Maček, et al. Protect our environment from information overload. *Nature Human Behaviour*, 8(3):402–403, 2024. https://doi.org/10.1038/s41562-024-01833-8.

[12] Felix Krieglstein, Maik Beege, Günter Daniel Rey, Christina Sanchez-Stockhammer, and Sascha Schneider. Development and validation of a theory-based questionnaire to measure different types of cognitive load. *Educational Psychology Review*, 35(1):9, 2023. https://doi.org/10.1007/s10648-023-09738-0.

[13] Chinmay Kulkarni, Kristine P. Wei, Harini Le, Daniel Chia, Katharine Papadopoulos, Justin Cheng, Daphne Koller, and Scott R. Klemmer. Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6):1–31, 2013. https://doi.org/10.1145/2505057.

[14] H. Li, Y. Xiong, C. V. Hunter, X. Guo, and R. Tywoniw. Does peer assessment promote student learning? a meta-analysis. *Assessment & Evaluation in Higher Education*, 45(2):193–211, 2020. https://doi.org/10.1080/02602938.2019.1620679.

[15] Ernesto Panadero, M. Romero, and J.-W. Strijbos. The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, 39(4):195–203, 2013. https://doi.org/10.1016/j.stueduc.2013.10.005.

[16] Kumar Raman and Thorsten Joachims. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1037–1046, 2014. https://doi.org/10.1145/2623330.2623654.

[17] Nihar B. Shah, James K. Bradley, Ameet Parekh, Martin J. Wainwright, and Kannan Ramchandran. A case for ordinal peer evaluation in moocs. In *NIPS Workshop on Data Driven Education*, 2013.

[18] Ivan Stelmakh, Nihar B. Shah, and Aarti Singh. Catch me if i can: Detecting strategic behaviour in peer assessment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4794–4802, 2021. https://doi.org/10.1609/aaai.v35i6.16611.

[19] Keith Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.

[20] Keith J. Topping. Peer assessment. *Theory Into Practice*, 48(1):20–27, 2009.

[21] Toby Walsh. The peerrank method for peer assessment. In *ECAI*, pages 909–914, 2014. https://doi.org/10.3233/978-1-61499-419-0-909.

[22] Bahman Zohuri and Farhang Mossavar-Rahmani. Revolutionizing education: The dynamic synergy of personalized learning and artificial intelligence.

International Journal of Advanced Engineering and Management Research, 9(1):143–153, 2024. https://doi.org/10.51505/ijaemr.2024.9111.

APPENDIX: QUESTIONNAIRE

Students completed a questionnaire immediately after finishing the peer assessment tasks. It included two demographic questions, six items from a validated satisfaction scale [4], four items from an information overload scale [3], and five items on fatigue adapted from Krieglstein et al. [12].

Response Scales:

- **5-point scale:** 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, 5 = Strongly Agree
- **7-point scale:** 1 = Strongly Disagree, 2 = Disagree, 3 = Somewhat Disagree, 4 = Neither Agree nor Disagree, 5 = Somewhat Agree, 6 = Agree, 7 = Strongly Agree

General Information

- Age
- Gender

Satisfaction [4] *(5-point scale)*

- R2R was more efficient for peer and self-evaluation than paper-and-pencil methods would be.
- The operation of R2R evaluations was easy and convenient.
- I tried to grade my peers appropriately and not to be picky or hypocritical.
- I worried that my classmates would give me low grades intentionally.
- I think the instructor's comments and grading would be more professional than my classmates'.
- I felt that I received biased low scores from peers.

Information Overload [3] *(7-point scale)*

- I cannot effectively process all the information I read.
- I am overwhelmed with the information I read.
- I feel anxious that I might have missed an important piece of information I read.
- I cannot assimilate all the information I read.

Fatigue [12] *(7-point scale)*

- The learning content was difficult to understand.
- The learning content included much complex information.
- I actively reflected upon the learning content.
- I made an effort to understand the learning content.
- I was able to expand my prior knowledge with the learning content.