

## Jeder bewertet jeden: Erfolgsquote einer Peer Assessment Methode in einem Serious Game zum Lernen

Linda Eckardt<sup>1</sup>, Adam Jankowiak<sup>1</sup>, Dennis Bergelt<sup>2</sup> und Susanne Robra-Bissantz<sup>1</sup>

**Abstract:** Peer Assessment entlastet Lehrende und bietet Lernenden eine schnelle Rückmeldung während des Lernprozesses. Der Erfolg von Peer Assessment Methoden wurde bislang jedoch noch nicht umfangreich genug betrachtet. Daher wird in diesem Beitrag die Erfolgsquote eines implementierten Peer Assessments in einer digitalen Game-based Learning Anwendung untersucht. Die Ergebnisse der Untersuchung zeigen, dass nur ein Drittel der insgesamt 4963 Reviews korrekt sind, wobei die bewerteten Abgaben in Textform häufiger richtig bewertet wurden, als die Abgaben per Datei. Insgesamt neigen Lernende demnach dazu bessere Bewertungen abzugeben, als verdient.

**Keywords:** Game-based Learning, Serious Game, Peer Assessment, Erfolgsquote

### 1 Einleitung

Der Einsatz von Spielelementen beim Lernen ist ein aktueller Trend und an vielen Universitäten mittlerweile verbreitet. Game-based Learning Anwendungen, die innerhalb der Lehre Spielelemente zur Steigerung der Motivation sowie des Lerneffekts einsetzen, eröffnen vielfältige Möglichkeiten, um die Lehre flexibler zu gestalten [HKS14]. Diese lassen sich mit der Methode des Peer Assessments vereinbaren, in der die Peers (hier: die Lernenden) durch Reviews ihr erworbenes Wissen anwenden und für eine exaktere Bewertung nutzen [FG00]. Bislang gibt es jedoch wenig Studien, die den Erfolg eingesetzter Peer Assessment Methoden evaluieren. Kulkarni et al. zeigen in einem iterativen Vorgehen auf, dass durch Peer Assessment die Bewertungen der Teilnehmenden sehr stark mit denen der Lehrenden zugewiesenen Bewertungen korrelieren [Ku13]. Im Durchschnitt bewerten sich die Teilnehmenden jedoch 7 % besser als die Lehrenden. Eine weitere Studie hat gezeigt, dass der Einsatz einer Peer Assessment Methode eine zeit- und ressourcenschonende Möglichkeit für den Lehrenden ist, um die Lernenden durch Feedback in ihrem Lernprozess zu unterstützen [Le15].

In dieser Studie werden hingegen Datensätze aus einer universitären Game-based Learning Lehrveranstaltung ausgewertet. Dazu wird im Rahmen dieser Studie die Erfolgsquote der Peer Assessment Methode untersucht, bei der einerseits Antworten per Texteingabe und andererseits Abgaben per Datei ausgewertet werden.

---

<sup>1</sup> TU Braunschweig, Institut für Wirtschaftsinformatik, Abt. Informationsmanagement, Mühlenpfordtstr. 23, 38106 Braunschweig, {linda.eckardt|a.jankowiak|s.robra-bissantz}@tu-bs.de

<sup>2</sup> TU Braunschweig, Institut für Wirtschaftsinformatik, Abt. Informationsmanagement, Mühlenpfordtstr. 23, 38106 Braunschweig, dennis.bergelt@t-online.de

## 2 Peer Assessment im Serious Game

Serious Game ist eine Form des Game-based Learning, wobei die Entwicklung eines vollwertigen Spiels mit festen Regeln und Zielen in einem spielfremden Kontext, z. B. der Lehre, gemeint ist [De11]. Das in dieser Studie verwendete Serious Game „Lost in Antarctica“, ist ein entwickeltes Open Source Browserspiel zum Erlernen von Informationskompetenz. In zwölf Levels haben Studierende innerhalb einer fiktiven Forschungs Expedition am Südpol die Möglichkeit verschiedene Teilgebiete der Informationskompetenz (z. B. Recherchestrategien und Literaturverwaltung) spielerisch zu lernen [ER16].

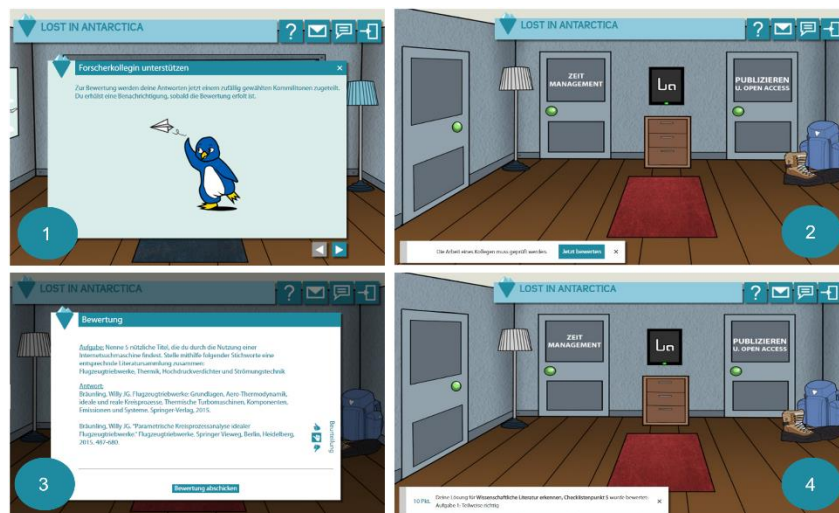


Abb. 1: Bewertung

In dem Serious Game ist ein Peer Assessment in Form eines Peer Ratings implementiert, d. h. Spielende werden anonym von Mitspielenden durch die Vergabe von Punkten bewertet [Ke05]. Dabei können nur diejenigen Studierenden abgegebene Antworten bewerten, die die Aufgabe selbst bereits erfolgreich gelöst haben. Sobald Studierende eine Aufgabe abschließen, die mittels Peer Assessment bewertet wird, erscheint eine Meldung (Screen 1 der Abb. 1). Über neu zu bewertende Abgaben informiert ein im Spiel auftauchender Hinweis (Screen 2 der Abb. 1). Diesen Hinweis kann der Reviewende ausführen oder das Fenster schließen und somit die Bewertung ablehnen. Der Reviewende bekommt die Antworten in einem Fenster zu sehen und soll diese dann anhand seiner Vorkenntnisse bewerten (Screen 3 der Abb. 1). Dabei entspricht der „Daumen nach oben“ und der „mittlere Daumen“ einer positiven Bewertung und der „Daumen nach unten“ hat eine Wiederholung der Aufgabe zur Folge. Nach abgeschlossener Bewertung erhalten die Studierenden eine Rückmeldung über die Ergebnisse (Screen 4 der Abb. 1).

### 3 Erfolgsquote des Peer Assessments

#### 3.1 Design der Untersuchung

Insgesamt wurden 4963 Reviews ausgewertet, wobei diese Lösungen zu Hauptaufgaben und zugehörigen Teilaufgaben umfassen. Überprüft wurde die Richtigkeit der durchgeführten Reviews (1 = „Daumen nach oben“; 0,5 = „mittlerer Daumen“; 0 = „Daumen nach unten“). Eine mittlere Bewertung bedeutet, dass mindestens 50 % korrekt sind. Das durchgeführte Review wird in vorher festgelegte Kategorien eingeordnet (Tab. 1). Dabei bestehen die Kategorien aus der Güte des Inhalts der Abgabe, die durch die Untersuchung festgestellt wurde und der verfassten Bewertung der reviewenden Personen. Beispielsweise entspricht ein „Pfeil nach oben“ einer richtigen Antwort. Die Bewertung gilt als erfolgreich abgeschlossen, wenn das Review in die Kategorien 1, 5 oder 9 fällt. Eine erfolglose Bewertung liegt bei den übrigen Kategorien vor.

| Kategorie | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-----------|---|---|---|---|---|---|---|---|---|
| Abgabe    | ↑ | ↑ | ↑ | ⇒ | ⇒ | ⇒ | ↓ | ↓ | ↓ |
| Bewertung | ↑ | ⇒ | ↓ | ↑ | ⇒ | ↓ | ↑ | ⇒ | ↓ |

Tab. 1: Kategorien für die Bewertung

#### 3.2 Ergebnisse der Untersuchung

Insgesamt haben 107 Studierende (87 Männer und 20 Frauen) mit einem Durchschnittsalter von 22 Jahren mit dem Peer Assessment gearbeitet. Tab. 2 veranschaulicht die Gesamtauswertung des Peer Assessments. 33,02 % der abgegebenen Reviews sind korrekt, d. h. die Mehrheit der Bewertungen durch die Studierenden ist fehlerhaft. Dies bestätigt auch die Bestehensquote, nach der 92,85 % der Antworten gemäß der Reviews zum Bestehen reichen, die Untersuchung jedoch gezeigt hat, dass nur 43,06 % ausreichend sind. Somit sind mehr als die Hälfte aller Abgaben nicht ausreichend und hätten zu einer Wiederholung führen müssen. In Kategorie 7 fallen die meisten Reviews. Demnach haben sehr viele Studierende falsche Antworten als richtig bewertet. Zu strenge Bewertungen (Kategorie 2, 3 und 6) sind am wenigstens vertreten. Mittlere Bewertungen (Kategorie 5) sind ebenfalls kaum vorhanden, d. h. die Studierenden bevorzugen die Bewertung mit Tendenz in positiver oder negativer Richtung.

| Kategorie | 1    | 2  | 3  | 4   | 5  | 6  | 7    | 8   | 9   | Summe |
|-----------|------|----|----|-----|----|----|------|-----|-----|-------|
| Reviews   | 1244 | 43 | 26 | 703 | 95 | 26 | 2339 | 187 | 300 | 4963  |

| Bewertung | Reviews | Untersuchung |
|-----------|---------|--------------|
| 1         | 4286    | 1312         |
| 0,5       | 322     | 825          |
| 0         | 355     | 2826         |

| Bestehensquote | Reviews          | Untersuchung |
|----------------|------------------|--------------|
|                | 92,85 %          | 43,06 %      |
| Erfolgsquote   | Misserfolgsquote |              |
| 33,02 %        | 66,98 %          |              |

Tab. 2: Gesamtauswertung der Peer Assessment Methode

Nachfolgend werden die Aufgabentypen §Abgabe per Datei (Tab. 3) und §Abgabe per Texteingabe (Tab. 4) vergleichend gegenübergestellt.

| Kategorie | 1   | 2 | 3 | 4   | 5 | 6 | 7    | 8  | 9  | Summe |
|-----------|-----|---|---|-----|---|---|------|----|----|-------|
| Reviews   | 223 | 2 | 2 | 125 | 9 | 0 | 1232 | 69 | 81 | 1743  |

| Bewertung | Reviews | Untersuchung |
|-----------|---------|--------------|
| 1         | 1580    | 227          |
| 0,5       | 79      | 134          |
| 0         | 84      | 1384         |

| Bestehensquote | Reviews          | Untersuchung |
|----------------|------------------|--------------|
|                | 95,18 %          | 20,71 %      |
| Erfolgsquote   | Misserfolgsquote |              |
| 17,96 %        | 82,0 %           |              |

Tab. 3: Auswertung der Abgabe per Datei

| Kategorie | 1    | 2  | 3  | 4   | 5  | 6  | 7    | 8   | 9   | Summe |
|-----------|------|----|----|-----|----|----|------|-----|-----|-------|
| Reviews   | 1021 | 41 | 24 | 578 | 86 | 26 | 1107 | 118 | 219 | 3220  |

| Bewertung | Reviews | Untersuchung |
|-----------|---------|--------------|
| 1         | 2706    | 1085         |
| 0,5       | 243     | 691          |
| 0         | 271     | 1444         |

| Bestehensquote | Reviews          | Untersuchung |
|----------------|------------------|--------------|
|                | 91,58 %          | 55,16 %      |
| Erfolgsquote   | Misserfolgsquote |              |
| 41,18 %        | 58,82 %          |              |

Tab. 4: Auswertung der Abgabe per Texteingabe

Von 1743 bewerteten Dateiabgaben wurden nur 17,96 % korrekt bewertet, wodurch die Bestehensquote aus der Untersuchung nur 20,71 % ergab. Die meisten Abgaben wurden fälschlicherweise als richtig bewertet (Kategorie 7). Mit 41,18 % richtig durchgeführter Reviews ist die Erfolgsquote bei 3220 bewerteten Abgaben per Texteingabe deutlich höher als bei der Abgabe per Datei. Die Bestehensquote aus der Untersuchung ist mit 55,16 % dadurch ebenfalls höher als bei der Abgabe per Datei, in Anbetracht der höheren Anzahl an Reviews sogar deutlich höher. Viele falsche Lösungen wurden als richtig bewertet (Kategorie 7) und viele richtige Lösungen wurden auch als richtig bewertet (Kategorie 1). Möglich ist, dass Studierende nicht in die Dateiabgaben hineinschauen und

daher falsch bewerten. Dahingegen werden Textabgaben eher gelesen und richtig bewertet, weil sie unmittelbar angezeigt werden.

Abb. 2 zeigt die Erfolgsquoten der einzelnen Reviewer (a) und die Anzahl verfasster Reviews pro Person (b). Erkennbar ist, dass die Erfolgsquoten der Reviewer deutlich verschieden sind. Während sehr wenige Reviewer alle Bewertungen falsch vorgenommen haben, haben auch nur wenige Reviewer mehr als 75 % korrekt bewertet. Die Mehrheit bewegt sich im Mittelfeld und hat zwischen 30 % und 60 % richtig beurteilt. Die Anzahl der abgegebenen Reviews pro Person schwankt ebenfalls stark. Folglich ist die Intensität, mit der die Auseinandersetzung mit der Peer Assessment Methode erfolgt ist, unterschiedlich. Beispielsweise haben drei Reviewer jeweils nur eine Bewertung abgegeben, wohingegen vier Reviewer jeweils über 100 Beurteilungen vorgenommen haben. Die hohe Anzahl an falschen Bewertungen lässt darauf schließen, dass die Seriosität, mit der die Studierenden die Reviews formuliert haben, zu gering war. Die Studierenden bewerten oftmals falsche Angaben oder sogar themenfremde abgegebene Dateien mit teilweise richtig oder richtig.

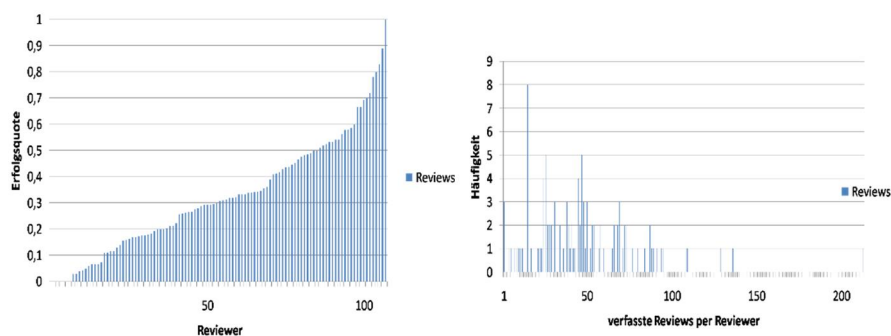


Abb. 2: Erfolgsquoten der Reviewer (a) und Anzahl verfasster Reviews pro Reviewer (b)

## 4 Zusammenfassung und Ausblick

Die Untersuchung des Erfolgs der Peer Assessment Methode hat gezeigt, dass nur ein Drittel der durchgeführten Reviews korrekt sind. Die Studierenden neigen demnach dazu bessere Bewertungen abzugeben, als verdient [WC86]. Zudem weichen die Erfolgsquoten und die Intensität der verfassten Reviews zwischen den einzelnen Reviewern stark voneinander ab. Die Abgabe per Textfeld erreichte mit 41,18 % eine bessere Erfolgsquote, als die Abgabe per Datei mit einer Quote von lediglich 17,96 %. Folglich eignet sich die Abgabe einer Lösung in Textform womöglich besser innerhalb einer Peer Assessment Methode, die online durchgeführt wird.

Eine unterschiedliche Auffassung der Bewertungskriterien zwischen den Studierenden und der Kontrolluntersuchung ist kritisch zu betrachten, lässt sich bei der Auswertung

offener Fragen jedoch nicht gänzlich vermeiden. In weiterführenden Studien kann beispielsweise durch eine Unterteilung in einen vorangestellten Wissenstest, der eigentlichen Verwendung der Anwendung sowie einem abschließenden Wissenstest untersucht werden, ob eine intensivere Auseinandersetzung mit dem Peer Assessment den Wissensgewinn positiv beeinflusst [WHC16]. Auch ist eine Untersuchung der Bewertungsqualitäten der Reviewer denkbar oder eine Unterscheidung zwischen Männern und Frauen. Studierende haben wenige mittlere Bewertungen vergeben. Auch dies könnte in weiterführenden Studien vertieft betrachtet werden.

## Literaturverzeichnis

- [De11] Deterding, S. et al.: From game design elements to gamefulness: defining gamification. In (Lugmayr, A.; Franssila, H.; Safran, C.; Hammouda, I. Hrsg.): Proc. of the 15th Int. academic MindTrek Conf.: Envisioning future media environments. ACM, Tampere, S. 9-15, 2011.
- [ER16] Eckardt, L.; Robra-Bissantz, S.: Design eines Spiels zum Lernen von Informationskompetenz. In (Lucke, U.; Schwill, A.; Zender, R. Hrsg.): Proceedings der DeLFI, LNI, Potsdam, S. 95-106, 2016.
- [FG00] Falchikov, N.; Goldfinch, J.: Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research* 70/3, S. 287-322, 2000.
- [HKS14] Hamari, J.; Koivisto, J.; Harri S.: Does gamification work? a literature review of empirical studies on gamification. In: Proc. of the 47th Hawaii Int. Conf. on Information Systems, IEEE, Waikoloa, S. 3025-3034, 2014.
- [Ke05] Kennedy, G. J.: Peer-assessment in group projects: is it worth it? In (Young, A.; Tolhurst, D. Hrsg.): Proc. of the 7th Australasian Conf. on Computing Education, ACM, Newcastle, S. 59-65, 2005.
- [Ku13] Kulkarni, C. et al.: Peer and Self Assessment in Massive Online Classes. *Transactions on Computer-Human Interaction* 9/39, S. 131- 168, 2013.
- [Le15] Lehmann, K. et al.: Der Wert von IT-gestütztem Peer Assessment zur Unterstützung des Lernens in einer universitären Massenlehrveranstaltung. In (Thomas, O.; Teuteberg, F. Hrsg.): Proc. der 12. Int. Tagung Wirtschaftsinformatik, Osnabrück, S. 1694-1709, 2015.
- [WC86] Weaver, R. L.; Cotrell, H. W.: Peer evaluation: A case study. *Innovative Higher Education* 11/1, S. 25-39, 1986.
- [WHC16] Wang, H. C.; Hsieh, Y. J.; Chen, W. F.: The Effect of Online Peer Assessment in Engineering Education: A Quasi-experimental Study. *Engineering Education* 32/1, S. 199-208, 2016.