Comparing Quantity vs. Quality of Assignment Feedback in Predicting Student Improvement in English Writing Skills

学習者への教育的なフィードバック効果：質的フィードバック及び量的フィードバックにおける差異～英作文授業の観点から

Alexander KRIEG　Fumihito “Andy” NAKAJIMA

アレクザンダー クリーグ　　中嶋アンディ史人

投稿日：2021 年 5 月16 日
受理日：2021 年 11 月19 日

グローバル・コミュニケーション学部

(Abstract)

The current study innovatively combines psychological and educational theories with an experimental design and modern grading technology to provide guidelines for instructors to maximize the feedback they give to their students. We sought to identify underlying behavior principles to maximize skill-based learning in the classroom. To this end, we examined the function of feedback frequency and feedback quality on student improvement in English-language writing skills. Our results demonstrated that both feedback quality and feedback quantity impact improvement in writing skills, with a cumulative effect of feedback quantity over time. The potential benefit of computer scoring systems was also highlighted by the results of the current study. By determining the overall and relative effectiveness of the feedback strategies, we hope that the current results can inform guidelines as to how instructors can maximize feedback to increase student skill development—even with limited time.

(要約)

この研究は､心理学及び教育理論を元に最新のコンピューター自動評価システムを使用することによって、教員がいかに効果的にフィードバックを学生に与え、英文ライティングのスキル向上に寄与できたかを示したものである｡ライティングの課題における質的･量的フィードバックの効果を測定したところ、質と量の両方がスキル向上に有益ではあるが、効果の持続度を考慮に入れると、量的フィードバックがより効果的であるという結果を得た。コンピューター自動評価システムが作成するフィードバックを採用することにより、学生の英文ライティングスキルの向上が見られた結果を踏まえ、教員の限られた時間でもライティングスキルの習得をより効果的に支援することが可能であるという結論に達した。

Key Words: academic writing, automatic grading system, effective feedback, Japanese University English Education

キーワード：アカデミック・ライティング、自動評価システム、効果的フィードバック、日本の大学における英語教育

INTRODUCTION

Within the field of psychology, there are numerous theories that attempt to explain the conditions under which people learn and develop new behaviors. Among these theories, none have been as successful at predicting behavior and behavior change as B.F. Skinner's Theory of Behaviorism (Baum, 2017; Skinner, 1953) and the subsequent science of Behavior Analysis (Austin & Carr, 2000; Skinner, 1938) that stemmed from it. Across a wide variety of species, settings, timeframes, and conditions, an organism's behavior in context can be both predicted to a high degree of accuracy and effectively changed using behavioral principles. Since the advent of this theoretical stance, applications of behavioral science have influenced the fields of biology, economics, applied psychology, anthropology, business, and education (Baum, 2017). Within the context of education, students are expected not only to grow their knowledge base about a given subject, but also improve in a variety of behavioral skillsets endemic to their academic field or trade. The current study seeks to identify underlying behavior principles to maximize skill-based learning in the classroom. To this end, we examined the function of feedback frequency and feedback quality on student improvement in English-language writing skills.

Overview of Behavioral Principles

Claiming academic heritage from Charles Darwin's Theory of Evolution, Behavior Analysis understands behavior as the outcome of the history of iterative interactions between an organism and its environment (Hayes, Sanford, & Chin, 2017). Consequently, Skinner identified two mechanisms by which behavior could influenced: (1.) changing the preceding events or environmental context (i.e., the antecedent), and (2.) changing the outcome of the behavior (i.e., the consequence). Some behaviors are highly unlikely or even impossible to perform in some antecedent contexts (e.g., singing is impossible underwater), while other behaviors are highly likely to occur in other contexts (e.g., greeting a friend during an unplanned encounter in the hallway). By manipulating the antecedent condition of a behavior, the behavior's likelihood and frequency subsequently changes. Outcomes of behavior (i.e., consequences) influence future behavior in a similar way as antecedents. When a given behavior is followed by an appetitive outcome (e.g., positive reinforcement) or the reduction/removal of an aversive outcome (e.g., negative reinforcement) the likelihood or frequency of the behavior increases. When a given behavior is followed by an aversive outcome (e.g., positive punishment) or the reduction/removal of an appetitive outcome (e.g., negative punishment) the likelihood or frequency of the behavior decreases (Skinner, 1938; 1963). Over and above the type of manipulation of behavioral antecedents and consequences, *how* these contingencies are manipulated is paramount to shaping new and complex behavior (Austin & Carr, 2000; Ferster & Skinner, 1957). In particular, three principles guiding the application of behavioral consequences (both reinforcement and punishment) to effectively change behavior have emerged (Thompson & Iwata, 2005). (1.) Consequences must

be contingent upon the target behavior. When consequences are applied to a wide variety of behaviors rather than the target behavior specifically, it takes longer to learn the association between the target behavior and their behavioral consequences. (2.) Consequences must be applied consistently. If consequences are sporadic, it takes longer for the target behavior to change initially. Finally, (3.) consequences must be immediate. A time delay between the behavior and the application of consequences weakens the strength of the reinforcement or punishment. Through effective management of antecedents and consequences, novel, complex behavior, including skill acquisition and skill improvement increases.

Adoption of Behaviorism in Education

Education can be thought of as a behavioral endeavor. Instructors create a set of antecedents and environmental conditions that increase the likelihood and frequency of desired classroom and skill-based behaviors. Instructors also apply consequences to behaviors in the forms of grades as well as verbal and written feedback. There is a rich literature related to behavioral interventions in the classroom with regards to learning, classroom behavior management, and skill acquisition (Sutherland, Lewis-Palmer, Stichter, & Morgan, 2008). Providing feedback to students has varying degrees of efficacy in improving student skills. It is likely that this variation is related to *how* the feedback is being applied. In a study by Gielen et al. (2010), feedback accuracy only predicted student performance when the feedback comments were frequent and justified (contingent on the individual's behavior). They concluded that to provide highly accurate and effective feedback, effort to frequently provide feedback and time to carefully justify the feedback were required. However, both classroom and instructor time is limited. In an ideal setting, students would have an overabundance of time to improve their skillsets, and instructors would have an overabundance of time to apply the most effective feedback possible to encourage learning and skill development. According to Voerman et al. (2012), an overabundance of time is far from reality and both instructor and classroom resources are stretched thin, which directly contributes to less effective feedback strategies. Because of this situation, it is especially important to maximize the effectiveness of behavioral consequences in the form of student feedback in order to support student learning in these conditions. To this end, there is an open question on whether instructors should increase feedback frequency or increase feedback quality (Gielen et al., 2010). If feedback frequency is more effective, instructors should spend their limited time and resources focusing on providing students feedback as often as possible. If feedback quality is more effective, instructors should spend their time providing higher quality feedback. Determining the functional relationship between feedback quality and feedback frequency on skill improvement within the underlying context of behavioral theory has the potential to map out strategies for effective teaching.

Current Study

The current study aimed to compare the degree to which feedback quality and feedback

frequency impact skill improvement. To this end, we examined and compared the impact of both high-quality and high-frequency feedback strategies on student English writing skill development in the context of an English writing course. In this natural experimental design, students randomly assigned to different classes completed five essays throughout the course of the semester, while receiving different types of feedback from their instructors. Each essay was objectively graded using an AI-based essay scoring system. By determining the overall and relative effectiveness of the feedback strategies, we hope to provide behavioral guidelines as to how instructors can maximize feedback to increase student skill development—even with limited time. By focusing on skill acquisition and development in general, it is our hope that the findings can extend beyond English writing proficiency and generalize to other skill-based learning common in other departments.

METHOD

The current investigation was carried out using a natural experimental design. In the Global Communication Department's English Course, approximately 100 students were randomly assigned to writing classes A, B, C, or D. During each of these writing classes, students wrote a set of approximately 5 essays as a part of their coursework. Each class met twice per week and was taught by different teachers on each day. There were four teachers total, two who teach classes A and C and two who teach classes B and D. At the time of this project, there was not yet an agreed-upon standard for feedback among all the English writing classes, and each teacher naturally provided feedback differently in terms of both quality and frequency. Two teachers provided written feedback approximately 2-4 times per semester with thorough comments on the essay content in its entirety. This type of feedback was categorized as high-quality/low-frequency. Another teacher provided individual feedback to each student on their essays every week (approximately 12 times per semester). However, this teacher used a computer-assisted AI algorithm to correct and score the essays. The algorithm catches most errors but is not as thorough as hand-grading. As such, this form of feedback would be considered computer-quality/high-frequency. The fourth teacher did not provide feedback to students outside of their final grade. This condition would be considered no-quality/no-frequency. Because each student is paired with two teachers, each student experiences two conditions. Thus, the resulting data was nested across three levels: approximately 500 essays nested within approximately 100 students, and students nested within each of the four classes. The breakdown is depicted in Table 1 below:

Measurement and Evaluation

In order to objectively assess changes in English writing ability, we used computerized scoring for all essays from all classes. By using a computerized scoring system, bias from human raters is all but eliminated, and the internal validity of the experiment was maintained. PaperRater (www.

paperrater.com) is an automated essay scoring system that uses a deep neural network trained on over 100 million student essays. The system evaluates spelling, grammar, transitional phrases, sentence length, passive voice, and vocabulary, generating a scaled score on each domain which are further combined into an overall score for the essay itself. We scored all essays from all participants using this system and using changes in the overall score as our marker for progress and improvement in English writing ability. A all essays were scored using the PaperRater algorithm, and the resulting information entered in a database to be statistically analyzed.

Analytic Strategy

In order to examine overall improvement in English writing abilities, we incorporated a linear mixed-effects model that regresses student scores[1] over time while allowing for random variation associated with each student's overall ability level, as well as classroom effects. The conditions of quality (none, human, computer) and frequency (0, 3, 12) were entered as main effects into the equation.

Equation 1. Linear mixed-effects model for overall skill improvement.

$$score_{ij} = \beta_{0j} + \beta_1 week_i + \beta_2 frequency_j + \beta_3 quality_j + \beta_4 quality * week_j + \beta_5 frequency * week_j + \mu_1 student_j + \mu_2 class_j + e_{ij}$$

RESULTS

Student and Essay Descriptive Statistics

Class A consisted of 52 second year students, while class B, C, and D consisted of 35, 49, 30 students respectively. In total, 665 essays were submitted by the 133 students (M = 2.51, SD = 1.35). 43.6% of students submitted all 5 essays, whereas 9.7% submitted 4 essays, and 46.7% submitted three or fewer essays. As can be seen in Table 1, the frequency of participation in the project was uneven across conditions.

Table 1. Number of students and essays in each condition

Class	Teacher 1 Computer-Quality 12x Feedback	Teacher 2 & 3 Human-Quality ~ 3x Feedback	Teacher 3 & 4 None 0x Feedback
A	---	52 students (89 essays)	---
B	35 students (140 essays)	---	35 students (95 essays)
C	---	49 students (85 essays)	---
D	30 students (163 essays)	---	30 students (93 essays)

Student essays scored an average of 83% (SD = 8.26) and varied slightly by class and cohort. These differences were statistically significant ($F[3, 661] = 9.90$, $p < .001$). Please see Figure 1 for a visual representation of the distribution of scores.

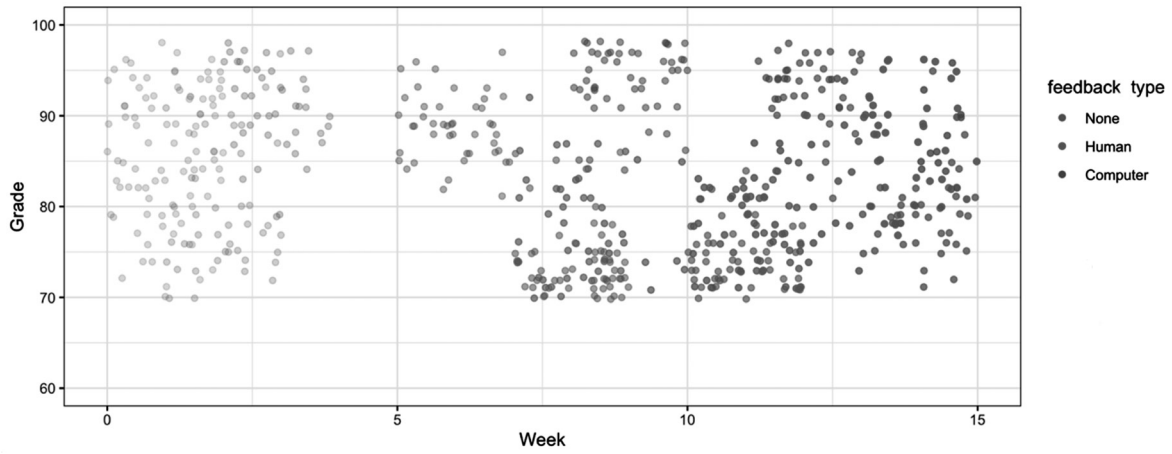

Figure 1. Essay scores per class

Impact of Feedback Quality and Quantity on Student Performance

In order to examine the impact of feedback quality (computer, human, none) and feedback quantity (0x, 3x, 12x; coded as 0, 1, and 2 to avoid singularity in the linear model), we examined student scores using the linear mixed effects model described above. As can be seen in Table 2, our model identified a statistically significant positive main effect for Feedback Quality (B = 57.67; $p < .001$) as well as a statistically significant negative main effects for Feedback Quantity (B = -25.38; $p < .001$) and no effect for Week (B = -.07; p = .45). Post-hoc analyses identified that computer quality (B = 10.37, $p < .001$) was related to higher essay scores as compared to human quality (B = -8.13, $p < .001$).

Table 2. Results from Linear Mixed-Effect Model

Variable	B	SE	DF	T-value	P-value
(Intercept)	78.29	2.38	494	32.84	< .001
Week	-0.07	0.09	494	-0.75	.45
Feedback Quality	57.67	9.95	494	5.79	< .001
Feedback Quantity	-25.38	5.37	494	-4.72	< .001
Week x Feedback Quality	-1.99	0.38	494	-5.19	< .001
Week x Feedback Quantity	1.15	0.21	494	5.49	< .001

The interaction effects with week helped nuance the main effect findings. Both the Week ^{x} Feedback Quality (B = -1.99; $p < .001$) and Week ^{x} Feedback Quantity (B = 1.15, $p < .001$) were statistically significant. The low-magnitude, negative relation with Feedback Quality over time indicated that, conversely, the positive relation of Feedback Quantity over time indicated that increased feedback facilitated score improvement over the semester cumulatively. A post-hoc analysis of Feedback Quality indicated that compared to Human Feedback Quality (B = 1.15, $p < .001$), Computer Feedback Quality did not result in as steep of an improvement slope (B = .38, p = .002). Please see the relative slopes as depicted in Figure 2.

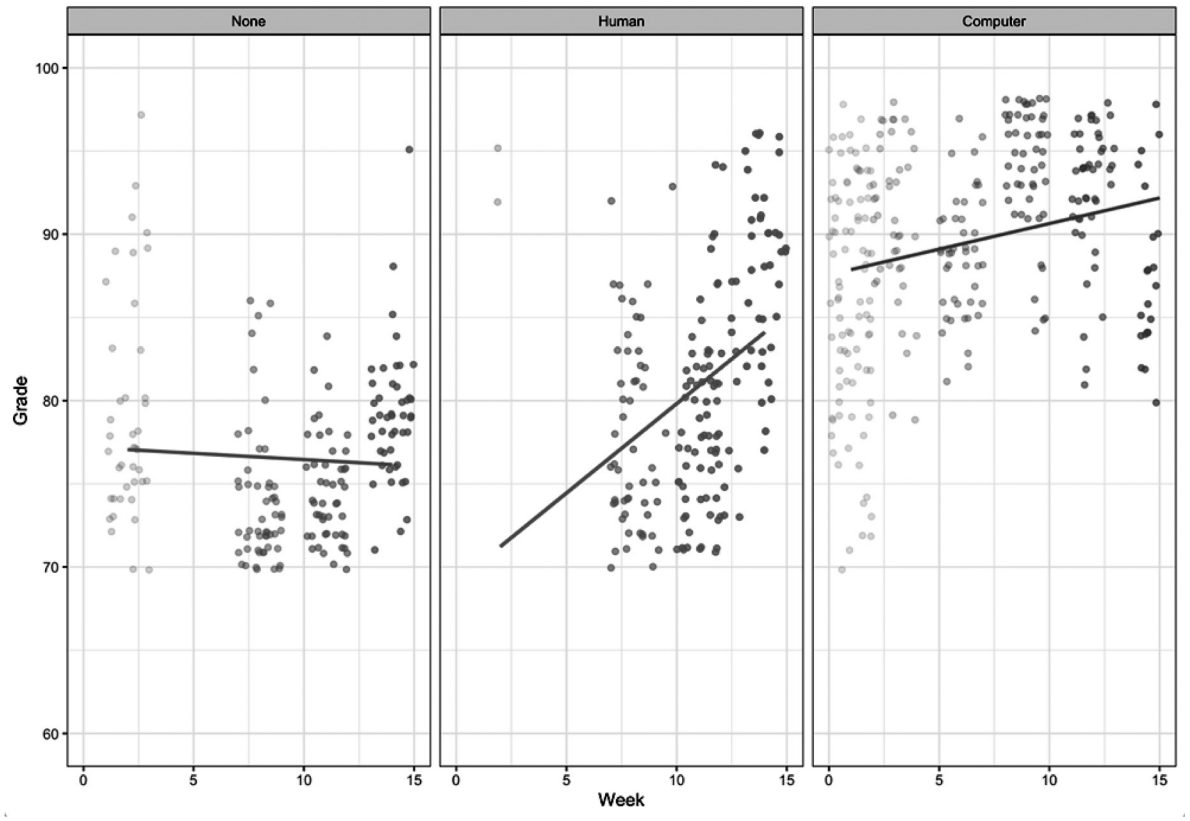

Figure 2. Improvement of Student Writing Scores Over Time per Feedback Category

DISCUSSION

The current study sought to examine the role of Feedback Quality and Feedback Quantity on student writing skill improvement over the course of the semester. We hypothesized that while carefully graded human-quality feedback is important, feedback quantity would better predict improvement in writing scores. Our results show a much more nuanced picture of the relationship between feedback quality/quantity and score improvement. First of all, not all students improved over the course of the semester. Specifically, in the "No Feedback Group", student scores stayed the same or decreased as the semester continued.

Our second finding was that students benefited from computer-quality feedback at every point in time. It appears that having a systematic, consistent grading format helped students know how to correct their essays. Third, feedback quantity was not beneficial overall, but had a cumulative effect in that high-frequency, consistent feedback facilitated better grades as time went on. Finally, there appears to be an interaction between Feedback Quality type (Human vs. Computer) and the cumulative effect over time, with human-quality feedback increasing student scores more over time compared to computer-quality feedback. An important qualifying factor, however, can be seen in Figure 2. In the later weeks of the semester, many students (38.46%) were maximizing their grade by scoring in above 93%[2], less improvement was possible given the ceiling effect. Whereas in the human-quality feedback group, only 6.97% reached this range in the final few weeks.

When examining the results overall, the core principles of behavioral psychology seem to be maintained. Frequent, consistent feedback, whether by a human or by a computer is key to skill improvement. Contrary to expectations, however, the results from the current study highlights that computer scoring, rather than human scoring may be a more effective avenue to attain the best results in a shorter period of time. However, this remains little more than a hypothesis given the current study's limitations. Specifically, the current study was a quasi-experimental design due to teacher's feedback styles being unable to be randomly assigned. Furthermore, the participation rate for each of the conditions is unbalanced, likely skewing the results. Future studies designed to test a causal model would also benefit from increasing the number of conditions and randomizing feedback quantity across feedback quality. Likewise, even if efficacy is established, it is essential to acquire use-case feedback from students in order to understand the phenomenological experience of receiving computerized feedback.

Conclusion

The current study innovatively combines psychological and educational theories with an experimental design and modern grading technology to provide guidelines for instructors to maximize the feedback they give to their students. By focusing on skill acquisition in general, we hope to provide guidelines that extend beyond English writing proficiency and can be readily incorporated into other skill-based learning endeavors across other departments. Specifically, computer scoring systems may be a fruitful area for further research on classroom interventions.

Footnotes

1: Please note that every reference to “student scores” or “student grades” reflect the score provided by PaperRater rather than the student's actual grade in the class, which was at the discretion of each individual teacher.

2: In the PaperRater scoring system, scores above 96% are very difficult to attain, and can be considered the ceiling.

謝辞

このプロジェクトは、神戸学院大学教育改革助成金を受けて実施されました。

REFERENCES

［1］ Austin, J., & Carr, J. E. (2000). *Handbook of applied behavior analysis*. Oakland, CA: Context Press.

［2］ Baum, W. M. (2017). *Understanding behaviorism: science, behavior, and culture. Understanding behaviorism: science, behavior, and culture*. Chichester, UK: John Wiley & Sons.

［3］ Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. East Norwalk, CT: Appleton-Century-Crofts. http://doi.org/10.1037/10627-000

［4］ Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20,* 304–315. http://doi.org/10.1016/j.learninstruc.2009.08.007

［5］ Hayes, S. C., Sanford, B. T., & Chin, F. T. (2017). Carrying the baton: Evolution science and a contextual behavioral analysis of language and cognition. J*ournal of Contextual Behavioral Science, 6,* 314–328. http://doi.org/10.1016/j.jcbs.2017.01.002

［6］ Skinner, B. F. (1938). *The Behavior of Organisms: An Experimental Analysis. New York.* New York: Appleton-Century-Crofts.

［7］ Skinner, B. F. (1953). *Science and human behavior.* New York, NY: Simon and Schuster.

［8］ Skinner, B. F. (1963). Operant behavior. *American Psychologist, 18,* 503.

［9］ Sutherland, K. S., Lewis-Palmer, T., Stichter, J., & Morgan, P. L. (2008). Examining the influence of teacher behavior and classroom context on the behavioral and academic outcomes for students with emotional or behavioral disorders. *Journal of Special Education, 41,* 223–233. http://doi.org/10.1177/0022466907310372

［10］ Thompson, R. H., & Iwata, B. A. (2005). a Review of Reinforcement Control Procedures. *Journal of Applied Behavior Analysis, 38,* 257–278. http://doi.org/10.1901/jaba.2005.176-03

［11］ Voerman, L., Meijer, P. C., Korthagen, F. A. J., & Simons, R. J. (2012). Types and frequencies of feedback interventions in classroom interaction in secondary education. Teaching and *Teacher Education, 28,* 1107–1115. http://doi.org/10.1016/j.tate.2012.06.006