Executive Summary

Evaluation of Packback

SCHOOL DISTRICT OF OSCEOLA COUNTY

2023

Authored by: David Maddock

Complete details of all analyses and literature references, can be requested from REA

Executive Summary

Evaluation of Packback

Background

A multitude of digital support tools exist for instruction and intervention. As part of the continuous evaluations of the various tools and resources that the School District of Osceola County spends money on, a program called Packback – a digital platform that purports to help students develop their writing and expression skills through the implementation of generative AI – was evaluated to determine its effects on students learning, and the cost associated with that learning. Packback began its operation in 2022 as a pilot program to be utilized in a select classrooms for the end of the 2021-2022 school year before the pilot was extended through the end of SY2023.

Digital platforms that provide differentiation through targeted interventions have become increasingly common in recent years. Despite their popularity, however, the effects of these programs are often low. In 2009, John Hattie established a baseline for the expected effects a teacher can provide a student in their growth as $d = 0.40$, a solid baseline expectation for interventions to meet. In 2019, after reviewing 747 randomized control trials, Matthew Kraft at Brown University proposed a new means of interpreting d effect size in relation to interventions, with less than $d = 0.05$ as a small effect, and greater than $d = 0.20$ as a large effect.

While Packback is relatively new, it's utilization of artificial intelligence (specifically, a pre-trained language prediction model) has led to a fair amount of attempts to evaluate its effects. Hudson, Archibald, & Heap (2020) found confounded and muted effects from Packback usage when compared to Canvas Discussion Boards, but did find that there was potentially a difference in implementation based on teacher familiarity with the platform. Rizzuto (2022) did not find evidence of Packback being associated with student's perceived learning. Within the larger context of AI-backed educational resources, Hwang (2022) stressed that the learning environment and implementation methodology of the teacher had a sizable impact on the effects on student achievement. Still, the overall body of research on AI generally leans towards larger effects, particularly when paired with gamification (which Packback also provides). For this reason, the expectation for the Packback platform should exceed $d > .20$.

Since Packback's introduction to Osceola County, a total of 1547 students entered the platform. During the current year, a total of 22 teachers logged on to the platform at least once. While Packback was unable to provide usage statistics for student participants (this evaluation will be updated in the future if those are furnished; raw data by teacher was provided), effects were isolated to determine the implications of Packback usage, and teachers were polled for their perceptions of the platform. Additionally, Packback provided their own results from teacher surveying in Osceola County, which are also provided here alongside internal polling.

Purpose

The purpose of this evaluation was to examine the effectiveness of the Packback for the cost per student.

The following evaluation questions were posed for each tool:

1) Is success on Packback associated with success on measures of standardized assessments?
2) What are the user perceptions of the platform?
    a. How often is Packback utilized in Osceola classrooms?
3) What is the cost per student of Packback?

Methodology

Quantitative methodologies via statistical analyses were utilized to examine the effects of each current program. For the evaluation of Packback, data were requested from Packback (curiosity score growth, citation usage, word count growth, usage statistics), however the only data furnished were teacher participants (*n* = 22), total number of students who've utilized the platform to date (*n* = 1547), and teacher perception surveys (*n* = 3). Students from the courses of the identified teachers had their individual data collected, including grade, school, NWEA results, overall course grade, and mock AP exam results. Assessment data were collected from SchoolCity, and internal data sources with each student in a row context. Data for the NWEA assessment were collected directly from the NWEA platform and matched to the student records. The data related to platform costs were collected via quote. Statistical tests were performed to compare differences among students. All statistical analyses were performed using SPSS 27.0.

Key Findings

Packback-NWEA Associations

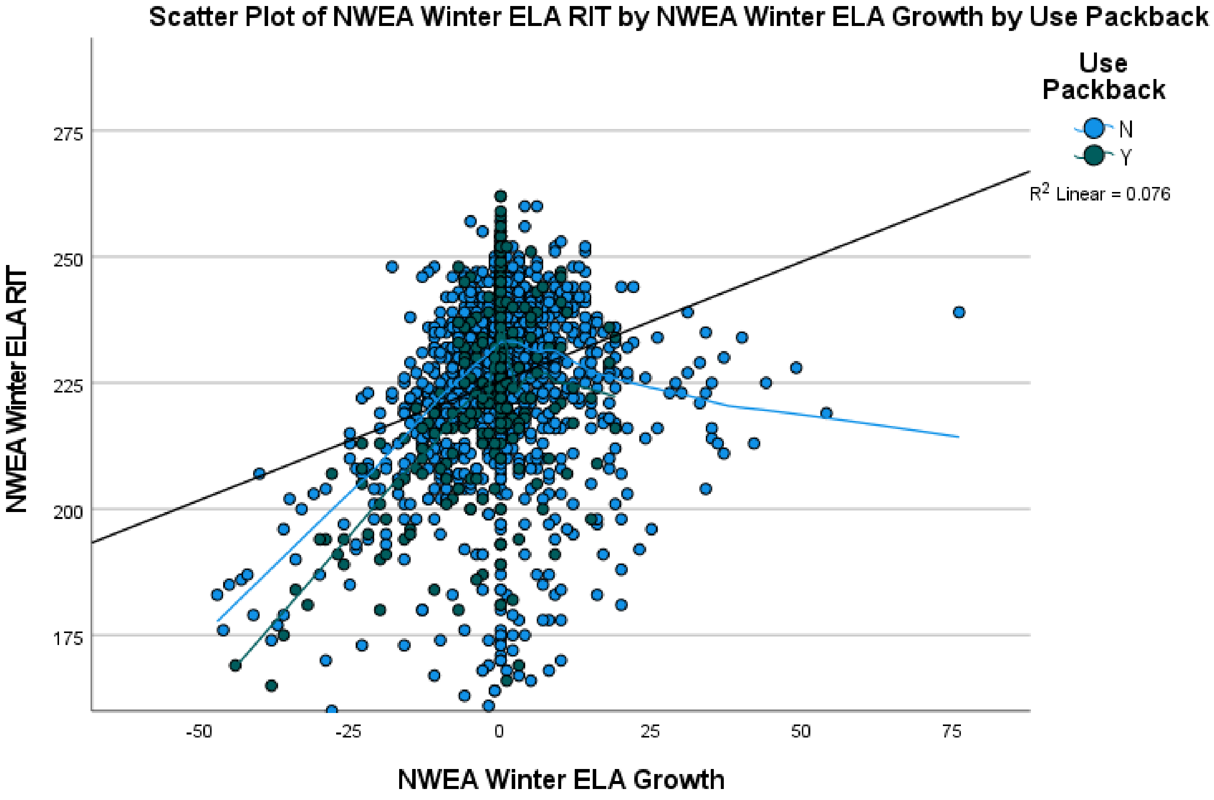
While all analyses would have been much richer if Packback could have identified which students were utilizing the platform and for what amount of time, students were divided into two groups: those that were in a course with a teacher that had more than five class posts, and everyone else. A total of 635 students were identified as exposed to Packback in this manner, and they were further demarked as whether they were exposed in either English, Biology, or US. Gov. While Packback reports that 1547 students have logged into the platform, since their data spans two years it likely includes many students who are no longer enrolled in the district, including previous year seniors.

For the first analysis, all students exposed to Packback had the results of their NWEA Winter assessment compared against students who were not exposed to Packback but were in similarly rigorous AP courses to those where Packback is most often used. This design was meant to focus the effects on a comparison of students with similar ability to those who are in classrooms that utilize Packback (since comparing to all students would result in students performing far below grade level dragging down the comparison mean). The effects for Packback are quite difficult to distinguish for a few reasons: 1) of the six active teachers on the platform, one teaches IB, one teaches CTE, and one teaches AVID which means that 2) there is not a single writing assessment that all of these students take since 3) these students are enrolled in grades that do not complete the state writing assessment. This means that the NWEA is the *closest* we can get to an analysis of effects on standardized measures, but there is still a large amount of assumption built into the analysis and therefor the results should be taken with an appropriate measure of skepticism.

It is also worth noting that in all further sub-analyses of effects, the students in the Packback group are often learning from a single teacher (and in most cases a highly-effective teacher at that). It is impossible to distinguish the Packback effects from the teacher effects (although a datapoint such as time on platform would allow for correlative analysis). This is to say, there's a distinct possibility that the students are growing because *their teachers are effective* rather than any effect from the platform.

Students in classrooms that utilized Packback were analyzed based on the NWEA MAP Reading performance. While Packback primarily asserts to assist in writing skill, the NWEA MAP Reading assessment is the closest and most reliable measure we have to the standardized assessments that these students will be measured on - the BEST, ACT, SAT, or AP exam – and is therefore an essential assessment to examine for changes in performance. A *t*-test was conducted between the RIT scores for students in classrooms that utilized Packback (*n* = 265, *M* = 224 RIT) versus students in other advanced placement courses that did not use Packback (*n* = 1191, *M* = 226 RIT). Students in classrooms that used Packback performed slightly worse than students in other advanced placement courses, *t*(1454) = -1.640, *p* = .051, *d* = -0.11 although the difference was just shy of statistical significance. When considering growth between the Fall assessment and the Winter assessment, students in classrooms that did not utilize Packback experienced growth that was generally in line with expectations (*M* = 0.24 RIT), while students that utilized Packback saw a drop in performance over the four months of exposure, *t*(1454) = -4.289, *p* < .001, *d* = -0.29, at a highly statistically significant level. This indicates that students in classrooms that utilized Packback grew less than students in advanced courses that did not use Packback.

Since there are students utilizing Packback in courses such as AVID, the potential exists that there was a moderating effect that led to the lower aggregate results. Therefore, a follow-up analysis was conducted comparing the NWEA scores for students who used Packback in an IB ELA classroom compared to students who did not use Packback but were in an AP ELA course. Students exposed to Packback (*n* = 109) has a mean Winter NWEA RIT of 239, while students who were not exposed to Packback (*n* = 242) had a mean RIT of 236. This difference, *t*(349) = 1.856, was statistically significant, *p* = .032, indicating that students in classrooms that utilized Packback score higher, an effect of *d* =0.21. However, when conducting a *t*-test based on student growth, it was seen that students in both groups of classes experienced nearly no growth (Packback growth *M* = -0.25, No Packback growth *M* = -0.61) with no statistically significant differences between groups (*p* = 0.297). A further follow-up analysis utilizing prior-year FSA scale scores found a similar three-scale-score-point difference, indicating the difference in student abilities existed in the year before the students were exposed to Packback, and that no major difference in growth on ELA measures has occurred since.



These results indicate that utilization of Packback is not associated with growth on measures of standardized assessments.

Packback and AP Mock Exams

Although the NWEA MAP is the most highly-aligned measure available to the district when comparing to required standardized assessments such as SAT, ACT, or BEST, it offers a somewhat obfuscated measure of the effects of Packback. For that reason, a different analysis was conducted comparing the results of the AP Mock exams between students who utilized Packback and those who did not. Given the extremely small quantity of teachers who utilized the platform, only two teachers could be identified who both utilized Packback *and* had their students take the mock exam: an AP Biology teacher, and an AP US Government teacher.

For AP Biology, 121 students completed the assessment within the district. Out of this population, 22 students were with a teacher who utilized Packback, while 99 were not. The mean AP raw score was higher for the students who used Packback (*M* = 32.3, *SD* = 8.2) than for students who did not (*M* = 30.3, *SD* = 10.3), but the difference was not statistically significant (*p* = .203, *d* = 0.19). This means that there is a likelihood that the difference between the scores exists due to either random chance, or some other effects (such as the ability of the teacher).

For AP U.S. Government, 159 students completed the assessment within the district. Out of this population, 40 students were with a teacher who utilized Packback, while 119 were not. The mean AP raw score was higher for the students who did not use Packback (*M* = 33.9, *SD* = 7.6) than for students who used Packback (*M* = 27.6, *SD* = 8.2). The difference was highly statistically significant, $t = -4.320$, $p < .001$, $d = -0.79$. Students in the classroom with Packback scored almost a standard deviation below those in other classrooms.

Yet, none of this evidence is conclusive. While the results overlap and conflict, the reason why remains plain: adoption for Packback as a platform is critically low, and therefore the effects are greatly exacerbated by the standard teacher effects and the standard student effects. These results are much more a reflection of the total teaching in a particular classroom than they are the effect of the platform. For that reason, it is also essential to analyze how many teachers are using the platform, and the why behind their usage.

Packback User Perception and Usage

Packback provided raw data on the total posts in each classroom each month between August 2022 and February 2023. While the data was not provided at the student level, it does provide some enlightenment about total usage of Packback within the county. Nine teachers had students create at least one post, with the lowest engagement being a teacher who had one student write one seven-word post, and the highest engagement being a teacher whose students wrote 6024 posts that consisted of 892,729 words. Among all users for the 22-23 school year, a total of 11,821 posts and 6148 replies were written, comprised of 1,778,369 words. It is clear that effective usage of the platform can result in large amount of writing, however it seems that it is uncommon for teachers to adopt platform usage into their classrooms.

To better understand this, two surveys were conducted on the 22 users in Osceola County of the Packback platform (user is defined as someone who has at least logged into the platform; only nine users assigned a question to students). The first survey, conducted by Packback, surveyed teachers across two years (n = 5). The second survey, conducted by Osceola REA, surveyed teachers in April of 2023 (n = 8). Both surveys used a mix of Likert-type response items and open-ended questions, although the Osceola Survey also used a Net Promoter Score item as well.

The survey responses provided to Packback were generally positive; in year one of surveying there was an average response of 5.7 (out of 7) to the item “Packback positively impacted my students” and a general agreement that platform made teachers lives easier. In the second year of surveys (n = 2), both respondents strongly agreed with a majority of statements about Packback, including that it improved students’ writing skills, that increased senses of belonging, engagement, and satisfaction, and that they would use the platform in the future.

The results to the internal Osceola REA Survey (n = 8) were considerably less favorable. A Net Promoter Score analysis was conducted to determine how likely participants were to recommend Packback to other teachers. Three teachers responded in the “promoters” (9 – 10), and no one responded in the “passives” (7-8) range. The remaining five respondents fell in the “detractors” (0-6) range, which resulted in a net promoter score (NPS) or -25, putting the platform in the “needs improvement” range, indicating that a majority of those responding to the survey are having a bad experience. Given that 16 of the 22 users who have interacted with the Packback platform fall in the range of “non-starters”, the survey results potentially illuminate why multiple teachers are not engaging with the platform.

The usage of Net Promoter allows for some rudimentary predictions of expansion beyond a pilot program, for example, an examination of how many likely users would adopt the platform if it were available to all high school classroom teachers. With 6 out of 22 users utilizing the platform for more than one month, extrapolating a similar trend to all 962 high school teachers, and applying an NPS effect of -25, the projected adoption rate for the platform (assuming no major changes in practice or implementation) would be 159 teachers. Packback’s proposed 2023-2024 plan covers 142 classrooms, which would, in this model, result in between 28 and 38 teachers adopting the platform. Not that this projection is anything like guaranteed, but it is mentioned to highlight the obstacles the platform would have to overcome to be successful in a wider adoption.

The respondents were often quite split on their responses as well. To the statement, “I like the Packback platform,” three respondents positively agreed, one negatively responded, and four said their neither agreed nor disagreed. The exact same spread was found on the item “Packback has a positive effect on my students’ performance”, although responses to “Packback is easy to use” were slightly more positive (five agreed), and the answers to “Packback fills a vital services” were slightly more negative (two agreed). Most respondents elected to skip the open response item “How do you use Packback in your classroom?”, but the responses were:

- I like using it as a previewing step, but sometimes there isn't a need for a previewing discussion. I'd like to use it the way they want us to use it, but I find it gets too chaotic in a high school class.
- My students use Packback to continue discussions and ask deeper questions for further analysis in an environment that allows outside of the box thinking and questionable interpretations. I also use deep dive to develop analysis skills and research skills.
- Like the idea of packback, but district does not pay for it

The Cost of Packback

The final question in the evaluation was related to cost per student and the potential return on investment from Packback. A total of 1547 students used the platform between January 1, 2022 and April 2023 at an initial cost load of $35,000 for user licenses across the two years of piloting. This results in an effective cost of $22.62 per student for the duration of the pilot (Packback actual rate of $25,000 for 1400 licenses in the 22-23 school year should result in a rate of $17.85 per student, meaning Osceola effectively paid more than expected due to some students not utilizing the platform). Based on the quote for the 23-24 school year, the actual cost per student would lower considerably since the proposal covers 3556 students and costs $49,383, which is an actual cost of $13.88 per student, similar in cost to platforms with ELA interventions such as DreamBox ($15 per student) and Freckle ($14 per student), although the target audience for Packback is considerably different from these platforms. It is worth noting that discounts were applied to this rate, and the cost before discounts was $59,000, which would be a cost of $16.59 per student were the discount to not be offered in future years.

When considering the platform based on teacher usage, only 22 teachers engaged with Packback at *any* level, according to data from the company. Although 16 of the 22 were considered “non-starters” who used the platform for less than one month, they can still be factored into the cost per teacher, resulting in a cost of $1,590 per teacher user. When only considering users who engaged with the platform regularly, the cost raises to $5,833 per teacher user. Another way to think of the cost is in dollars-per-post: accounting for all posts and replies, 17,969 piece of content were created. Given the 22-23 cost of $25,000, Packback cost $1.39 every time a student used the platform. While this cost would obviously come down drastically with greater penetration of the platform, the negative NPS for the platform in Osceola indicates that is unlikely to achieve the desired penetration for the program to be cost efficient.

One final note: it is generally the recent opinion that program evaluations in education should determine a return on investment in the platform. Given the extremely small student population exposed to the platform, and the general lack of fully aligned assessment, however, it was not possible to conduct an ROI analysis to the depth expected from the ROI Institute. Therefore, an ROI analysis was not conducted for Packback.

Conclusion

It is clear from the research that the future of the classroom *will* involve some form of AI-driven partner in education. An array of platforms are racing to offer AI classroom aides; just next year alone Khan Academy will provide an AI tutor for students on SAT work (included in the current contract), while platforms like Julian, Bard, and ChatGPT continue to expand their roles in schools. This only underscores the fact that extreme care must be taken to identify a solution that is both effective for students and well-liked by teachers.

The results from the quantitative analysis were fully in line with research from Hudson et. al (2020) and Hwang (2022). Effects on the typical measures of academic success were muted and conflicting, likely due to both aspects of the platform itself, and the significantly limited adoption of the platform. As Hwang noted, the implementation of Packback is a significant factor in its success in the classroom, and the fact is that many teachers did not engage with platform very well. Effects from the analysis ranged wildly, from a low d = -0.79** to a high of d = 0.21*. It was difficult to conclusively state that Packback has any association with student learning (positive or negative) simply because the utilization was so limited.

For the cost of the platform (quoted at $49,383 in 2023-24) it is likely that adoption of Packback would come with significant risk. The potential for the platform to be transformative exists: students in the three classrooms that utilize it consistently have some of the highest measured quantities of writing seen in Osceola program evaluations. Yet a majority of respondents did not have positive perceptions of the platform on the internal survey (it is highly likely that the three positive respondents on the internal survey were the same three respondents on Packback's external survey, who were the three users with high utilization of the platform) so the trend would indicate that that full and transformative adoption of the platform within the county is highly unlikely. While common applications of the Net Promoter Score would indicate that there could be a negative effect from the number of detractors that exist on the platform, the fact is that the current utilization of the platform is so slight (only 2% of high school teachers, or or 0.5% of Osceola teachers have even logged into the platform) that it is unlikely their detraction would affect the expansion or implementation of the platform much at all.

While it is possible that Packback could have an impact on Osceola students, the fact is that most users who were exposed to it elected not to engage. This drives up the relative cost of the platform and drives down the potential effects. Given the lack of engagement with the platform currently observed, and the high relative price in the field, it is not recommended that schools utilize the Packback platform in the coming year.