Where AI meets the learner: Classroom as a mediator

Shiyi Shao, Beata Beigman Klebanov, Anastassia Loukina, Priya Kannan, Paola Heincke

Educational Testing Service, Princeton, NJ
{sshao, bbeigmanklebanov, aloukina, pkannan, pheincke}@ets.org

Abstract

We report on a pilot of Relay Reader, an app built to help developing readers improve their reading skills. The pilot was run in 5 summer camps with 133 3rd to 5th graders led by 13 camp instructors. Analyzing children's use of the app, we observed substantial differences across classes in the fidelity of implementation. We investigate variation in instructor attitudes and experiences as well as classroom settings, using interview, demographic, reading log, and skill assessment data collected during implementation.

Summer Camp Pilot of Relay Reader

Relay Reader™ is a reading and listening app designed to help developing readers successfully engage with long and potentially challenging books with the goals of helping them (a) enjoy the experience and improve their attitude towards reading, and (b) improve reading skill through sustained listening and reading. With this app, the child takes turns reading out loud with a pre-recorded skilled adult narrator (audiobook). As children read, the app logs their actions, with timestamps. Audio from the child's reading turns is recorded, stored, and processed for feedback and analysis. For more information, see Beigman Klebanov et al. (2019).

The app was piloted in 4 summer camps in the greater Washington DC area and 1 camp in New Jersey. All camps belong to the same national system, hence they share the same curriculum, with some differences. In all camps, children were organized into grade-based levels; we worked with the middle level – 3rd to 5th graders. All participants read Collodi's 'The Adventures of Pinocchio' with the app. At the start of the program, participants took a standardized reading comprehension test (see Sabatini et al, 2013).

Ahead of the pilot, the project staff traveled to Washington DC to train prospective camp instructors and site supervisors. The NJ supervisor was familiar with the app from a prior trial; he trained his staff (all new to our program) using our training materials. The app was presented and the target implementation schedule was discussed. The app was to be used for about 20 minutes 3 times a week for the duration of the camp (6 weeks). Each instructor was sent a weekly report of the children's reading activity. The research team interviewed instructors during the 1st, 3rd, and 5th weeks, and a final interview. The sites, the instructors, and the supervisors were compensated; children received a paperback copy of Pinocchio as a participation award.

A Measure of Fidelity of Implementation

Our first task was to measure the fidelity of implementation – whether kids read the book. In prior work (Beigman Klebanov et al, 2019), we devised a method to estimate reasonable duration of a reading turn, based on norms of oral reading rates by grade and estimates of within-reader variation across texts. As fidelity index (**FI**), we calculated, for every classroom, the **percentage of children who mostly read within reasonable durations**; more precisely, the proportion of children with at least 60% reasonable turns. FI is shown in column 2 in Table 1. Note that this is not a performance measure, since reasonable durations include rates expected for children at the lowest 10th percentile in 3rd grade distribution who are reading the slowest-to-read text, as well as the highest 10th percentile of 5th graders who are reading the fastest-to-read text. The duration-based measure is robust to problems in recording or processing oral reading; it estimates whether bona fide oral reading could have happened given the turn duration.

Exploring Variation Across Classes

We observed a substantial variation in FI across classrooms (see column 2 in Table 1). In 4_2, none of the children read consistently; in 1_1 all but one child did. We then explored whether there is a correlation between FI and various factors that could potentially impact success of implementation. FI was not correlated with class **size** (column 3). In contrast, we see a correlation with **participation** (column 4, $r = 0.67$). Children often attend camp on a flexible schedule – only

certain days of the week and/or certain weeks. Inconsistent participation could detract from the child's immersion in a story. We also observe a correlation with **grade** (column 5): classes with older children had better FI ($r = 0.72$). The potential impact of grade could be related to skill – older children are expected to be better readers and find it easier to persist with reading. Indeed, an estimate of **skill** is correlated with FI ($r = 0.78$); yet, the partial correlation between grade and FI controlling for skill is still $r = 0.53$ ($p = 0.08$). It is also possible that class management is more challenging with younger children. Maturity could also be related to the child's ability to use the app independently and consistently.

Class	FI	Size	Participation	Grade	Skill	Training	Experience	Involvement
4_2	0	10	.5	2.7	236		1o	1
5_1	10	9	.6	4	249	S	*	1
5_3	10	10	.7	3	243	SI	3c1o	3
4_1	14	7	.71	3.1	246			2
3_2	21	14	.46	4	245	I	1o	3
2_1	25	12	.63	3.7	251	SI	1o2c	4
1_3	33	9	.5	3.9	243	S	*	*
1_2	55	11	.82	3.7	247	S		2
1_4	55	11	.73	4.1	250	S		2
3_1	56	9	.67	4.2	246	I	7c	1
5_2	60	11	.7	4.2	256	SI	3e6c	0
2_2	64	11	.73	4.3	253	SI	1o	2
1_1	89	9	.83	4.1	273	S	*	1
r		.04	.67	.72	.78			-.34
p		.90	.01	<.01	<.01			.26

Table 1: Classroom data sorted by FI. In **Class**, the first number indicates camp, the second – classroom within the camp. **FI**: fidelity index. **Size**: number of readers in a class. **Participation:** median proportion of children using the app during a session. **Grade**: av. most recent completed grade. **Skill**: av. reading comprehension test score. **Training** attendance: S (supervisor), I (instructor). **Experience**: number of years and context (**e**lementary teacher, **o**ther teacher, **c**amp); empty cell means no experience. Instructor **involvement** in implementation (see text). Asterisk indicates missing data. ***r*** is Pearson correlation with FI; ***p***-value is shown underneath the correlation.

To explore additional factors, we turn to information collected in our communications with instructors and supervisors. The column **Training** shows that in only 4 out of 13 cases both the instructor and the supervisor attended training. The column **Experience** shows that camp instructors generally were not very experienced educators: 3 reported no prior teaching experience (4_1,1_2,1_4), 3 had one year in another context (afterschool, high school). The instructor of 2_2 stands out as an inexperienced teacher with a higher FI. She reported working on her M.Ed., including coursework in children's literature; she mentioned bringing *Harry Potter* and *Narnia* books to class and has perhaps succeeded better than most others in creating a good atmosphere around independent reading of novels for enjoyment.

While all instructors observed the children, answered their questions, and managed the classroom, some instructors reported more extensive **involvement**, such as explicitly helping weaker readers by sitting with them and breaking down the reading turns for them; looking at the reports, keeping notes, and communicating the feedback to the class; motivating children through discussions and incentives. The last column in Table 1 shows a quantification of the reported involvement, with 1 point to each behavior type. Closer instructor involvement does not seem to have a clear relationship with fidelity, perhaps because some involvement strategies were better than others, as well as due to interaction with other factors.

Technical problems could also affect fidelity. Poor WiFi plagued multiple sites, especially camp 5. According to an instructor, children were initially excited but became tired and distracted due to the long wait for the app to load. Class 5_2 had better FI in spite of these problems, possibly due to the instructor's experience, among other reasons.

Conclusion

We presented an analysis aimed at understanding the observed variation in implementation fidelity of a reading program with an app during a trial in summer camps. Factors related to the composition of the class, attendance, technical problems; instructor training, teaching experience, and involvement patterns were considered. While certain practical implications for subsequent implementations can be derived from our study (load-testing WiFi; hosting multiple training sessions; using longer or shorter stories depending on expected extent of participation), our bigger point is that a large variation across many dimensions along with strong correlations with fidelity for some of them point to the necessity to better understand the impact of context of use on the extent to which AI has the intended educational effect.

References

Beigman Klebanov, B., Loukina, A., Madnani, N., Sabatini, J., Lentini, J. 2019. Would you? Could you? On a tablet? Analytics of children's e-book reading. In *Proceedings of the Learning Analytics and Knowledge Conference,*106-110. Tempe, AZ: ACM.

Sabatini, J., Bruce, K., Steinberg, J. (2013). Sara Reading Components Tests, RISE Form: Test Design and Technical Adequacy. *ETS Research Report Series*, 2013: i-25.