Methods for Language Learning Assessment at Scale: Duolingo Case Study

Lucy Portnoff
Duolingo
lucy@duolingo.com

Erin Gustafson
Duolingo
erin@duolingo.com

Joseph Rollinson
Duolingo
joseph@duolingo.com

Klinton Bicknell
Duolingo
klinton@duolingo.com

ABSTRACT

Students using self-directed learning platforms, such as Duolingo, cannot be adequately assessed relying solely on responses to standard learning exercises due to a lack of control over learners' choices in how to utilize the platform: for example, how learners choose to sequence their studying and how much they choose to revisit old material. To provide accurate and well-controlled measurement of learner achievement, Duolingo developed two methods for injecting test items into the platform, which combined with Educational Data Mining techniques yield insights important for product development and curriculum design. We briefly discuss the unique characteristics and advantages of these two systems - Checkpoint Quiz and Review Exercises. We then present a case study investigating how different study approaches on Duolingo relate to learning outcomes as measured by these assessments. We demonstrate some of the unique benefits of these systems and show how educational data mining approaches are central to making use of this assessment data.

Keywords

online learning; language learning; assessment; regression

1. INTRODUCTION

Online learning platforms have at their disposal large volumes of data about how students engage with learning material, how they navigate educational software, and how the learning process unfolds over time. Using a variety of methods - machine learning, statistics, psychometrics, etc. - Educational Data Mining (EDM) and Learning Analytics (LA) researchers identify students at risk of dropout from a course [e.g., 13], detect changes in study behavior [e.g., 11], predict exam performance [e.g., 1, 4, 12], and characterize the different learning strategies that learners adopt [e.g., 1, 12].

Duolingo is a learning platform that provides free language education through mobile apps and a website. With around 40 million users active on the platform each month, Duolingo may well possess the largest language learning dataset of any company or research institution. Researchers at Duolingo leverage EDM/LA methodologies to mine datasets - including internal assessment and log data - for insights that inform improvements to the learning experience, help identify opportunities for changes to curriculum design, and fuel research on second language (L2) learning more generally.

Due to the self-directed nature of the Duolingo learning platform and the desire for holistic learner assessment, we have developed two assessment systems - the Checkpoint Quiz and Review Exercises - that allow for carefully controlled measurement of learner achievement. These two assessments were designed with the challenges gamified platforms struggle with in mind, including ensuring the learning experience remains motivating and maintaining a scalable content creation process.

The utility of the Checkpoint Quiz and Review Exercises for assessing learner achievement depends, at least in part, on the high volume of data collected from Duolingo learners and the EDM methodologies that can be applied to that data. By leveraging predictive modeling and natural language processing (NLP) methods, we are able to control for the various ways that learners choose to navigate through the platform. Further, these methods allow us to uncover useful insights into how this variation in user navigation relates to learning outcomes - insights that we can leverage for product development and curriculum design. In this paper, we present two of our assessment systems and a case study highlighting the importance of applying EDM methodologies to derive insights from Duolingo assessment and log data.

2. RELATED WORK

Most EDM/LA applications at Duolingo focus on pedagogy-oriented issues [10] or computer-supported predictive analytics [2]. Most relevant to the current work are studies focused on predicting performance on upcoming course exercises [9] and predicting performance on an assessment [1, 4, 12]

Rather than relying on assessment data, some systems discussed in other studies instead model student interaction with and performance on individual course exercises. Knowledge tracing [7] is a popular approach for maintaining a model of whether students have learned specific concepts in a course. One system [9] compared the performance of a Bayesian Knowledge Tracing (BKT) model with a Deep Knowledge Tracing (DKT) model using Long Short-Term Memory (LSTM) to better capture longer-

Lucy Portnoff, Erin Gustafson, Joseph Rollinson and Klinton Bicknell "Methods for Language Learning Assessment at Scale: Dulingo Case Study". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 865-871. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

term learning. These models predicted future performance on exercise x_{t+1} given the previous performance record for a student (x_0, ..., x_t). This system treats every student interaction as an opportunity for assessment and the model output was used for developing student-facing modules for progress tracking and content recommendation. However, knowledge tracing approaches primarily focus on characterizing mastery of specific concepts rather than providing a holistic assessment of knowledge or achievement.

Other studies rely both on knowledge tracing and assessment data to analyze course effectiveness and provide this more holistic view. This approach is especially useful in more self-directed learning platforms. One study [4] used BKT to characterize learning using a digital game and used outputs from these models to predict post-test scores following a period of learning with the game. They found that mastery scores for two knowledge components (output from BKT models) had positive and significant association with post-test scores. Insights from the BKT model itself were also useful for identifying concepts that are difficult for students to master, which highlights opportunities for improving course effectiveness. This study also found evidence that learners have poor meta-cognition about their mastery of key concepts; when left to use the learning platform freely, many students continue to practice concepts the BKT model predicts they have mastered rather than moving on to new material.

Knowledge tracing is not the only approach used for characterizing student behavior using clickstream or log data. To make log data useful for predictive modeling, many researchers turn to methods from NLP to aggregate events [1, 12]. Simple methods include calculating n-grams for particular event types. For example, unigrams can capture the number of times a student completes a particular learning module and bigrams can capture the number of times students complete two modules in sequence [12]. Such data can be used as inputs into predictive models either relying solely on raw n-gram counts [12] or by processing the data further using unsupervised machine learning methods - such as hierarchical clustering - to identify common sequence patterns [1].

3. DUOLINGO ASSESSMENT SYSTEMS

3.1 Duolingo Course Structure

Duolingo courses are organized into a series of *units*, each of which concludes with a *Checkpoint*. Courses used by the majority of learners have the following structure: 25-30 *skills* per unit with five difficulty *levels* per skill and 5-6 *lessons* per level. Skills are designed around a particular theme (e.g., Travel). The vocabulary taught in the skill is aligned around that theme (e.g., hotel, airport, passport) and grammatical topics tend to be consistent across lessons within a skill. Lessons typically consist of 12-15 exercises designed to teach some vocabulary and/or grammatical concept. Duolingo curriculum designers incorporate aspects of spiral curriculum [5] to revisit familiar concepts in more complex contexts in future skills. See Figure 1 for an example of the typical Duolingo course structure.

The five levels for each skill provide a scaffolded learning experience, where learners review the same vocabulary or grammatical concepts in increasingly difficult contexts. All skills start with a foundational Level 0 and as learners "level up" a skill they see the same sequence of lessons teaching the same content but using different exercise types. Early levels include exercises that focus on passive recognition, such as matching a second language (L2) word/picture pair with the corresponding word in the first language (L1); see Figure 2). Exercises in later levels are more difficult, as they require recall and production in the L2 (e.g., translating an L1 sentence into L2; see Figure 2). The level achieved for a given skill is indicated in the user interface with a number inside a crown icon (see Figure 1).

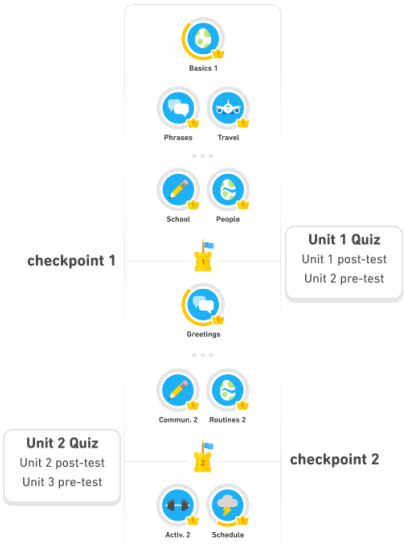

Figure 1. Duolingo course and Checkpoint Quiz design.

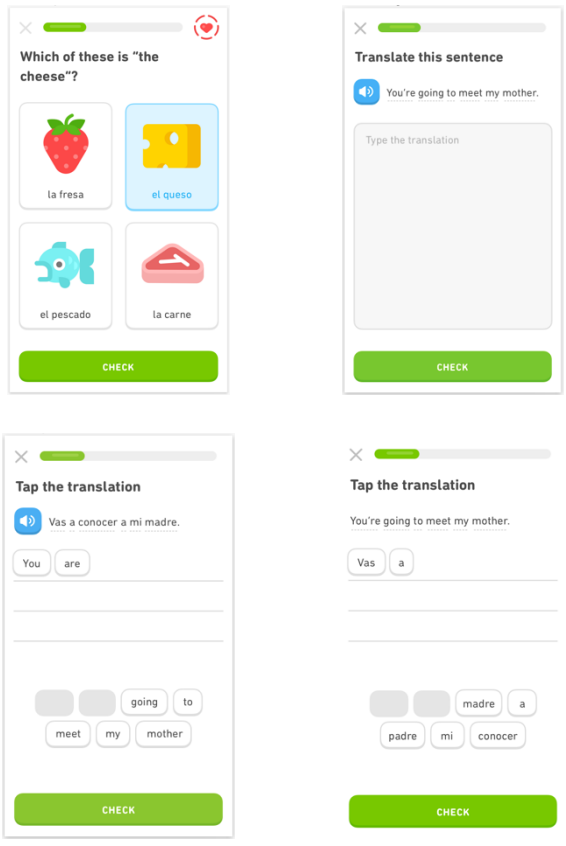

Figure 2. Example exercise types. Top left: passive recognition; top right: recall and production. Bottom left: recall L2→L1; bottom right: recall L1→L2.

When learners begin a Duolingo course, not all skills in the first unit are immediately available; a row unlocks once Level 0 is complete for all skills in the prior row. For example, only the Basics 1 skill is available at first and the next set of skills in the row below Basics 1 (e.g., Phrases and Travel; see Figure 1) will only unlock once Basics 1 reaches Level 1. Once skills are unlocked, learners are free to return to them to practice previously studied material and "level up" the skill. Duolingo learners are, therefore, given agency to choose their learning path. Some learners prefer to attempt only the foundational level in a skill (Level 0) before moving on to new material, while others prefer to level up all skills. Leveling up is entirely optional and learners are required to complete only the foundational level for each skill before they can move on to the next unit of content. This self-directed nature of the learning platform provides challenges for assessing learner achievement.

Other modes of learning are available to users outside of course skills. Learners can build reading and listening proficiency through the Stories feature, which reinforces unit content through interactive dialogues with exercises to check comprehension. Learners can also complete generalized practice sessions, which drill users on content they have already studied from throughout the course. Further, after learners have leveled a skill up all the way, they can return for skill practice to reinforce their knowledge. If learners find skill material too easy, they also have the option to "test out" of a level and jump to harder exercises at the next level.

We use a variety of methods to assess learner achievement and proficiency throughout a Duolingo course. In the sections below, we describe two of the core assessments in use today: Checkpoint Quiz and Review Exercises.

3.2 Checkpoint Quiz

For a subset of Duolingo's courses, learners must complete a custom-built assessment once they finish a unit and reach a Checkpoint. The Checkpoint Quiz is an achievement test that measures the extent to which our learners have achieved the objectives for each unit of a course. Checkpoint Quiz items are independent from the items used in course skills and users are only exposed to the quiz items during the assessment. This ensures that learners do not have the opportunity to learn the items in the assessment while studying course content and is important for test validity. Checkpoint Quiz items were designed by curriculum experts and Duolingo assessment scientists have conducted analyses to ensure their quality.

Learners do not receive corrective feedback or a final grade for the assessment and may only take the quiz once. At each Checkpoint, learners complete a randomly generated quiz consisting of 15 items (sampled from a larger pool of items). Seven items are pre-test items that test the next unit of the course that the learner is about to start and another seven are post-test items that test the unit the learner just completed (critically, the same seven items the learner saw in the previous quiz as a pre-test). This pre-test / post-test design allows us to establish a baseline level of performance so we can later assess gain in accuracy from pre-test to post-test. The final item is a self-directed writing item designed to assess the current unit (with no pre-test). See Figure 1 for an illustration of Checkpoint Quiz design.

The assessment tests knowledge of vocabulary, grammar, listening comprehension, reading comprehension, and free-form writing using separate items designed to test one of these language skills and components. Vocabulary and grammar items are a combination of multiple choice and fill-in-the-blank questions (i.e., learners type the missing word), listening and reading are exclusively multiple choice, and writing questions are free-response. Each item is accompanied by a set of curated tags for grammatical concepts and communicative components.

3.3 Review Exercises

Review Exercises prompt learners to review content from a skill earlier in their course. A single Review Exercise is inserted into randomly selected lessons in the foundational level of a skill (only for skills beyond the first five in the course). These exercises are randomly and uniformly sampled from the pool of available exercises from either three skills or five skills earlier in the course. For example, randomly selected exercises from the Animals skill are injected into Level 0 lessons seen by learners studying the Places skill (see Figure 3). These exercises are inserted into the lesson in a random position, as long as it is not among the first two or last two exercises. Therefore, lessons with Review Exercises will be one exercise longer than a standard lesson. Review Exercises come in two forms: assisted recall and translation from L1-to-L2 or vice versa (see bottom row of Figure 2).

Figure 3. Review Exercise design for testing five skills earlier in course.

Review Exercises as a form of assessment have a number of advantages over the Checkpoint Quiz: 1) Review Exercises are available in all courses; 2) they allow us to measure learning at every skill in a course, rather than just at unit-terminal Checkpoints; and 3) they provide an order of magnitude more data than Checkpoint Quizzes.

However, Review Exercises have a few disadvantages over the Checkpoint Quiz. One key disadvantage is that the items used for Review Exercises overlap with items used for lessons with skills; therefore, we sacrifice some test validity in order to be able to use the assessment at scale across all courses and all skills in a given course. Further, the sentences used as Review Challenges have not been assessed for their quality as measures of learning. Another disadvantage is that the data is not tagged for grammatical concepts or communicative components, which

limits the insights this assessment can provide for informing curriculum design.

Table 1. Key differences between the Checkpoint Quiz and Review Exercises.

Checkpoint Quiz	Review Exercises
Slow data collection (only at Checkpoints)	Fast data collection (at every skill in a course)
Tagged and calibrated by curriculum experts	Not tagged or calibrated
Items siloed from course	Items sampled from course
Only certain courses	All courses

4. CASE STUDY

Learners on Duolingo use the platform in a variety of different ways. In this case study, we investigate how learning decisions impact outcomes, so that we could "nudge" learners to use the app more effectively.

This case study demonstrates how EDM methodologies allow us to investigate the various ways that learners choose to navigate through the platform - focusing on differences in "leveling up" behavior - and how course navigation relates to learning outcomes. We show a correlation between leveling up and higher accuracy on the Checkpoint Quiz. Complementary modeling with Review Exercise data establishes a causal link between completing sessions in higher levels and accuracy on assessments.

4.1 Checkpoint Quiz

4.1.1 Data

Our work uses four months of Checkpoint Quiz data. For every learner completing at least two consecutive Checkpoint Quizzes within this timeframe, we collected the pre-test / post-test item response pairs (e.g., the pre-test responses collected at Checkpoint 1 and the corresponding post-test responses collected at Checkpoint 2) as well as summary statistics on learners' studying behavior in the unit the items assess (e.g., number of lessons completed at each level across skills in Unit 2, number of Stories completed between pre-test and post-test). Responses to free-form writing items were not included in this analysis.

4.1.2 Methods

To isolate the impact of lessons completed at each level on Checkpoint Quiz outcomes, we built a logistic regression model to predict post-test scores for items that were answered incorrectly in the pre-test (a measure of learning gain). Primary variables of interest capture the number of lessons learners completed at a given level for each skill in the unit of interest (frequency counts for Level 1 through Level 4; e.g., a learner completed 20 Level 1 lessons, 15 Level 2 lessons, etc.). Although Duolingo has five levels for all skills (starting with Level 0), we exclude counts for the foundational level because all learners must complete the same number of Level 0 lessons to finish a unit. The model controls for item and user covariates: language component of the item (e.g., vocabulary), unit (e.g., Unit 2), course (e.g., French for English Speakers), number of sessions completed for other types of study material (e.g., Stories, generalized practice, test-outs), self-reported prior proficiency (0-10), and subscriber status[1] (non-paying or paying learner).

4.1.3 Results

We found that average post-test item accuracy increases linearly with every skill-level completed (Figure 4). In other words, each additional level completed across all skills increases the odds of answering a Checkpoint Quiz item correctly by the end of the unit.

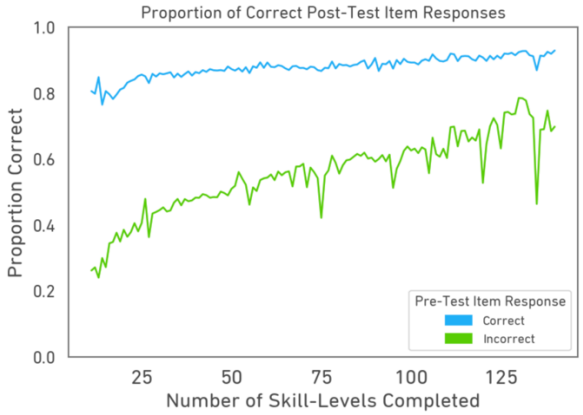

Figure 4. Average post-test accuracy by the number of skill-levels completed as a function of pre-test accuracy.

This finding was supported by the results of our logistic regression model (summarized in Figure 5). We observed that the probability of answering a post-test item correctly increases with every additional lesson in Levels 1, 2, and 4. Level 3 has a negative coefficient, but this is likely an artifact of variable suppression[2].

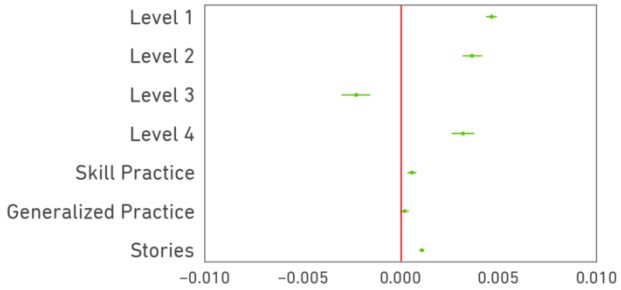

Figure 5. Checkpoint Quiz logistic regression model output. Coefficients of the number of times a user completed seven different session types in a model including other user and item covariates (see Section 4.1.2).

[1] Duolingo offers a paid subscription that removes ads, allows offline access, and includes additional features and learning modes. All learners have access to the same course content.

[2] Because learners tend to complete the same number of lessons in Levels 3 and 4, we attributed the negative coefficient to the statistical consequence of highly collinear relationships existing in the correlation matrix, which can cause variable suppression and model instability [8]. To verify that this multicollinearity did not result in model instability, we repeatedly fit the model on bootstrapped samples of the original data. We found that small changes to the data do not cause any erratic changes in the coefficients, so we concluded that our model estimates are stable.

We also compared the magnitudes of the leveling up effects with those of other types of learning modes, specifically Stories (interactive dialogues to practice reading and listening skills), skill practice, and generalized practice (see Section 2 for more details about these learning modes). Coefficients capturing leveling up behavior show dominant effects in the model; one additional skill-level has a greater impact on Checkpoint Quiz scores than one additional Story, skill practice, or generalized practice.

The Checkpoint Quiz findings show that providing learners with multiple difficulty levels to practice study material improves learning outcomes. Further, we found evidence that completing lessons at Levels 1, 2, 4 is not only positively associated with learning outcomes, but is *more* positively associated than any other activity. However, the Checkpoint Quiz analysis is not necessarily causal. The findings could also be due to self-selection biases, wherein the type of learner that is motivated to complete additional (non-required) levels is likely to perform better in general. A complementary analysis is required to establish a causal link.

4.2 Review Exercises

We utilized Review Exercise data to establish a causal link between leveling up and better learning outcomes. Review Exercises are better suited to this complementary analysis than the Checkpoint Quiz because each Review Exercise targets material from a single source lesson. This design allows us to compare learners who exhibit the same studying behavior except for the completion of one additional level for that lesson. Isolating the change in accuracy from one additional level means that we have controlled for self-selection biases and can interpret the change as causal.

4.2.1 Data

For the Review Exercise analysis, we collected all Review Exercises completed over the course of approximately two months. Data comes from all Duolingo courses. Along with Review Exercise response accuracy, we collect important control variables: whether the exercise came from 3 or 5 skills earlier in the course, exercise type, and the skill the exercise was sampled from (see Figure 3 for Review Challenge design).

4.2.2 Methods

Using logistic regression and a regression discontinuity design (RDD) [3, 6], we are able to model the impact of completing higher levels on Review Exercise accuracy while controlling for self-selection bias that may occur for learners who choose to level up vs. those who do not. An RDD is a quasi-experimental approach where a synthetic treatment condition is assigned to observations that fall above or below a certain "cut-off" point. We achieve this by first identifying learners who have completed any lessons at a given level for the skill a Review Exercise was sampled from (e.g., learners who have completed at least one Level 1 lesson). Among those learners, we define a cut-off point to compare those who have completed that level for the Review Exercise source lesson (e.g., Level 1) to those who have completed that level for the lesson that *immediately precedes* the source lesson but who have not yet completed that level for the source lesson itself (e.g., preceding lesson to Level 1, but source lesson to Level 0). This approach controls for most potential self-selection bias in deciding to level up (all comparisons include learners who have chosen to level up the skill) and can provide stronger evidence for a causal relationship between leveling up and Review Exercise accuracy.

We created a variable with eight levels for use in the regression model to capture 1) the highest level a learner has leveled up the Review Exercise source lesson to and 2) whether the learner studied the source lesson to the same level as the preceding lesson (e.g., both at Level 1) or studied the source lesson one time less than the preceding lesson (e.g., preceding lesson at Level 2 but source lesson at Level 1). For example, this scheme yields coefficients of the form `Level 1:Same Level`, indicating learners for whom both the source lesson and preceding lesson were at Level 1, or `Level 1:Lower Level`, indicating learners for whom the source lesson was at Level 1 and the preceding lesson was at Level 2. This coding scheme required excluding certain observations. Cases where the learner had completed the highest level possible for the Review Exercise source lesson (i.e., Level 4) is not included because it is impossible for the lesson preceding the source lesson to be leveled up any higher. We also exclude observations where the source lesson is the first lesson of a skill because there will be no preceding lesson to serve as a control comparison.

In addition to this main variable, we also control for other factors that influence Review Exercise accuracy: the number of skills away from the source skill (three or five), and the exercise type of the Review Exercise (L1-to-L2 translation or vice versa), and the difficulty of the source skill. We defined difficulty of source skills by computing the log-odds of answering a Review Exercise correctly in each skill in the data overall[3]. This allows us to control for the fact that, all else being equal, accuracy is likely to be lower overall for Review Exercises sampled from more difficult skills, which increases the power of the analysis.

4.2.3 Results

If leveling up causes higher Review Exercise accuracy, we expected to see that the `Level N:Same Level` (source lesson and preceding lesson to Level N) coefficients were significantly larger than the `Level N-1:Lower Level` (source lesson one level lower than preceding lesson; Levels N-1 and N, respectively) coefficients. Such an effect would indicate that - controlling for leveling up behavior overall - completing higher levels of the lesson a Review Exercise came from yields significant improvements in Review Exercise accuracy.

Figure 6 summarizes the results of our logistic regression model. We can see that `Level 1:Same Level` is significantly higher than `Level 0:Lower Level`. This effect indicates that learners who have studied a Review Exercise source lesson twice (at Level 0 and Level 1) are more likely to provide a correct response on their Review Exercise than learners who have studied a Review Exercise source lesson once (only at Level 0) but already had studied the previous lesson twice (at Level 0 and Level 1). This result provides evidence for a causal relationship between leveling up study material and assessment performance, at least for the first time learners level up. The model shows similar trends for leveling up beyond Level 1 (e.g., `Level 2:Same Level` is numerically higher than `Level 1:Lower Level`), suggesting this relationship continues to exist as users study the Review Exercise source lesson additional times (although perhaps with diminishing returns).

The regression results also show significant differences between `Level 0:Same Level` / `Level 0:Lower Level` and

[3] Empirical log odds defined as *log((correct + 1) / (incorrect + 1))*.

`Level 1:Same Level` / `Level 1:Lower Level`. Although the learners captured in the `Lower Level` coefficients had not leveled up the source lesson to Level 1, we see clear improvements in Review Exercise accuracy stemming from leveling up any lessons preceding the source lesson. These learners will not have had additional opportunity to study the exact exercise used for the Review Exercise, but the content and concepts in other lessons in the skill will have been related. Therefore, the benefit of studying in one lesson transfers to other lessons.

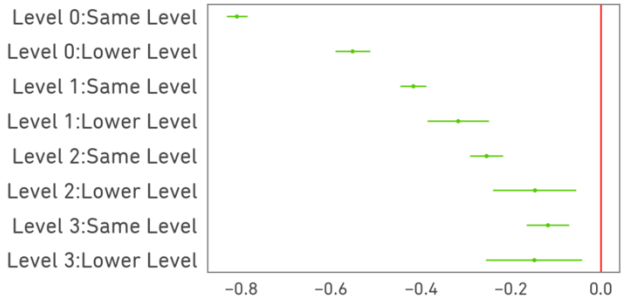

Figure 6. Review Exercises model output. Coefficients of leveling up behavior in a model including other item covariates (see Section 4.2.2).

5. CONCLUSIONS

In a case study of the levels mechanic, wherein learners study content in increasingly difficult contexts by "leveling up", complementary analyses of the Checkpoint Quiz and Review Exercises showed that completing sessions in higher levels leads to stronger performance on assessments. Analyzing accuracy rates on the Checkpoint Quiz by the number of skill-levels completed in the course unit revealed a strong positive trend. Because variation in how learners navigate the platform may introduce self-selection bias and complicate interpretation of these results, we conducted an additional analysis of Review Exercises that controlled for this bias. The Review Exercises analysis supports a causal link between leveling up and improved assessment performance, showing that completing additional levels for a skill (beyond the foundational level) has measurable learning value.

Together, these results directly motivated the implementation of a number of interventions that encourage learners to reach higher levels. For example, because learner awareness of the existence and purpose of levels was relatively low, we added design elements that give learners a visual stand-in for how the levels system works. Learners also now receive a pop-up with a redirect button upon finishing a level prompting them to start the next level in the skill. Randomized controlled experiments (i.e., A/B tests) introducing these changes showed >10% increases in the number of lessons completed in each level beyond the required foundational level and significantly more studying activity on the app overall. These interventions exemplify how insights from the Checkpoint Quiz and Review Exercises have lasting impact on the Duolingo learning experience.

This study focused on one type of variation in how learners choose to navigate the Duolingo learning platform, namely leveling up. Learners can additionally choose their own study sequence for the skills (e.g., completing all the levels in a skill before starting the next skill, completing the entire course unit one level at a time, leveling up clusters of skills within a unit), as well as which types of learning material to study (e.g., course skills, generalized practice, Stories). Future iterations of this work will aim to capture such variation, thereby improving model fit and deepening our understanding of how other types of navigational choices relate to learning outcomes. Previous EDM studies [1, 9] provide methodologies that can be used to characterize this variation.

Future work will also continue to explore the utility and limitations of the Review Exercise assessment system. For example, data from Review Exercises show promise as a method for measuring learning improvements over the course of an A/B test due to the high volume of daily data generated, highly localized measurement (i.e., testing learning of content from specific course skills), and the distributed nature of the assessment (i.e., testing learning in all course skills). Future work could also consider whether Review Exercise accuracy can be predicted based on engagement with (and accuracy on) source lessons in the past.

Self-directed learning platforms such as Duolingo require accurate and well-controlled assessments to measure learner achievement. Because learners exercise a high degree of agency in how they navigate the courses, achievement cannot be adequately assessed by analyzing exercise responses alone. Duolingo developed two forms of assessment - the Checkpoint Quiz and Review Exercises - to capture insights about how different study approaches relate to learning outcomes. Applying EDM techniques to these assessments yields useful insights that inform our understanding of how the navigation of course content relates to learning outcomes and how we can leverage these insights to improve the learning experience on the platform.

6. ACKNOWLEDGMENTS

Special thanks to Daniel Falabella, Xiangying Jiang, Geoff LaFlair, Bozena Pajak, and Karin Tsai for helpful comments on this work.

7. REFERENCES

[1] Nil-Jana Akpinar, Aaditya Ramdas and Umit Acar. 2020. Analyzing Student Strategies in Blended Courses Using Clickstream Data. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020),* July 10-13, 2020, 6-17.

[2] Hanan Aldowah, Hosam Al-Samarraie, & Wan Mohamad Fauzy. 2019. Educational data mining and learning analytics for 21st century higher education: A review and synthesis. *Telemat Inform*, 37 (Apr. 2018), 13–49. https://doi.org/10.1016/j.tele.2019.01.007

[3] Joshua D. Angrist & Jörn-Steffen Pischke. 2014. *Mastering Metrics: The Path from Cause to Effect*. 2014. Princeton University Press, Princeton, NJ.

[4] Huy Anh Nguyen, Xinying Hou, John Stamper, & Bruce M McLaren. 2020. Moving beyond Test Scores: Analyzing the Effectiveness of a Digital Learning Game through Learning Analytics. In *Proceedings of the 13th International Conference on Educational Data Mining* (*EDM 2020)*, July 10-13, 2020. 487–495.

[5] Jerome S. Bruner. 1960. *The Process of Education*. Harvard University Press, Cambridge, MA.

[6] Thomas D. Cook, Donald T. Campbell, & William Shadish. 2002. *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin, Boston, MA.

[7] Albert T. Corbett & John R. Anderson. 1995. Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User Model User-Adapted Interaction* 4, 4 (March 1995), 253–278.

[8] Lynn Friedman & Melanie Wall. 2005. Graphical Views of Suppression and Multicollinearity in Multiple Linear Regression. *Am Stat* 59, 2, 127-136. https://doi.org/10.1198/000313005X41337

[9] Tao Huang, Zhi Li, Hao Zhang, Huali Yang, & Hekun Xie. EAnalyst : Toward Understanding Large-scale Educational Data. In *Proceedings of the 13th International Conference on Educational Data Mining (EDM 2020)*, July 10-13, 2020, 620–623.

[10] Zacharoula Papamitsiou, & Anastasios A. Economides. 2014. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology and Society* 17, 4, 49–64.

[11] Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, & Mark Warschauer. 2017. Detecting Changes in Student Behavior from Clickstream Data. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference (LAK 2017)*, March 2017, 21-30. https://doi.org/10.1145/3027385.3027430

[12] Bertrand Schneider, & Paulo Blikstein. 2015. Unraveling Students' Interaction Around a Tangible Interface Using Multimodal Learning Analytics. *Journal of Educational Data Mining* 7, 3, 89-116. DOI: https://doi.org/10.5281/zenodo.3554729

[13] Wanli Xing & Dongping Du. 2019. Throughput Prediction in MOOCs: Using Deep Learning for Personalized Intervention. *J Educ Compt Res* 57, 3, 547-570.