Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

SLP: A Multi-Dimensional and Consecutive Dataset from K-12 Education

Yu LU[a], Yang PIAN[a*], Ziding SHEN[b], Penghe CHEN[a], Xiaoqing LI[a]
[a]*Advanced Innovation Center for Future Education, Beijing Normal University, China*
[b]*University of California, Los Angeles, USA*
* bianyang@mail.bnu.edu.cn

Abstract: Learning is a complicated process jointly influenced by multiple factors, such as learner's personal characteristics, family background and school environment. However, the existing public datasets in K-12 education domain seldom fully cover the heterogeneous dimensions, which greatly hinders the research on fully analyzing and understanding the learners and their learning process. In this work, we report a dataset that includes the learners' demographic information, psychometric intelligence scores as well as their family-school background information. Furthermore, the dataset records the learners' academic performance data on 8 different subjects in 3 consecutive years. This multi-dimensional dataset from K-12 education can be a valuable information source for learning analytics and would benefit the cross- disciplinary research in education on a broader canvas. The dataset has been publicly available for the research purpose at https://aic-fe.bnu.edu.cn/en/data/index.html.

Keywords: Educational dataset, learning analytics, K-12 education

1. Introduction

Driven by the fast advancements of online learning platforms and educational data collection techniques, learner modelling and learning analytics (Ferguson, 2012) have recently attracted much attention from both academia and industry. Proper collection and analysis of educational data have then become crucial, as they provide opportunities to fully understand the human learning process and precisely identify its underlying key factors, which may ultimately help improve human's learning experience, efficiency and efficacy.

Specifically, the Massive Open Online Course (MOOC) platform and Intelligent Tutoring System (ITS) (Polson & Richardson, 2013) have significantly facilitated the collection of learner's online assessment data and their interaction with the corresponding virtual learning environment. For example, the ASSISTments dataset (Selent, Patikorn & Heffernan, 2016) includes the assessment data of secondary school students and the log data recording their interactions with an ITS for math study. The KDD Cup dataset (Qiu et al., 2016) contains the MOOC learners' assessment and interaction data on 40 different online courses. The EdNet dataset (Choi et al., 2020) gathers learner's online interaction in 4 different levels of abstractions from a multi-platform service. The Open University Learning Analytics Dataset (OULAD) (Kuzilek, Hlosta & Zdrahal, 2017) introduces learner's demographic information, including gender, geographic region and education background. The Programme for International Student Assessment (PISA) dataset (OECD, 2015) covers the participants' assessment data, and school-family information, which mainly used the one-time test (performed every three years on 15-year-old students) without continuous tracking on the participants. A public data repository, called PSLC DataShop (Stamper, 2010), has been established to host the education-related datasets, particularly for the learning interaction data.

However, there seldom exists dataset that provides multiple heterogeneous dimensions and could be directly employed for interdisciplinary research purpose. Past literature has already demonstrated that psychometric intelligence is a strong predictor of student's scholastic success, which can be measured by the Intelligence Quotient (IQ) test (Downey et al., 2014; Eccles & Harold, 1993). Several external factors have then been proven to influence IQ score, in which family and parental

factors might directly impact it, especially for K-12 learners (Haimovitz & Dweck, 2016; Melby et al., 2008). For example, students who are from a higher socioeconomic background, a smaller-sized family or with a better educated parent, might be more intelligent and perform better at school (Von Stumm & Plomin, 2015). Lack of such heterogeneous dimensions could hamper the further exploration on learning analytics, and thus would possibly prevent researchers from discovering the essential laws of education.

To fully cover all these heterogeneous dimensions, our dataset collected from an online learning platform called smart learning partner (SLP), intentionally recorded the learner's data from the five dimensions above to provide a wealthy content for learning analytics and educational data mining. It thus has two unique characteristics:

- It explicitly covers the data from five different dimensions, namely student demographic information, psychometric intelligence information, academic performance information, family information and school information;
- It automatically captures the learner's academic performance data during their three-year study (mainly from 7th grade to 9th grade) on 8 different subjects, namely Math, Physics, Chemistry, Biology, History, Chinese, Geography and English.

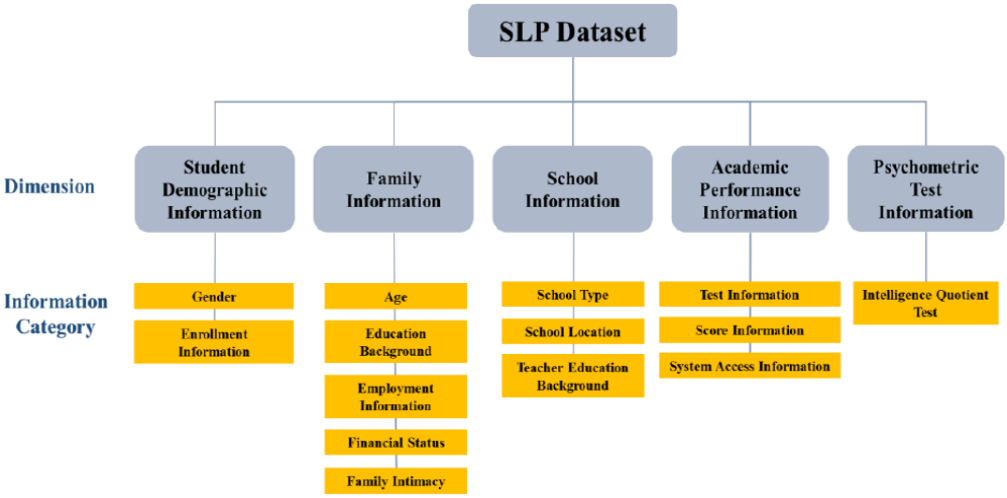

Figure 1. Overall Structure of SLP Dataset.

Figure 1 gives the overall structure of the SLP dataset. Briefly speaking, the SLP dataset are collected from 4830 students in 32 local secondary schools. The entire data collection process has lasted approximately three years (from November 2016 to June 2019) by leveraging on an online learning platform, called SLP, which currently serves more than 140,000 secondary school students. The students employ this platform to conduct regular unit tests and term tests since their first year in secondary school (i.e., 7th grade) to the last year (i.e., 9th grade). Up to June 2019, a total of 832 unit tests and 54 term tests on 9 different subjects have been designed and implemented on the platform. Specifically, each unit test is associated with a knowledge concept in a given subject, while each term test normally covers the knowledge concepts that have been learned during the entire semester. In addition, 14,027 learning resources (mainly in the form of micro-lecture, i.e., a short video in 5 to 10 minutes) would be automatically recommended to students based on their unit test or term test results.

The SLP dataset could be used in different studies, such as predicting student's academic performance in a fine-grained manner (Chen et al., 2018), or evaluating the influences of pertinent factors (e.g., intelligence or family factors) on different subject learning. Our aim is to encourage researchers from diverse fields to engage in understanding and modelling learners in K-12 education.

2. Data Collection Workflow

Figure 2 depicts the data collection process from different sources: the student basic information and the school information were directly acquired from the local education bureau, mainly consisting of the student demographic and enrolment information as well as the corresponding school information. Such information was also utilized to create and validate the individual user's account on the online learning

platform, which guaranteed all the platform users were the authenticated local students. When a student logged into the platform for the first time, he/she was required to complete an online survey to collect his/her family information, where the questionnaire contains the questions about family intimacy, parent education background and other pertinent family information. The first-time login and the data collection process were usually guided by local teachers during class time, and it thus, to a certain extent, ensured the high quality of the collected family information data.

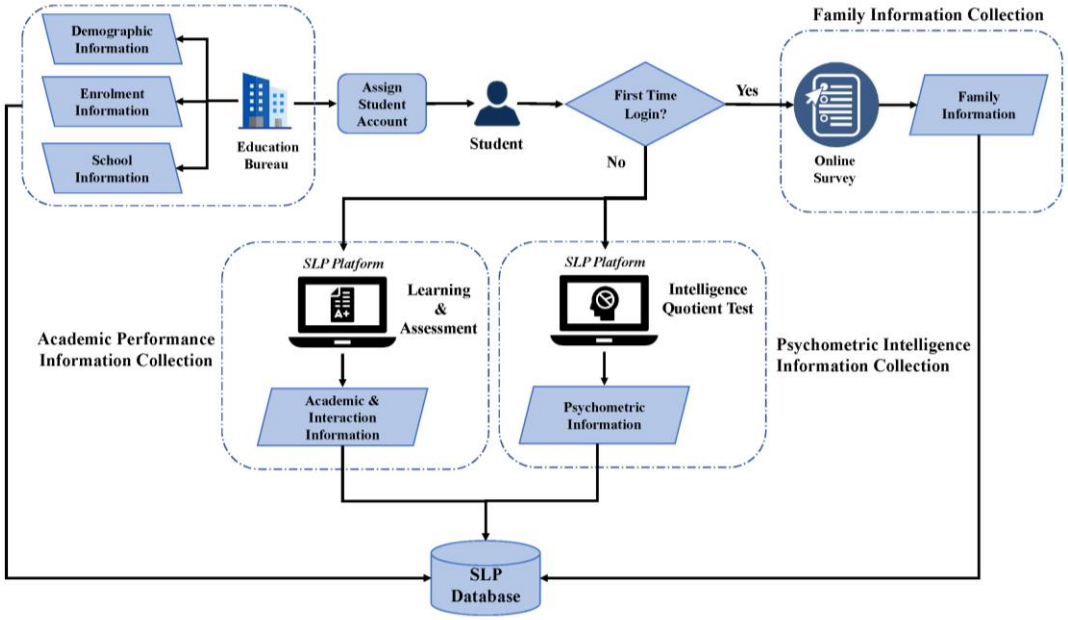

Figure 2. SLP Data Collection and its General Workflow.

2.1 Assessment Data Collection

After the first-time login and the online family survey accomplished, students were asked to regularly use the platform to conduct the assessment tests, and to freely choose the personalized learning resources on different subjects. All of the test results, including both unit tests and term tests, and the important interaction log data are automatically recorded and stored by the platform. A unit test normally contains 9 questions on the same knowledge concept and mostly in the form of multiple-choice questions (MCQ) or short-answer free text questions. When students were doing a unit test, the platform recorded students' access time. Most of the unit tests could be auto-scored by the system. On the other hand, the term test typically contains more questions on multiple knowledge concepts in a similar form, and the access time was also recorded. The SLP dataset includes all the above information, as well as the name of the knowledge concept tested by each question and the name of the corresponding subject.

2.2 Psychometric Intelligence Data Collection

On the SLP platform, students were offered with multiple psychometric tests and could choose to take any of these based on their preference. In the current SLP dataset, 1161 students participated and responded to one part of the IQ test online. The test was based on the classical Raven's Progressive Matrices Test (Raven & Court, 1938) with minor modifications to fit with the cultural and linguistic context. Specifically, students answered a 40-item (α = .80) question set. Scores were then computed based on norms, and the total score was 140 with 70 as the passing line. A high score would indicate a mastery of reasoning ability, demonstrating the student's ability to deduce, summarize, and exchange the pattern with others solely based on the provided information.

2.3 Data Usage and Privacy Issues

Before the data collection process, all the participants, including the students, their parents and schools, have been explicitly informed that the collected data on the platform would be directly used for the

research purpose and shared publicly. All the parents were given the option to opt out of the data sharing before they signed the data usage agreement. After the data was submitted, we also carefully notified each parent via both Email and SMS, and removed the ones who felt uncomfortable to be involved in the SLP dataset. To further protect the privacy of the opt-in participants, we had irreversibly anonymized their identity, and randomly sampled a subset of the participants.

3. Data Description

The SLP dataset is available in the form of multiple CSV files (value is separated by comma, and the first line in each table shows the column name), and can be accessed from the website[1]. Figure 3 illustrates tables and corresponding data fields in the SLP dataset, which would be elaborated as below.

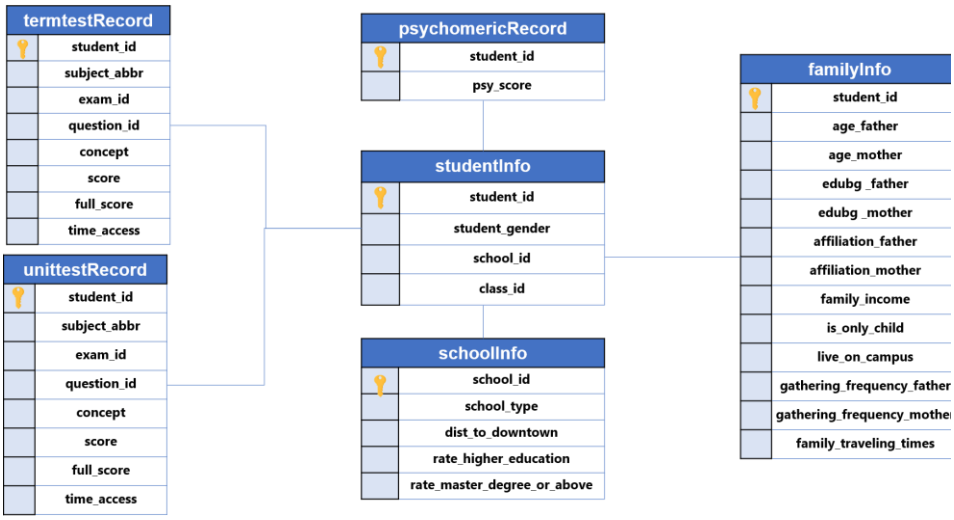

Figure 3. Tables and Data Fields of SLP Dataset.

- **Academic Performance Information**: The two file folders, named "termtestRecord" and "unittestRecord", contain multiple tables with the identical table structure as above, which provide the student academic performance information. Specifically, these two folders collect the term test and unit test data respectively. In each folder, each sub-table contains individuals' subject score information. For example, the sub-table "termtestRecord-BIO" inside the folder "termtestRecord" stores the term test results of Biology, which currently consists of 500,562 rows. Similarly, the sub-table "unittestRecord-MATH" inside the folder "unittestRecord" keeps the unit test results of Math, which currently consists of 57,244 rows. Note that multiple concepts may be associated to one question, which would be separated by semicolon, and all the associated concepts are explicitly shown in the table (e.g., "line segment" or "intersecting lines" in math questions, "digestive system" or "urinary system" in biology questions).
- **Psychometric Test Information**: Table *psychomericRecord* contains the student's psychometric intelligence score, and it consists of 1161 rows. Note the SLP dataset only includes participants who agreed to take and share their psychometric test results.
- **Family Information**: Table *familyInfo* contains student's family information, including parents' age, education background, employment information, financial status and family intimacy (e.g., how often the student meets with father or mother). The table consists of 3189 rows, and only includes the students who completed the family information survey and confirmed the truth of it.
- **School Information**: Table *schoolInfo* contains the school's demographic information, including the school type, location information and teacher's education background. It consists of 32 rows, and for teachers' education background, 3 school's data are missing (marked as n.a. in the table).
- **Student Demographic Information**: Table *studentInfo* contains the basic demographic information of students, including their gender, school and class ID. The table consists of 4830 rows, and all the school and class information are anonymized.

[1] https://aic-fe.bnu.edu.cn/en/data/index.html

4. Technical Validation

The SLP dataset contains learners' information from multiple dimensions, ranging from academic achievement, psychometric intelligence, to student's family and school information. We thus tentatively provide some illustrations and simple analysis results to validate the quality and value of the dataset, while deeper and broader studies could be further conducted.

We first looked at the SLP data from the perspective focusing on the relationship between family factors and intelligence. Figure 4 compares the mean of IQ test score within different family information categories. Each bar and the number above it represent the mean psychometric score in one group within one category. The analysis showed that there was a significant difference of psychometric intelligence score between students from one-child family (M = 87.40, SD = 9.99) and students from multiple-child family (M = 85.81, SD = 9.35). In addition, there was a significant positive correlation between students' psychometric test score and family's annual income (r = .07, n = 1161, p <.05), showing that a higher socioeconomic family background might be beneficial on children's intelligence development. More interestingly, when we analyzed the relationship between parents' education level and students' intelligence, only mother's education level was found to be significantly correlated with the psychometric test scores in a positive way (r = .12, n = 1161, p <.05), indicating that mothers with higher education background might have kids who also score higher in the intelligence test.

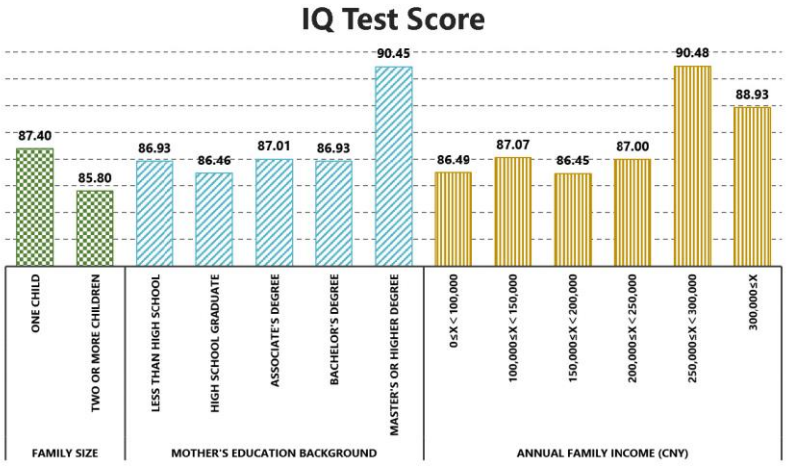

Figure 4. Mean Psychometric Test Scores Grouped by three Categories of Family Information.

Secondly, we focused on the relationship between school and academic performance. After removing the schools with data size smaller than 100 students, we portraited a scatter plot with different points representing each school's average term test score and the distance from school to downtown area, as shown in Figure 5. The line represents the significant negative correlation (r = -.87, n = 11, p <.05). This analysis result might reveal a unique pattern in China that most of the large and high-quality primary and middle schools aggregate around the center of the city.

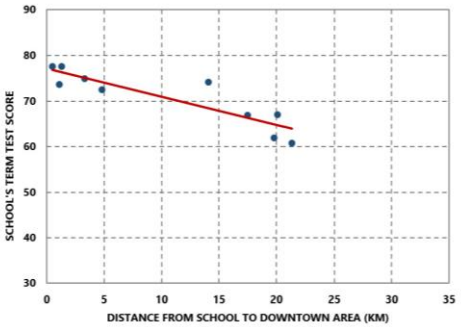

Figure 5. Scatter Plot of School's Term Test Scores and Its Distance to Downtown.

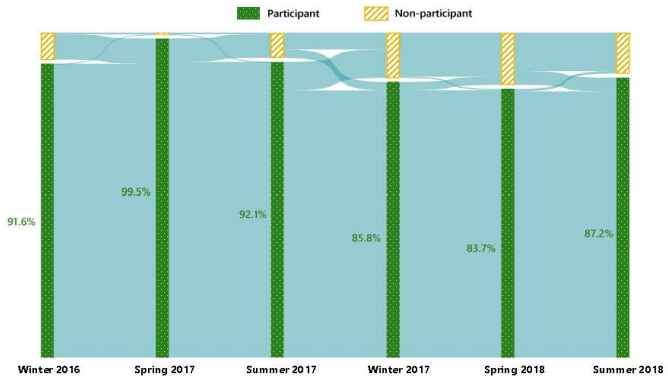

Figure 6. Sankey Diagram Describing the Term Test Participants' Flow Over Time.

Lastly, we validated the continuous collection of the academic performance data from the same student population across three years. Figure 6 shows the Sankey diagram that portrays how the student population has changed from 2016 to 2018 in 6 different term tests, where the width of the flow is proportional to the flow quantity. As could be seen from Figure 6, a majority of the students took all 6 term tests: for example, around 91.6% and 99.5% students took the term test in winter of 2016 and spring of 2017, whereas more than 90% of students took both tests. The diagram clearly validated the consistency of the academic performance data in terms of participants.

5. Conclusion

In this paper, we present an education-oriented 3-year consecutive dataset from k-12 learners. This dataset consists of multi-dimensional data including the learner's demographic information, academic performance, psychometric intelligence scores, and family-school background. We tentatively conduct some preliminary studies and provide some simple analysis results to validate the quality and value of the dataset. We believe the SLP dataset would provide valuable information and foster the cross-disciplinary research for learning analytics on a broader canvas.

Acknowledgements

This research is supported by the National Natural Science Foundation of China (No. 62077006, 61807003), and the Fundamental Research Funds for the Central Universities.

References

Chen, P., Lu, Y., Zheng, V.W., Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In *2018 IEEE International Conference on Data Mining (ICDM)* (pp. 39-48). IEEE.

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., ... & Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education* (pp. 69-73). Springer, Cham.

Downey, L. A., Lomas, J., Billings, C., Hansen, K., & Stough, C. (2014). Scholastic success: Fluid intelligence, personality, and emotional intelligence. *Canadian Journal of School Psychology*, *29*(1), 40-53.

Eccles, J. S., & Harold, R. D. (1993). Parent-school involvement during the early adolescent years. *Teachers college record*, *94*, 568-568.

Ferguson, R. (2012). Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5-6), 304-317.

Haimovitz, K., & Dweck, C. S. (2016). Parents' views of failure predict children's fixed and growth intelligence mind-sets. *Psychological science*, *27*(6), 859-869.

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. *Scientific data*, *4*(1), 1-8.

Melby, J. N., Conger, R. D., Fang, S. A., Wickrama, K. A. S., & Conger, K. J. (2008). Adolescent family experiences and educational attainment during early adulthood. *Developmental psychology*, *44*(6), 1519.

OECD. (2015). Programme for international student assessment (PISA). Retrieved April 30, 2021, from http://www.oecd.org/pisa/publications/.

Polson, M. C., & Richardson, J. J. (Eds.). (2013). *Foundations of Intelligent Tutoring Systems*. Psychology Press.

Qiu, J., Tang, J., Liu, T. X., Gong, J., Zhang, C., Zhang, Q., & Xue, Y. (2016). Modeling and predicting learning behavior in MOOCs. In *Proceedings of the ninth ACM international conference on web search and data mining* (pp. 93-102).

Raven, J. C., & Court, J. H. (1938). Raven's progressive matrices. Los Angeles, CA: Western Psychological Services.

Selent, D., Patikorn, T., & Heffernan, N. (2016). Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale* (pp. 181-184).

Stamper, J., Koedinger, K., d Baker, R. S., Skogsholm, A., Leber, B., Rankin, J., & Demi, S. (2010). PSLC DataShop: A data analysis service for the learning science community. In *International Conference on Intelligent Tutoring Systems* (pp. 455-455). Springer, Berlin, Heidelberg.

Von Stumm, S., & Plomin, R. (2015). Socioeconomic status and the growth of intelligence from infancy through adolescence. *Intelligence*, *48*, 30-36.