The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment

Conference Paper · April 2016

DOI: 10.1145/2883851.2883872

6 authors, including:

Korinn Ostrow
Worcester Polytechnic Institute
28 PUBLICATIONS **129** CITATIONS
SEE PROFILE

Neil T. Heffernan
Worcester Polytechnic Institute
235 PUBLICATIONS **4,630** CITATIONS
SEE PROFILE

Joseph Jay Williams
Harvard University
54 PUBLICATIONS **1,411** CITATIONS
SEE PROFILE

The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment

Korinn S. Ostrow, Doug Selent, Yan Wang,
Eric G. Van Inwegen, Neil T. Heffernan
Worcester Polytechnic Institute
Worcester, MA 01609
{ksostrow, dseslent, ywang14,
egvaninwegen, nth} @wpi.edu

Joseph Jay Williams
VPAL Research
Harvard University
Cambridge, MA 02138
joseph_jay_williams@harvard.edu

ABSTRACT

Researchers invested in K-12 education struggle not just to enhance pedagogy, curriculum, and student engagement, but also to harness the power of technology in ways that will optimize learning. Online learning platforms offer a powerful environment for educational research at scale. The present work details the creation of an automated system designed to provide researchers with insights regarding data logged from randomized controlled experiments conducted within the ASSISTments TestBed. The Assessment of Learning Infrastructure (ALI) builds upon existing technologies to foster a symbiotic relationship beneficial to students, researchers, the platform and its content, and the learning analytics community. ALI is a sophisticated automated reporting system that provides an overview of sample distributions and basic analyses for researchers to consider when assessing their data. ALI's benefits can also be felt at scale through analyses that crosscut multiple studies to drive iterative platform improvements while promoting personalized learning.

Categories and Subject Descriptors

K: Applications to Education. K.3: Computers and Education. I.2.2: Automatic Programming. G.3: Probability and Statistics.

General Terms

Measurement, Documentation, Experimentation, Standardization.

Keywords

Assessment of Learning Infrastructure, Automated Analysis, Randomized Controlled Experiments at Scale, The ASSISTments TestBed, Universal Data Reporting, Tools for Learning Analytics.

1. INTRODUCTION

An immense community of researchers, educators, and administrators seeks to enhance the effectiveness of educational practices. Those invested in K-12 education struggle not just to enhance pedagogy, curriculum, and student engagement, but also to harness the power of technology in ways that will optimize learning. Researchers often fall back on observational studies or turn to data mining large longitudinal datasets due to the difficulties inherent to conducting student-level randomized controlled experiments (RCEs) in authentic learning environments. Software for sharing educational data has driven tremendous progress in educational research and best practices. For instance, the Pittsburgh Science of Learning Center's DataShop [8], funded by the National Science Foundation, provides an extensive database of educational datasets for post hoc data mining and analysis. However, the pace and power of educational research would increase drastically if researchers had easier access to environments in which they could design, implement, and analyze hypothesis driven experiments. The RCE remains the "gold standard" in determining causal relationships and was referred to when the U.S. Department of Education advocated for K-12 schools to apply basic findings from cognitive science to improve educational practices [16]. Without the assistance of scalable technologies, it has been difficult for researchers to answer the call to conduct RCEs within authentic academic settings [6] due to the high cost of establishing and maintaining sample populations, the complications inherent to randomization at the teacher-level (i.e., vast samples are required), and the often invasive curriculum restrictions necessary to establish sound controls.

When designed with flexibility and collaboration in mind, online learning platforms offer a unique and scalable approach to educational research and data analysis. Users of online learning platforms (i.e., students and teachers) create hundreds of thousands of data points each day, with databases of rich learner information growing exponentially as platforms gain popularity and validity as powerful learning aids. Beyond achievement measures, these systems provide opportunities to collect information including (but not limited to) behavior and affect [2, 17], learning interventions within content or feedback [14, 15], and interactions between skill domains that help guide curriculum development [1]. Through flexibility in content design, manipulation, and delivery, researchers are able to tap into the elements that drive effective learning within authentic K-12 classroom environments. When content can be manipulated to include parallel assignments, fashioned as conditions within RCEs, researchers are able to determine best practices and work toward personalized learning. Further, designing these environments with the open, collaborative, and perhaps even competitive design of RCEs in mind can strengthen internal validity and promote open source data reporting for review and replication of findings upon publication [11]. By allowing data scientists, educational researchers, and K-12 educators to work collaboratively within online learning platforms, all are empowered to dynamically evaluate and improve the effectiveness of the platform and its content while fostering growth in learner analytics.

LAK '16, April 25-29, 2016, Edinburgh, United Kingdom
© 2016 ACM. ISBN 978-1-4503-4190-5/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883872

1.1 Research in the ASSISTments TestBed

ASSISTments is a unique online learning platform that was designed with educational research as one of its primary goals [5]. The platform is used for both classwork and homework by over 50,000 users around the world, and provides students with immediate feedback and rich tutorial strategies and teachers with powerful assessment through a variety of reports that pinpoint where students are struggling and empower data driven teaching [5]. Recent funding from the NSF has allowed ASSISTments to promote educational research at scale through the development of the ASSISTments TestBed (www.ASSISTmentsTestBed.org). External researchers can use the TestBed to embed studies within ASSISTments content and non-invasively tap into our user population at virtually no cost and in a fraction of the time previously required to run experiments within K-12 environments.

The process of conducting an RCE within the TestBed typically involves researchers modifying preexisting certified content to include treatment interventions and student-level random assignment. The latter feature makes the TestBed a unique and robust tool for conducting research; rather than delivering the same treatment condition to all students within a particular class, students in the same class will be randomly assigned to different conditions while participating in the same assignment (i.e., content, feedback, or delivery may vary from student to student). The library of certified ASSISTments content consists primarily of middle and high school mathematics skills, with content organized and tagged by Common Core State Standard [10]. However, this library has grown to include content in physics, chemistry, and electronics, and researchers are able to develop their own content for experimentation in other domains.

Figure 1 depicts a simple study design implemented within the ASSISTments TestBed. Inclusion of a student in this type of study is dependent on her ability to access video content (note that many schools block video servers like YouTube). When the student begins her assignment, she must first pass a "Video Check," or a standard problem that serves as password protection to study participation. If the student can access video, she enters the 'password' provided in the short clip as her answer, and her correct response serves as the "Then" in an "If-Then" routing structure. If the student enters anything other than the password as a response, she is provided a default assignment without video content and is not considered a study participant. While this process attempts to control for technical issues, it does not demand the fidelity of study participants (i.e., we cannot currently track viewing statistics for embedded videos). Upon being routed into the study depicted in Figure 1, students are randomly assigned into one of two conditions using a "Choose Condition" routing structure. Note that although two conditions are presented here for simplicity, the system is able to compare any number of conditions. The platforms approach to random assignment will be discussed further in Section 3.1.2.

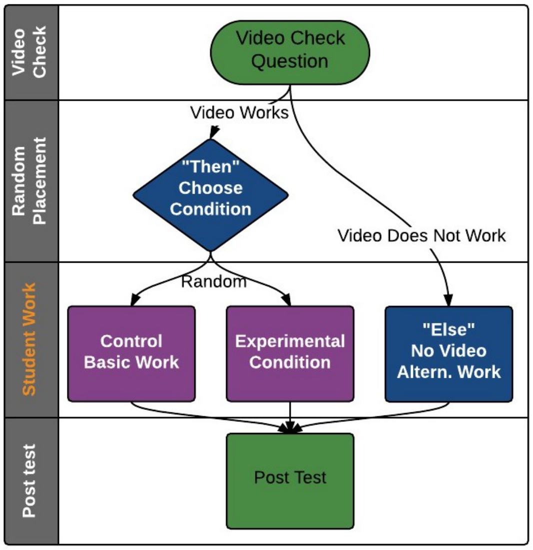


Figure 1: A simple research design that can be built using the ASSISTments TestBed to compare learning interventions.

In the present example, there are three possible paths that a student may follow as she progresses through her assignment (the specific trace of these paths will become important in the automated reporting and analysis of student performance presented in Section 3). For each student, regardless of path, ASSISTments logs substantial data detailing performance as the student progresses through the assignment. This data includes binary measures of problem accuracy (i.e., a correct or incorrect first response), the students first action (i.e., an attempt vs. requesting tutoring), the number of attempts per problem, the number of feedback interactions per problem (i.e., hints requested or scaffolds seen), whether or not the student saw the bottom out hint (i.e., the correct answer, provided to keep the student from getting stuck within the assignment), and start and end times for each problem. For researchers with a fine-toothed comb, ASSISTments can also provide logged information at the action level, detailing each step taken within a problem. ASSISTments is also able to track user information that is ultimately helpful to researchers, including data on the students performance in the system prior to their inclusion in a study, student characteristics (i.e., gender, age), and additional variables at the class and school levels. Through use of the TestBed, this information is consolidated, anonymized, and provided to researchers through unified reports (depicted in Section 3.1.1) to enhance the ease with which RCEs are conducted at scale.

1.2 Utility of Automated Data-Preprocessing

With students accessing experiments naturally in authentic learning environments, sample populations increase as a function of time. For instance, within three months of deploying a study within ASSISTments, a researcher may accrue 740 participants. This process does not require direct interaction between researcher and teachers, although some researchers choose to work directly with local classrooms to establish stronger controls. As external researchers are unfamiliar with the ASSISTments database and the inner workings of the platform, universal data reporting and preprocessing techniques were designed to ease the hurdle of interpreting system output. Without preprocessing, a researcher analyzing data from the study depicted in Figure 1 would need to use raw data to decipher whether students should be included in analyses, what condition each student experienced, details pertaining to each students experience within that condition (i.e., how many problems were completed, their content, and all associated performance data), and how each student performed at posttest. While such rich information is helpful in analyzing a study, providing researchers with a surplus of data necessitates larger and more complex datasets that must still meet ease of use requirements. Although different researchers focus on different information (as it applies to their particular hypotheses), an infrastructure for data preprocessing, restructuring, and reporting was necessary to bring ASSISTments to the next level as a shared scientific instrument for educational research.

In the following sections we discuss the creation of an automated reporting and analysis system built to provide researchers with data logged from RCEs conducted within the ASSISTments TestBed. The Assessment of Learning Infrastructure (ALI) builds upon existing technology to foster a symbiotic relationship beneficial to students, researchers, the platform and its content, and the science of learning. Evolving from a universal data logging and retrieval tool, ALI is quickly becoming a sophisticated system for automated analysis, offering researchers an overview of their sample population and conducting a selection of analyses for consideration when assessing data. The benefits of ALI can also be felt at scale, with analyses spanning content to drive platform improvements with the long-term goal of personalizing learning.

2. ALI IN THEORY

The Assessment of Learning Infrastructure is an automated research assistant that, while not meant to replace the researcher, is meant to lighten the load of working with large data files output from RCEs conducted within the ASSISTments TestBed. ALI alerts the researcher to new data, presents that data in a meaningful way, tentatively examines effects observed between conditions, and flags potential threats to validity. On a weekly basis, as well as on demand, ALI consults all logged information pertaining to a study and conducts preliminary analyses on student participation and performance (described further in Section 3). The potential benefits of automated reporting and analysis are broad; in the next four sections we briefly discuss how ALI's success will affect ASSISTments and its users, researchers and the Testbed, and the greater learning analytics community.

2.1 Benefits to ASSISTments Users

ALI's work at scale will help to guide the development of stronger learning interventions and, eventually, drive personalized learning within ASSISTments. Research conducted within the TestBed is unique in that while researchers are able to alter content and deliver versatile interventions as previously exemplified in Figure 1, such manipulations are not invasive. Study participation and student performance within an assignment is passively logged. A student may notice that some of her assignments include video feedback or have extra survey questions while others do not, but she is not informed that she is participating in an RCE. A primary goal driving the TestBed's ability to implement RCEs within ASSISTments is the provision of normal instructional practice and interventions that do not compromise learning.

ALI is also beneficial to teachers, as the infrastructure is able to separate rich study information from daily assessment data. Teachers are responsible for assigning content within ASSISTments to their students. Although it seems as though research designs created in the TestBed would complicate daily assessment, class and student reports have been designed such that teachers are provided pertinent information in a clean and concise manner. This low profile approach to conducting research maintains a highly participatory subject pool. Teachers wishing to conduct action research within their classes may do so by working with the TestBed as well, although most prefer to use day-to-day reports to guide their teaching practices rather than large automated data files.

2.2 Benefits to the Researcher

For those conducting RCEs within the ASSISTments TestBed, ALI plays the role of research assistant. The infrastructure intelligently communicates with researchers when new data is available for analysis and provides an overview of the sample distribution across conditions to signify the power of current analyses. Although researchers will undoubtedly run their own in depth analyses, standard high-level analyses can be automated to save time and reduce monotony. For example, ALI's ability to trace a student's path through an assignment allows the infrastructure to infer what condition the student experienced. This allows ALI to test for differential attrition rates across conditions and notify the researcher of apparent selection biases. This simple analysis can serve as a beneficial warning against analyzing posttest results due to potential threats to internal validity. Combined with the data preprocessing and sophisticated reporting that ALI's analytics are built upon, these notifications are often enough to save researchers from hours of wasted labor.

2.3 Benefits to the Platform

When considered at scale, ALI's capabilities for data reporting and analysis contribute to the enhancement of the ASSISTments platform by supporting practical improvements to content and feedback without interrupting student learning. As researchers collaborate and compete to design interventions within the ASSISTments TestBed, it will grow increasingly possible to evaluate interventions at scale, both across skills and longitudinally within students. Ideally, the best version of content and delivery observed (to date) for a particular skill would be delivered to students as the control condition in new RCEs. Through this approach, each study offers the potential for iterative improvement as experiments are launched and re-launched, capturing key features of design-based educational research methodology [3]. Such improvements additionally benefit users through the predicted outcome of enhanced learning gains and researchers through the rapid succession and enhanced validity of positive findings.

ALI's ability to analyze at scale will also help the ASSISTments team to quickly isolate and remove ineffective interventions. It is our goal that in the near future, ALI will conduct robust analyses across multiple studies while considering student, class, and school level characteristics. Roughly speaking, ALI will allow ASSISTments to personalize learning by better understanding why certain educational practices and interventions work for certain students but not for others.

2.4 Benefits for Learning Analytics

How can ALI and the promotion of infrastructures like ALI within other learning platforms benefit the learning analytics community? At its very core, ALI answers the general call of learning analytics, in that the infrastructure "emphasizes measurement and data collection as activities that institutions need to undertake and understand, and focuses on the analysis and reporting of the data" [20]. A strong focus on providing universal measures of learning garnered from authentic learning environments will strengthen the validity of findings from a broad range of interventions that seek to isolate best practices in education.

Further, much attention in the broader scientific and psychological research communities has recently befallen the general inability to replicate research findings [7, 11]. The same is likely true for educational research, with little emphasis placed on data accountability. Perhaps the best outlet for promoting open data, the Pittsburgh Science of Learning Center's Data Shop [8] takes a number of steps in the right direction with regard to shared datasets that promote open, replicable, and sound science. ALI builds upon the PSLC's model of open data reporting by establishing stable, timestamped links to every data analysis report ever provided to a researcher throughout the duration of their work within ASSISTments. Researchers are asked to cite the report from which they draw data for final analyses and publication (explained further in Section 3.1.5). References to these reports will also drastically increase the availability of preprocessed and anonymized educational datasets for researchers wishing to mine big data without designing specific interventions.

In some ways, ALI is also an extension of industry track research focused on learning analytics; companies like Google and Microsoft increasingly implement large-scale experimentation in online learning environments to consider reporting metrics and analytic methods that meet practical goals rooted in scientifically sound evidence [9]. If infrastructures like ALI were incorporated into other learning platforms, similar large-scale experimentation could easily be promoted for its importance to learning analytics.

3. ALI IN PRACTICE

The Assessment of Learning Infrastructure has grown considerably over the past year. ALI began as a robust SQL query to the ASSISTments database to retrieve unified information across multiple studies and to present it to researchers in a single format. Ease of use requirements, communication considerations, and feedback from external researchers has helped ALI to grow beyond data preprocessing and reporting into a tool for learning analytics at scale. The following sections discuss how ALI has evolved and provides examples of the infrastructure's current capabilities in reporting, analyzing, and communicating data from RCEs conducted within the ASSISTments TestBed.

3.1 ALI's Current Capabilities

3.1.1 Data Reporting at Scale

When a researcher submits a study to the ASSISTments TestBed, details about the study and the researcher's contact information are entered into ALI's study repository. Although researchers can request immediate data analysis reports on demand, ALI defaults to a weekly inspection of each study in the database and makes a decision regarding whether or not to process a data analysis report for the researcher. This decision is based on measured increases in sample size. Due to common curricula structures, certain skills are only used at specific times of year and thus, an assignment with an embedded study may be highly popular during the Fall term but not the Spring term. When ALI inspects the study's logged data, at least three new participants since the last ALI communication are required to trigger a new data report.

As teachers using ASSISTments are able to make copies of assignments and alter their content, ALI is also able to detect when teachers have assigned a copy of a study. ALI is sophisticated enough to recognize when a copy is identical to the original study and include data associated with the copy in each report. If a copy of the study has been altered (i.e., problems were removed or sections were changed), ALI does not report data associated with the copy. This ensures that researchers receive all data associated with their experiment without corrupt data.

Once ALI has determined that new data is available, several robust SQL queries are run on the ASSISTments database. Three major queries are used to a) retrieve student data detailing student, class, and school level characteristics for each student recorded prior to random assignment (see Table 1; field definitions are beyond the scope of this paper but are available in our glossary at [13] for additional reference), b) retrieve problem level data (see Table 3), and c) detect the problem set structure (i.e., the paths depicted in Figure 1) for each student with logged data. These three queries provide ALI with the information necessary to establish reports and conduct automated analysis. By working closely with researchers throughout the development of ALI, we have designed four different universal data representations in an attempt to meet dynamic research needs. Subsets of data exemplifying each type of report are provided below. Table 2 shows fields typical to the Action Level file. This file offers the finest granularity of data logged by ASSISTments as a student works through an assignment. Each row provides information pertaining to a single step within a problem (i.e., when the problem is initiated, or when the student asks for a hint). A subset of the Problem Level file is depicted in Table 3. This file provides the same data as that found in the Action Level file, but the granularity has increased. Each row provides information pertaining to a single problem, with actions collapsed across columns. Student Level files, as depicted in Table 4, offer the coarsest granularity of data reporting. In this type of file, each row provides information pertaining to the entire assignment for a single student. For each feature or action, problem information is presented across columns in the order in which the student experienced the assignment, with the number of columns for each feature extrapolated to the maximum number of problems experienced by any student in the file. An alternative version of Student Level data is also provided in which each student assignment is represented by a series of rows, each representing a feature for problems displayed across columns (akin to a pivot table of the file described in Table 4). Full examples of each data file are available at [13] for further consideration. Links to each data file are gathered and presented to the researcher in a single, organized communication, depicted in Figure 2 and discussed further in Section 3.1.5.

When preprocessing is complete and all data files have been compiled, ALI sends analytic commands to Rserve, an extension to the R programming language that allows for other applications to call R functions via a TCP/IP connection [19]. The ASSISTments team created a client side API to interact with Rserve, allowing ALI to send requests to R. Because Rserve is not multithreaded, several instances of Rserve run on separate ports on the ALI server. The server is designed to recycle existing connections, with a connection pool equal to the maximum number of threads used by ALI. This allows several data

Table 1: A theorized subset of student historical data. Each row contains student, teacher, and school characteristics linked to a particular student, using information sourced prior to random assignment

Student	Class ID	Grade	School ID	Guessed Gender	Birth Year	Prior HW Completion %	Prior Class HW Completion %	Normalized HW Mastery Speed
A	1007475	8	5597	Male	2001	0.83	0.88	0.33
B	1180278	8	5597	Male	2001	0.76	0.88	0.03
C	1180278	8	5597	Male	2001	0.76	0.88	0.03
D	1322778	7	2342	Female	2002	0.95	0.97	-0.39

Table 2: A theorized subset of an action level data file. Each row represents a single action within a single problem as experienced by a student. This is the finest granularity of data reported by ALI

Student	Problem ID	Sub-Problem ID	Order	Action Type	Timestamp	Answer	Correctness
A	PRAUVJS	806533	1	Start	08/26/15 15:25:26	--	--
A	PRAUVJS	806533	2	Hint	08/26/15 15:25:52	--	--
A	PRAUVJS	806533	3	Answer	08/26/15 15:26:40	18.2	TRUE
A	PRAUVJS	806533	4	End	08/26/15 15:26:42	--	--
A	PRAVKJX	833840	1	Start	08/26/15 15:26:43	--	--

Table 3: A theorized subset of a problem level data file. Each row contains all the information linked to a single problem as experienced by a student. This is a popular form of data for student modeling and analytics

Student	Assignment ID	Problem ID	Correct	Answer	Hints	Attempts	Start Time	End Time
A	1007475	PRAUVJS	1	18.2	0	1	08/26/15 15:25:26	08/26/15 15:26:42
A	1007475	PRAVKJX	1	14.3	0	1	08/26/15 15:26:43	08/26/15 15:27:45
A	1007475	PRAVKHT	1	6.4	0	1	08/26/15 15:27:50	08/26/15 15:28:47
B	1180278	PRAUVJX	0	22.8	2	3	08/26/15 17:14:22	08/26/15 17:15:42
B	1180278	PAVKGZ	0	7.2	0	2	08/26/15 17:15:43	08/26/15 17:17:31

Table 4: A theorized subset of a student level data file. Each row contains all information linked to a single student's experience of the problem set. Assignment information is presented across columns in the order in which the student experienced problems

Student	Assignment ID	Late	Mastered	Correct Q1	Correct Q2	Correct Q3	Answer Q1	Answer Q2	Answer Q3
A	1007475	1	1	1	1	1	18.2	14.3	6.4
B	1180278	0	0	0	0	1	17	14.1	6.4
C	1180278	1	0	0	1	--	24.6	14.3	--
D	1322778	0	1	1	1	1	18.2	14.3	6.4

analysis reports to occur simultaneously, all using different Rserve connections. This approach lowers the turnaround time when a researcher actively requests data. It also keeps weekly reporting as efficient as possible, as all datasets in ALI's study repository are assessed weekly for potential reporting.

3.1.2 Smart Structures

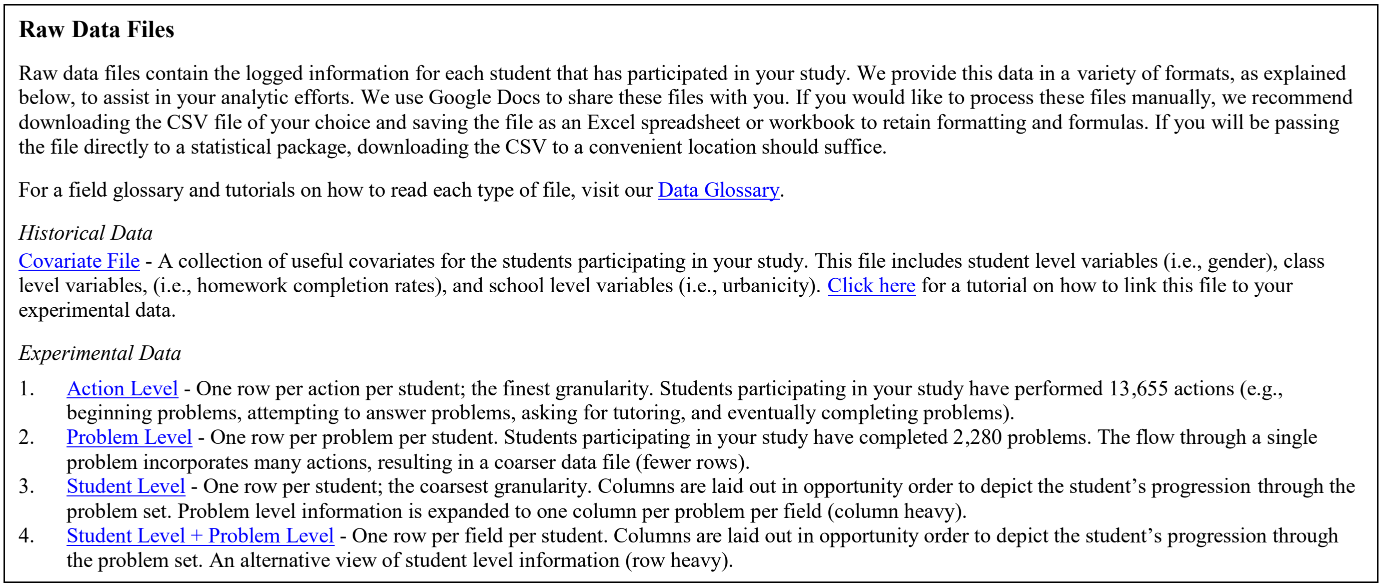
In order to determine *what* to analyze, ALI must first process the structure of a study and trace each student's path through the assignment (as previously discussed in relation to Figure 1). As ALI parses the assignment's structure, the infrastructure is able to make intelligent decisions upon meeting certain section types within the design. This is accomplished by recursively generating the assignment's reported structure into tree form. Within the Problem Level data file presented in Table 3, each problem is labeled with a path, similar to that used when traversing a set of folders within an operating system. ALI steps through each problem path for each student to establish an intuitive structure of the study and to cluster students by condition.

RCEs within the ASSISTments TestBed are designed by taking advantage of a variety of section types offered by the platform. The "If-Then" routing discussed in Section 1.2 was an example of a section type. When ALI observes an If-Then structure that issues a routing standard like a "Video Check," the infrastructure intelligently conducts its analyses on students assigned to the study and disregards students routed to alternative content. Similarly, studies often employ parallel experimental and control conditions delivered using a section type referred to as a "Choose Condition." This section type is used to drive random assignment. The "Choose Condition" depicted in Figure 1 included two parallel conditions: an assignment with video content and a control assignment with traditional text content. Currently, in order for ALI to recognize an assignment as a research study, a "Choose Condition" must be present when mapping the assignment's structure. ALI then assesses logged data within each condition and considers any section immediately following these conditions as a subsequent posttest (see Figure 1). Using this information, ALI is able to aggregate statistics and perform a selection of simple analyses across problems and students.

It is important to note that research designs within the ASSISTments TestBed can grow far more complex than the simple structure presented herein. When assignments include nested section types and multiple "If-Then" routing standards, ALI currently has difficulty interpreting condition and isolating posttest content. In its current form, ALI is only meant to assist researchers with the analysis of common design patterns. Future work, discussed in Section 5, will expand ALI's ability to intelligently parse studies using tagging rules set forth by the researcher.

3.1.3 Selection Bias

After establishing a study's structure and sample distribution, ALI is able to assess assignment completion rates across conditions and alert researchers to potential threats to internal validity due to selection bias. ALI records the observed number of students in each condition that began the assignment, and considers logged assignment end times to consider the proportion of students that ultimately completed the assignment. The observed distribution is then compared to the expected distribution of proportional attrition in a normal sample. A Chi-squared analysis is used to determine if the observed distribution of attrition significantly differs from the expected distribution. ALI then flags conditions that have a reliably different attrition rate and alerts the researcher of a potential threat to internal validity. Without considering differential attrition across conditions, an analysis of posttest performance may inaccurately suggest the significant effect of a particular condition that was actually driven by the disproportionate loss of weaker students. This simple analysis, presented to researchers as shown in Figure 3, may help even the most seasoned experts to accurately assess their sample. It is important to note that while ALI provides this warning, the infrastructure still releases all data to the researcher and never prohibits the researcher from further analysis. The goal of ALI's selection bias assessment is not to impede or prevent analysis, but rather to advocate sound analytic practices.



Raw Data Files

Raw data files contain the logged information for each student that has participated in your study. We provide this data in a variety of formats, as explained below, to assist in your analytic efforts. We use Google Docs to share these files with you. If you would like to process these files manually, we recommend downloading the CSV file of your choice and saving the file as an Excel spreadsheet or workbook to retain formatting and formulas. If you will be passing the file directly to a statistical package, downloading the CSV to a convenient location should suffice.

For a field glossary and tutorials on how to read each type of file, visit our Data Glossary.

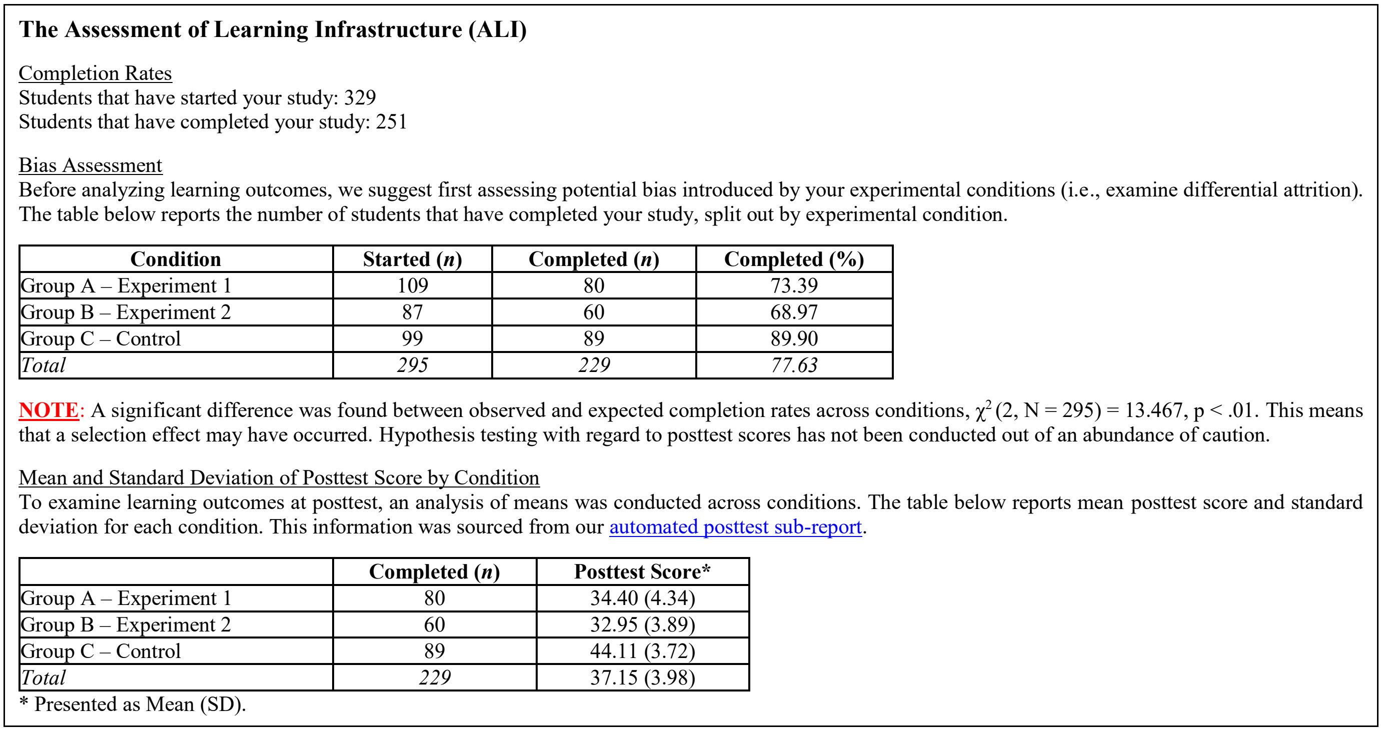
Historical Data

Covariate File - A collection of useful covariates for the students participating in your study. This file includes student level variables (i.e., gender), class level variables, (i.e., homework completion rates), and school level variables (i.e., urbanicity). Click here for a tutorial on how to link this file to your experimental data.

Experimental Data

1. Action Level - One row per action per student; the finest granularity. Students participating in your study have performed 13,655 actions (e.g., beginning problems, attempting to answer problems, asking for tutoring, and eventually completing problems).
2. Problem Level - One row per problem per student. Students participating in your study have completed 2,280 problems. The flow through a single problem incorporates many actions, resulting in a coarser data file (fewer rows).
3. Student Level - One row per student; the coarsest granularity. Columns are laid out in opportunity order to depict the student's progression through the problem set. Problem level information is expanded to one column per problem per field (column heavy).
4. Student Level + Problem Level - One row per field per student. Columns are laid out in opportunity order to depict the student's progression through the problem set. An alternative view of student level information (row heavy).

Figure 2: A thoroughly developed universal reporting of logged data from students participating in RCEs. Each file presented here is discussed further, including depictions of file subsets, in Section 3.1.1.



The Assessment of Learning Infrastructure (ALI)

Completion Rates
Students that have started your study: 329
Students that have completed your study: 251

Bias Assessment
Before analyzing learning outcomes, we suggest first assessing potential bias introduced by your experimental conditions (i.e., examine differential attrition). The table below reports the number of students that have completed your study, split out by experimental condition.

Condition	Started (*n*)	Completed (*n*)	Completed (%)
Group A – Experiment 1	109	80	73.39
Group B – Experiment 2	87	60	68.97
Group C – Control	99	89	89.90
Total	*295*	*229*	*77.63*

NOTE: A significant difference was found between observed and expected completion rates across conditions, χ^2 (2, N = 295) = 13.467, p < .01. This means that a selection effect may have occurred. Hypothesis testing with regard to posttest scores has not been conducted out of an abundance of caution.

Mean and Standard Deviation of Posttest Score by Condition
To examine learning outcomes at posttest, an analysis of means was conducted across conditions. The table below reports mean posttest score and standard deviation for each condition. This information was sourced from our automated posttest sub-report.

	Completed (*n*)	Posttest Score*
Group A – Experiment 1	80	34.40 (4.34)
Group B – Experiment 2	60	32.95 (3.89)
Group C – Control	89	44.11 (3.72)
Total	*229*	37.15 (3.98)

* Presented as Mean (SD).

Figure 3: Current ALI analytic reporting. Available analyses include a Chi-squared test comparing the observed and expected sample distributions, simple hypothesis testing, and an analysis of means on posttest performance between conditions. Note that these analyses are currently driven by the structure of the assignment as parsed by ALI from Problem Level data. Future work includes allowing researchers to tag their study with items of interest to automate analysis with greater sophistication.

3.1.4 Simple Hypothesis Testing

After conducting a selection bias assessment, ALI progresses to a set of simple hypothesis tests with regard to posttest performance. If ALI detects a posttest section when parsing an assignment's structure, the infrastructure compares performance across conditions by referring to the previously aggregated group distributions. ALI approaches posttest analysis much like a researcher would: if only two conditions are detected within the study, ALI conducts a t-test, while if more than two conditions are detected, ALI conducts an Analysis of Variance (ANOVA). ALI currently has the API to support simple univariate and multivariate analyses including ANOVA, ANCOVA, MANOVA, and MANCOVA. ALI stores all input parameters for a given statistical test in a single object. The parameters are extracted from this object and transformed into the appropriate R function calls through the Rserve API communication. Results are accumulated and presented to the researcher alongside an analysis of means, as shown in Figure 3, allowing the researcher to observe the direction of the reported effect. Note that in the present example, ANOVA results are not presented to the researcher out of an abundance of caution due to ALI's detection of a potential selection bias. Our goal in restricting this information is strictly in the promotion of sound scientific inquiry. It should also be noted that covariates are not presently considered in ALI's hypothesis testing. Future work will control for student, class, and school level characteristics sourced from the historical student data file (see Table 1) by using ANCOVA or MANCOVA approaches in an attempt to explain additional variance in learning outcomes.

3.1.5 Data Storage and Researcher Output

When ALI's automated analysis is complete, ALI stores all data files and analytic output on Google Drive in archival quality. This data cannot be altered but can be downloaded by anyone. For active studies, copyright protection will be placed on new data analysis reports for one year from the study's initial run date. This means that researchers will have a full calendar year to publish on their findings before their data becomes freely available to the public.

ALI communicates to researchers via email, providing a link to a stable URL for a Google Doc housing that week's data analysis report. The Doc contains links to all raw data files, as shown in Figure 2, and provides automated analysis as depicted in Figure 3. The creation of this Google Doc is automated, based on an HTML template file that uses custom tagging conventions to insert variables with dynamic text or data. Using this method, the same report can be generated multiple times or across multiple assignments with changes to only the pertinent information. This allows for customized reporting based on the results of ALI's analysis. The Google Doc report also provides researchers with links to additional resources including a glossary explaining features of the data and video tutorials on how to understand each file type (available at [13]).

When researchers are ready to publish findings, a condition of working with the ASSISTments TestBed requires that they include a reference in their work to the stable record from which they sourced the data files used for final analyses. This approach allows reviewers and secondary researchers to gain access to raw study data, thereby encouraging replication and open science [11]. In addition to the raw data, secondary researchers will also be able to use these references to access ALI's analytic report, including all automated analyses.

4. ANALYSIS AT SCALE

Although ALI's analytic structure is still somewhat rudimentary, considered at scale, comparisons of findings from multiple studies can offer substantial insights for the ASSISTments platform and in more general terms, for the learning analytics community. By simultaneously examining attrition outcomes across studies it becomes possible to make claims about the quality of interventions that crosscut multiple skills. As ALI's analytical capabilities increase, analysis at scale will grow even more powerful.

As a proof of concept of the potential benefits of automated analysis at scale, ALI was run across a special dataset including 25 studies that are currently running within ASSISTments. This file was created for another sophisticated approach to modeling student performance across multiple studies [18], but serves as a perfect example of ALI's capabilities at scale. In the spirit of open data, this file is available for reference at [12]. The studies in this file were selected from a group of 126 studies currently running within the ASSISTments platform based on the following criteria:

- Studies selected contained at least 50 students within each condition that completed the assignment.
- Studies selected were designed within Skill Builders, a mastery learning based assignment that considers predefined thresholds for student completion (i.e. by default, to complete the assignment the student must solve three consecutive problems accurately).

As most of the studies in this file were built prior to the implementation of automated path-logging (which drives ALI's ability to read in the structure of the study and infer a condition for each student), condition was manually traced and logged for each student based on his or her observed problem sequence. A number of these studies were also built before the availability of If-Then routing and subsequent checks for internal validity (i.e., the "Video Check" explained in connection to Figure 1). As such, it is difficult to tell if students experienced technical difficulties during the course of a condition. To analyze this dataset using all of the capabilities that ALI has with recently designed studies, we manually notated flags regarding the observed fidelity of conditions. This flagging also included whether students 'tested out' of the condition experience (i.e., if a student was assigned to a condition in which the treatment was presented through feedback but answered the first 3 consecutive problems accurately, they did not ultimately experience the treatment). As only three of the studies in this file contained valid posttest information, we only present ALI's selection bias assessment for consideration at scale (see Table 5).

The 25 studies presented in Table 5 span a variety of investigations including: assessing the effect of various types of video tutoring (i.e., pencasts, teacher recorded instruction, online resources) compared to traditional text-based tutoring across multiple designs (i.e., using scaffolding, using hints, as an intervention to wheel-spinning [2], or provided based on student choice), investigating the manipulation of content (i.e., interspersing learning with humor through comics in content or feedback, asking students to gauge their confidence in solving problem content, and altering student mindset (as inspired by [4]), and challenging cognitive principles (i.e., mental representations, and alterations in the consistency of math equations). Assignment names, as presented in Table 5, are tagged with the grade level and domain of the skill content as defined by Common Core State Standards [10]. Despite differences in domain and experimentation, ALI is able to provide a sense of condition quality across studies at scale.

The results of the simple Chi-squared analyses in Table 5 may not seem significant at first, but are actually quite insightful at scale. In studies with two conditions, experiment vs. control (20 comparable sets of the 25 shown in Table 5), the control groups showed less attrition in 15, while the experimental groups showed less attrition in only five. On its own, this comparison suggests that experimental conditions correlate with higher attrition rates. However, this attrition is only significantly different than that of a normally distributed sample in five studies ($p < .05$), with experimental conditions showing significantly more attrition than expected in four studies, and control conditions showing significantly more attrition than expected in only a single study.

At scale, these analyses can help researchers and developers determine which interventions are effectively retaining students,

Table 5: ALI's Bias Assessment at Scale - Observed Distributions and Chi-Squared Analyses Across 25 Problem Sets

Problem Set by Condition	*Started (n)*	*Completed (n)*	*Completed (%)*	*df*	χ^2	*p*
Multiplying Mixed Numbers 5.NF.B.4a	**775**	**466**	**60.13**	**1**	**5.30**	**0.021***
Control	403	258	64.02			
Experiment	372	208	55.91			
Understanding Vocabulary About Circles G-C.A.2	**695**	**674**	**96.98**	**1**	**4.87**	**0.027***
Control	330	325	98.48			
Experiment	365	349	95.62			
Equivalent Expression 6.EE.B.4	**273**	**240**	**87.91**	**1**	**0.39**	**0.532**
Control	138	123	89.13			
Experiment	135	117	86.67			
Writing Inequalities from Situations 6.EE.B8	**627**	**539**	**85.96**	**1**	**2.21**	**0.138**
Control	338	297	87.87			
Experiment	289	242	83.74			
Dividing Mixed Numbers 6.NS.A.1	**1864**	**1285**	**68.94**	**1**	**0.99**	**0.321**
Control	943	660	69.99			
Experiment	921	625	67.86			
Finding Expected Value SS.MD.B.5	**457**	**337**	**73.74**	**1**	**0.06**	**0.802**
Control	224	164	73.21			
Experiment	233	173	74.25			
Conditional Probability SS-CP.A.3	**515**	**366**	**71.07**	**1**	**0.70**	**0.401**
Control	281	204	72.60			
Experiment	234	162	69.23			
Permutations and Combinations SS-CP.B.2	**540**	**456**	**84.44**	**1**	**0.00**	**0.958**
Control	265	224	84.53			
Experiment	275	232	84.36			
Basic Logarithm Manipulation F-BF.B.5	**136**	**121**	**88.97**	**1**	**0.21**	**0.645**
Control	62	56	90.32			
Experiment	74	65	87.84			
Properties of Exponents 8.EE.A.1	**545**	**435**	**79.82**	**1**	**0.24**	**0.626**
Control	264	213	80.68			
Experiment	281	222	79.00			
Intermediate Logarithm Manipulation F-BF.B.5	**205**	**169**	**82.44**	**1**	**8.44**	**0.004****
Control	102	92	90.20			
Experiment	103	77	74.76			
Solving $ab^{ct} = d$ LE.A.4a	**147**	**122**	**82.99**	**1**	**0.01**	**0.914**
Control	72	60	83.33			
Experiment	75	62	82.67			
Finding Inverse Functions F-BF.B.4	**301**	**143**	**47.51**	**1**	**3.32**	**0.068†**
Control	145	61	42.07			
Experiment	156	82	52.56			
Composition of Functions F-BF.A.1c	**219**	**173**	**79.00**	**1**	**0.86**	**0.354**
Control	118	96	81.36			
Experiment	101	77	76.24			
Sequences F-BF.A.2	**382**	**241**	**63.09**	**1**	**0.20**	**0.658**
Control	198	127	64.14			
Experiment	184	114	61.96			
Comparing Values - Multiplying by Fractions 5.NF.B.5a	**129**	**121**	**93.80**	**1**	**1.59**	**0.208**
Control	69	63	91.30			
Experiment	60	58	96.67			
Converting Radians to Degrees F-TF.A.1	**245**	**226**	**92.24**	**1**	**0.23**	**0.631**
Control	129	120	93.02			
Experiment	116	106	91.38			
Trigonometric Ratios G-SRT.C.8	**307**	**266**	**86.64**	**1**	**0.91**	**0.341**
Control	141	125	88.65			
Experiment	166	141	84.94			
Pythagorean Theorem – Finding the Hypotenuse 8.G.B.7	**447**	**349**	**78.08**	**1**	**6.40**	**0.011***
Control	237	174	73.42			
Experiment	210	175	83.33			
Solving 1-Step Equations 7.EE.B.4a	**928**	**818**	**88.15**	**1**	**0.01**	**0.934**
Control	459	405	88.24			
Experiment	469	413	88.06			
Prime Factorization 6.NS.B.4	**1238**	**1058**	**85.46**	**2**	**0.97**	**0.616**
Control	430	369	85.81			
Experiment 1	399	345	86.47			
Experiment 2	409	344	84.11			
Order of Operations (No Exponents) 7.NS.A.3	**1231**	**1172**	**95.21**	**2**	**4.50**	**0.105**
Group A - Consistent/Neutral	597	574	96.15			
Group B - Inconsistent	300	287	95.67			
Group C - Mixed	334	311	93.11			

Note. †p < .10, *p < .05, **p < .01. *df* = Degrees of Freedom.

Table 5: ALI's Bias Assessment at Scale - *Continued*

Problem Set by Condition	*Started (n)*	*Completed (n)*	*Completed (%)*	*df*	χ^2	*p*
Multiplying Simple Fractions 5.NF.B.4a	**598**	**559**	**93.48**	**3**	**1.54**	**0.673**
Group A – No Choice + Text	142	131	92.25			
Group B – Choice + Text	222	211	95.05			
Group C – Choice + Video	76	71	93.42			
Group D – No Choice + Video	158	146	92.41			
Rotations 8.G.A.3	**306**	**186**	**60.78**	**1**	**0.82**	**0.365**
Experiment 1	145	92	63.45			
Experiment 2	161	94	58.39			
Reflections 8.G.A.3	**239**	**171**	**71.55**	**1**	**0.17**	**0.680**
Experiment 1	125	88	70.40			
Experiment 2	114	83	72.81			

Note. †p < .10, *p < .05, **p < .01. *df* = Degrees of Freedom.

and more importantly, critical design issues that drive students away. As many of these 25 studies were designed prior to the implementation of internal validity checks (i.e., assessing a student's technical abilities with video content), we believe that the analyses in Table 5 suggest higher attrition in experimental conditions because certain students were assigned to content that they had difficulty accessing. This finding would not likely hold true when considering studies run more recently, suggesting the importance of the recent implementation of If-Then routing. Future work with ALI at scale will help to confirm this hypothesis. Usability is a concern within any online learning system, and providing students with access to default assignments when they cannot access enriched content is a safe practice.

It is also important to consider the percentage of students excluded from analysis prior to the assessments presented in Table 5. Within all sets, an average of 22.85% of students did not actually experience condition and were removed from the sample prior to analysis. Students that fail to experience interventions implemented within feedback (due to mastery or performance at ceiling) provide valuable information to researchers regarding the raw (inflated) sample size required to achieve statistical power. Certain elements of a study's design, including the content domain (i.e., some topics are easier than others and students require less feedback on average), and the type of feedback provided (i.e., on demand feedback requires a larger raw population than feedback provided automatically upon the student's incorrect response), can have a significant impact on the raw sample size required to attain enough treated students to reliably detect effects. RCEs that consider interventions implemented strictly within problem content have fewer issues with regard to raw sample sizes as all students experience the intervention regardless of performance, easing potential issues surrounding intent-to-treat analyses.

Finally, analyzing the selection effects inherent to multiple assignments simultaneously allows ASSISTments to evolve more rapidly, providing benefits to users, researchers, and the learning analytics community. As the experimental conditions in Table 5 exhibited only 1.5% greater attrition on average than control conditions, it is possible that the benefits of these experimental interventions may still outweigh the increase in attrition. Additional data mining would be necessary to determine a standard at which the potential for emphasized learning gains within an experimental condition no longer outweighed the potential for increased attrition. However, regularly conducting this type of broad scale analysis across assignments could quickly isolate studies with conditions considered extremely detrimental, and the condition could be discontinued in order to limit the intervention's negative impact on students. ALI's automated analysis makes the process of intervention validation dramatically more efficient and robust. From these findings, and from future, more powerful iterations of ALI's at-scale capabilities, ASSISTments will be able to deliver rapid iterations of interventions with the goal of optimizing students' interactions with the system through enhanced usability and strengthened content and delivery methods.

5. LIMITATIONS & FUTURE WORK

As ALI is constantly evolving and gaining new capabilities, the version of the infrastructure presented here carries a number of limitations. As made apparent by the complex methods applied to consider ALI's effects at scale, the infrastructure is currently only able to recognize studies with logged path information. The implementation of path logging occurred in March 2015, and ALI is only able to reliably analyze studies that were created after this implementation. This limitation is compounded by ALI's inferences of the study design and posttest items. As studies within the ASSISTments TestBed can be designed using a number of complex, nested structures, ALI's current decisions about study designs are not exceptionally intelligent. A serious limitation of the work presented herein is that the infrastructure is currently only able to reliably recognize and analyze study designs with simple structures (i.e., "If-Then" routing, a single "Choose Condition," and a clear cut posttest section that directly follows an intervention).

While these limitations influence ALI's significance for the learning analytics community, they can easily be resolved through future work. One of our current focuses is the implementation of a tagging system that will allow researchers to identify pertinent sections of a study prior to its distribution. Using unified naming structures for the design of assignment sections within the building process (e.g., [experiment], [control], [posttest]), researchers will essentially be able to tell ALI exactly how to approach analysis. This will allow ALI to provide customized analysis and, potentially, refined data files that are preprocessed according to the researcher's distinct needs. Tagging will also allow for analyses that collapse similar treatment groups (i.e., experimental group 1 and experimental group 2 could both be tagged with [experiment] to denote that ALI should collapse these conditions), that isolate unconventional posttest problems (i.e., problems falling within a section that does not immediately follow a "Choose Condition"), and that assess growth models of student performance (i.e., by measuring pre- to posttest gains, or through more complex hierarchical models).

Future work for the ALI team also includes defining a powerful list of student, class, and school level variables for use as covariates in statistical analyses. Variables that have already been established include measures of each student's prior performance within ASSISTments, measures of their completion rate on classwork and homework assignments, and normalized values that compare the student's performance and attrition against that of their class. As such, future iterations of ALI's at-scale capabilities will also be able to control for particular student characteristics in order to assess the true variance established by experimental interventions. Additional content is also being built into ASSISTments and made available in the TestBed to collect self-report measures from students for use as possible covariates. Rich covariates will provide ALI with the ability to examine the effects of experimental interventions across groups while controlling for substantial variance, making automated analyses far more robust.

6. CONTRIBUTION

The learning analytics community will benefit greatly from the Assessment of Learning Infrastructure (ALI) and the promotion of similar infrastructures for other online learning platforms. Currently, very few learning technologies serve as scientific tools for researchers to conduct and communicate the findings of sound educational research at scale. By allowing researchers to conduct research within authentic learning environments through classwork and homework completed within online learning platforms, it is possible to collect rich log files that can be reported in universal formats and analyzed using automated processes. As a community, a strong focus on providing universal measures and analyses from these platforms will strengthen the validity of findings from a broad range of interventions that seek to isolate best practices in education. The broad dissemination of vast anonymized educational datasets will also propel the field toward more transparent, replicable, and reputable scientific practice, improving learning analytics for all.

7. ACKNOWLEDGMENTS

We acknowledge funding from multiple NSF grants (1440753, 1252297, 1109483, 1316736, 1535428, 1031398), the U.S. Dept. of Ed. (R305A120125, R305C100024, P200A120238), and ONR. We also thank our colleagues Yang Lu and Anthony Botelho for their roles in designing ALI. Thanks to S.O. & L.P.B.O.

8. REFERENCES

[1] Adjei, S.A. & Heffernan, N.T. (2015). Improving Learning Maps Using an Adaptive Testing System: PLACEments. In Conati, Heffernan, Mitrovic, & Verdejo (eds.) Proc of the 17th Int Conf on AIED. Springer, 517-520.

[2] Beck, J.E. & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. In Lane, Yacef, Mostow & Pavlik (eds.) Proc of the 16th Int Conf on AIED. Springer-Verlag, 431-440.

[3] Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *J of Learning Sciences*, 2(2), 141-178.

[4] Dweck, C.S., Chiu, C., & Hong, Y. (1995). Implicit theories and their role in judgments and reactions: A world from two perspectives. *Psychological Inquiry*, 6(4), 267-285.

[5] Heffernan, N. & Heffernan, C. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *Int J of AIED,* 24(4), 470-497.

[6] Institute of Education Sciences. (2003). Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide. U.S. Dept of Ed. Washington, D.C.

[7] Ioannidis J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Med 2*(8): e124.

[8] Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: The PSLC DataShop. *Handbook of EDM*, 43.

[9] Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, *18*(1), 140-181.

[10] National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). Common Core State Standards. Washington, DC: Authors.

[11] Open Sci Collab. (2015). Estimating the reproducibility of psychological science. *Science*, 349 (6251).

[12] Ostrow, K. (2015). Data for "The Assessment of Learning Infrastructure (ALI): The Theory, Practice, and Scalability of Automated Assessment." Accessed from http://tiny.cc/LAK2016-ALI

[13] Ostrow, K. & Heffernan, C. (2014). How to Create Controlled Experiments in ASSISTments. Retrieved from https://sites.google.com/site/assistmentstestbed/

[14] Ostrow, K.S. & Heffernan, N.T. (2014). Testing the Multimedia Principle in the Real World: A Comparison of Video vs. Text Feedback in Authentic Middle School Math Assignments. In Stamper, et al. (eds.) Proc of the 7th Int Conf on EDM, 296-299.

[15] Ostrow, K., Heffernan, N., Heffernan, C., & Peterson, Z. (2015). Blocking vs. Interleaving: Examining Single-Session Effects within Middle School Math Homework. In Conati, Heffernan, Mitrovic, & Verdejo (eds.) Proc of the 17th Int Conf on AIED. Springer, 388-347.

[16] Pashler, H., Rohrer, D., Cepeda, N. & Carpenter, S.K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review. 14* (2), 187-193.

[17] San Pedro, M., Baker, R., Gowda, S., & Heffernan, N. (2013). Towards an Understanding of Affect and Knowledge from Student Interaction with an Intelligent Tutoring System. In Lane, Yacef, Mostow & Pavlik (eds.) Proc of the 16th Int Conf on AIED. Springer-Verlag, 41-50.

[18] Selent, D., Patikorn, T., Heffernan, N. (Under Review). ASSISTments Dataset from Multiple Randomized Controlled Experiments. Submitted to the 3rd Annual ACM Conference on L@S.

[19] Urbanek, S. (2003). Rserve—a fast way to provide R functionality to Applications. In Hornik, Leisch, & Zeileis, Proc of the 3rd Int Workshop on DSC, ISSN 1609-395X. http://rosuda.org/rserve.

[20] U.S. Department of Education, Office of Educational Technology. (2012). Enhancing teaching and learning through educational data mining and learning analytics: An Issue Brief. Washington, DC