CLUSTERING STUDENT ERRORS

Morten Elkjær[1] & Christian Hansen[2]

[1]*Danish School of Education, Aarhus University, Emdrup, Denmark; me@edu.au.dk*

[2]*University of Copenhagen, Copenhagen, Denmark; christianh@edulab.dk*

This paper gives an in-depth description of the research design in the pursuit of clustering of students' performance when solving different types of linear equations. The student performances are clustered using data from 457,185 answers to equation tasks, made by 37,585 students, distributed across 3,438 unique linear equations in a digital learning environment. The tasks consist of different categories of linear equations. The clustering analysis contributes to the development of an online tool to provide the teachers with easy accessible formative assessment. At this point, the attempt to cluster the students' performance have not yet been successful, meaning that no clusters are found. Instead, a description of how the pursuit of these clusters will continue is presented alongside the research design.

Keywords: Students' difficulties, linear equations, unsupervised learning, clustering, formative assessment.

INTRODUCTION

This paper presents the research design for utilizing a large amount of data to assess and accommodate students' mathematical difficulties when they are working with linear equations in Danish lower secondary school, more specifically the 6th and 7th grade (12-14 years old). Overall, this paper describes a research design that is based around the clustering of students working with linear equations in an online learning environment. An initial attempt was made, that unfortunately did not bear fruit. The project stems from a collaboration between two doctoral candidates associated with two independent industrial PhD projects. The authors collaborate across academic institutions with the common goal of generating knowledge about whether or not standard digital mathematical tasks on an online platform can serve as a non-disruptive diagnostic tool to generate easily accessible formative assessment for the teachers in Danish lower secondary school. The ideas and the design described are research work in progress. Therefore, only initial results and anticipated outcomes will be presented alongside the design.

The Danish private company Edulab develops and maintains an online mathematical learning platform for the Danish K-10 schools. In Denmark, 75% of the K-10 schools subscribe to Edulab's learning platform, which effectively means that 600,000 students have access to these digital learning materials. During the last 12 years, Edulab has developed the online mathematical learning platform, called *matematikfessor.dk*. The platform performs primarily as a teacher-driven supplement to a learning material, but the students also have access in order to explore the mathematical content on their own. Every day Danish school students collectively give answers to 1.5 million tasks on the online learning platform. This creates a unique opportunity to implement didactical research results regarding students' difficulties with equations directly into practice. The two industrial PhD projects aim to provide Edulab with a research-based tool to reveal and capture students' mathematical difficulties when working with linear equations in order to provide teachers with valuable information hereof. The assumption is that standard digital mathematical tasks (already implemented on the

platform) can serve as a substitute for a diagnostic test, based on the idea that a large enough amount of data together with research-based diagnostic verification can generate significant information on students with difficulties when working with linear equations. This leads to the following overall research aim:

How can an online diagnostic tool for lower secondary school be designed, utilizing existing research findings on mathematical difficulties when working with equations in order to provide the teachers with significant dynamic formative assessment?

In order to approach the answer to the above question, we have chosen to explore possibilities based on the vast amount of data that Edulab has collected so far from students answering textbook like standard tasks, involving equations, in their online environment. Therefore, we have posed the following research question that will be the main research question for this paper:

To what extent can the existing categorization of standard textbook linear equations serve as a mean for generating clusters of students with a large amount of data?

By existing categorization, we mean the levels of distinctions we are able to make based on the tasks already implemented on the planform.

THEORETICAL BACKGROUND

Terms like misconceptions, alternative conceptions, or errors have been used in the mathematics education literature in the past describing children's beliefs, conceptions and problem solving strategies (J. L. Booth, McGinn, Barbieri, & Young, 2017; L. Booth, 1984; Linsell, 2009). When we mention children's difficulties working with equations in this particular context, the main focus is to describe common misinterpretations or common conceptual limitations that children have about linear equations, for example, the role of the equal sign, the role of the literal symbols or the ability to apply different problem solving strategies.

From reviewing the literature, it is concluded that the mathematics educations community knows a great deal about children's conceptions and errors working with the elements of algebra (Rhine, Harrington, & Starr, 2018). Linear equations have not received the same attention, but some studies have dealt with the strategic aspect and the concrete errors children make when working with equation solving (Herscovics & Linchevski, 1994; Kieran, 1985, 1992; Linsell, 2009). Linsell (2009) presents tests and interviews with the purpose of uncovering the difficulties children have with solving equations as well as the origin of these difficulties. Together with the strategic difficulties that arise when students have to solve equations, there are of course also difficulties with the individual sub-concepts of linear equations. The interpretation of the equal sign becomes a natural conceptual centrepiece when addressing the overall concept of equations (Kieran, 1981). Alongside the equal sign comes the algebraic elements of literal symbols, numbers and operators (L. Booth, 1984; Küchemann, 1981). When talking about the concept of equations, one also has to take more intangible aspects such as truth-value and solution into consideration.

Following categorization of the types of linear equations is adapted from Vlassis (2002) based on abstraction level and the partition presented in Filloy and Rojano (1989).

- Concrete arithmetical equations: Arithmetical equations that consist only of natural numbers and only include a single occurrence of the unknown. (e.g. $ax + b = c, \quad a, b, c \in N_0$)

- Abstract arithmetical equations: Equations with the unknown in one member, which require certain algebraic manipulations, because of the presence of negative integers or several occurrences of the unknown. ($a_1x \pm \cdots \pm a_n \pm b_1x \pm \cdots \pm b_nx \ \pm c, \quad a_1 \ldots a_n, b_1 \ldots b_n, c \in Z, \ n \in N$)

- Algebraic equations: Equations similar to the abstract arithmetical equations, but with the exception of occurrences of the unknown on both sides. ($ax \pm b = cx \pm d, \quad a, b, c, d \in Z$)

Linsell (2007) lets us divide this categorization, into equation types based on the number of steps it will take to solve the equations by normal transformations. Within the concrete arithmetical equations, we find the one-step equations, divided into containing small and large numbers. The abstract arithmetical equations contain (in this context) two- and three-step equations. It is also possible to have two-step equations within the algebraic equations, but for the most part in this context, they will require further transformational steps. The categorization of the equations used for analysis, taken from the online platform, is presented in table 1.

Label	**Form**	**Steps**	**Number of tasks**
1	$ax + b = c, \quad x, a, c \in \{1,2, \ldots ,9\}, b \in \{0,1, \ldots ,9\}$	One/Two-step	106
2	$ax + b = c, \quad x, a, c \in N \ b \in N_0$	One/Two-step	812
3	$ax + b = c, \quad x, c \in Z \ b \in N_0, a \in N$	One/Two-step	501
4	$ax + b = cx + d, \quad a, b, c, d, x \in N$	Two/Three-step	1478
5	$ax + b = cx + d, \quad a, c \in N \ x, b, d \in Z$	Two/Three-step	519

Table 1 - Categorization of linear equations on the online platform

The reason for having this categorization of the linear equation is that the online platform currently only give us the opportunity to make the above distinction at this point. Notice that equations labelled 1, 2 and 3 are arithmetical, but are only considered abstract if e.g. *c* is a negative number. The equations labelled 4 and 5 are considered algebraic equations.

METHODOLOGICAL CONSIDERATIONS

Design research (Barab & Squire, 2004) will pave the way for the knowledge generated through iterations of the work with the initial clustering of students performance solving different types of equations and the selection of the standard textbook tasks, about equations, implemented on the online learning platform.

In order to answer the main research question for this paper, the design goal is the analysis of student interaction with the chosen items on the platform, fitting the categorization based on the literature. The categorization of the equations mentioned in the theoretical background is to be utilized in order to analyse the clusters when found. In the following section, we describe methodological considerations as well as the process of and theory behind the approach for the clustering of the

different types of students based on the answers given to the equations presented in the theoretical background and the specific errors the student make while working with these equations.

CATEGORIZING AND CLUSTERING OF ERRONEOUS ANSWERS AT EDULAB

This section contains our current approach to the categorizing and clustering of the errors the students make when using the portal. Before the approach can be described in detail, it is necessary to shortly establish how the current system at Edulab works.

The primary activity students engage in at Edulab's online learning platform is answering tasks. The tasks can be either multiple-choice or writing an answer in an input field. The multiple-choice tasks are structured such that each task always has five possible answers, while the free form input fields can only contain a number. Each task is associated with a lesson e.g. "Equations with a single unknown, plus minus", which is the lesson introducing them to equations with only one unknown, containing only simple plus or minus operations. A lesson is a categorized element consisting of a video-clip introducing the content and related tasks. Each lesson is further associated to a topic, e.g. "Equations with unknowns" which is more a general element than the lesson.

Edulab's platform contains a collection of over 1 million tasks. Each task is therefore not "handmade", and there is no meta-information related to the task on what different erroneous answers can indicate of possible underlying mathematical difficulty related to equations. Despite the lack of meta-information, we hypothesize that with a large number of erroneous answered tasks from the users on the portal, we can infer what the erroneous answers might indicate of student difficulties.

Distribution of erroneous answers

Before we elaborate on how we intend to use the erroneous answers, we first investigate how the erroneous answers are distributed for each task to ensure that not all erroneous answers are equally likely to be chosen by the students, as this will require us to infer the meaning of a very large number of erroneous answers. If some erroneous answers are very unlikely, we will not try to utilize these, as the information for these are very rare.

We will focus on the distribution of erroneous answers for a lesson related to equations, which only contain multiple-choice tasks. We use log data collected for the school year 2018-2019 for students attending 6^{th} and 7^{th} grade. The log contains 457,185 answers to tasks, provided by 37,585 students, in the lesson containing the equation tasks mentioned in table 1, distributed across 3,416 tasks. We remove all tasks that received less than 150 answers leaving us 197 tasks and 379,315 answers. To investigate the distribution of erroneous answers we apply the following for each task:

1. Compute the number of times each answer is answered by the students for each task
2. Normalize it to a probability distribution
3. Sort the probabilities from most probable to least probable (On all answers the most probable answer is the correct answer)

We are left with a probability distribution for each task, which we want to cluster, to see how the answers of the students vary across tasks. To do the clustering we use affinity propagation (Frey & Dueck, 2007), which finds candidate "exemplars" of the data points, which can be used as a general example for a cluster. We use standard parameters with damping being 0.5, and the preference is the median of the affinities. For the affinity measure, we use the well-known Jensen-Shannon divergence, as the data points we wish to cluster are probability distributions. Doing this we find 3 clusters:

1. Cluster 1 are tasks that primarily are always answered correctly, with very few errors.
2. Cluster 2 are tasks that have 1 particular wrong answer, which receives most of the erroneous answers, when a student chooses an incorrect answer.
3. Cluster 3 are tasks that have 2-3 wrong answers, which receive most of all the erroneous answers, when a student chooses an incorrect answer.

Cluster 1 have 88 of the tasks, cluster 2 have 39 and cluster 3 have 70. Our assumption that not all erroneous answers are equally likely is thereby supported.

We have done a similar analysis on a collection of other lessons, which also have input field answers, which shows that the answers from input fields also are focused on few very likely erroneous answers, and a large tail distribution of almost random answers.

Co-occurrence of erroneous answers

In the previous section we established that not all erroneous answers are equally likely, therefore the erroneous answers are focused on a smaller subset of possible erroneous answers. In the following, we establish how the erroneous answers are going to be utilized.

Each erroneous answer to a task gets a unique ID, if the erroneous answer has received more than some threshold, $\boldsymbol{P}$, of the answers for the given task. All erroneous answers, which occur often, will therefore have a unique ID, while we ignore the erroneous answers, which receive little attention. The reason for this is twofold, if all erroneous answers get an ID, the space of erroneous answers will be very large, and it is thus difficult to infer the meaning of each erroneous answer as they occur rarely.

Doing this, each student will have a sequence of erroneous answers which have been given an ID, $e_1, e_2, \dots, e_k$, ordered by when the student gave the answer. Based on this sequence we can now construct a co-occurrence matrix, which is a MxM where M is the number of erroneous answer IDs. Each row and column correspond to an erroneous answer, and the entry at the i'th row, j'th column is a counter of how often the i'th erroneous answer occur together with the j'th erroneous answer in a student sequence. What it means for two IDs to occur together is a matter of choice, e.g. if there is a sequence of student errors over a long time horizon, then the definition of two erroneous answers to occur together can depend on the time between the two erroneous answers. By doing this, only erroneous answers, which occur closely together in time, are considered a pair. On the other hand, all IDs for the same student can also be considered as occurring together if only a small number of lessons are considered, or the time horizon is small.

We use data for a single lesson (the same as in the previous section) which focuses on equations to construct the co-occurrence matrix, where we set the threshold $\boldsymbol{P}$=5%. We only include students who have made more than three mistakes on the lesson. This is early work, and for the final work, we wish to include a long series of lessons related to equations, but these are currently being deployed, and the data needs to be collected.

For this project we have no labels and are therefore limited to unsupervised learning (Hastie, Tibshirani, & Friedman, 2009), which is the paradigm of machine learning of learning some underlying structure of the data. As an initial experiment, we wanted to investigate if there was any clusters in the co-occurrence matrix, as this would indicate the existence of errors that mostly occurred together.

To explore the existence of clusters, we employed t-SNE (Maaten & Hinton, 2008), which is a powerful non-linear clustering technique, which tries to find a mapping for each data points to a new space, such that elements which are close in the original space, are also closed in the mapped space.

This exploration did not reveal any clusters, and we therefore did not manage to find any errors that occur primarily together.

When we get the full dataset, where students interact with equations over a much larger variety of lessons, we will repeat the following exploration. This is further elaborated on in the final section of the paper. In addition, the co-occurrence matrix allows for embedding based strategies of the errors. An embedding of an error would be some function $F(e_i)$ which maps the error to some real space of dimension N. The embedding would be such that errors who occur often together will be more similar than errors than does not occur together. This can be done using the algorithm Glove (Pennington, Socher, & Manning, 2014) which is used widely for finding embeddings for words, and work directly on a co-occurrence structure like ours.

CONCLUSION AND FUTURE DESIGN

The main focus of this paper was to answer the question; to what extent a categorization of standard textbook linear equations could serve as a mean for generating clusters of students with a large amount of data. Unfortunately, the clusters are not yet found. In the following section, we elaborate on the next step of the research design in order to accomplish finding the clusters.

To link the content of this paper to the overall research question, the following section will describe the future possibilities for finding clusters. With the right strategy, it will be possible to push these new tasks to the users of the platform, both for the teachers to assign to their students and for the students to explore on their own. In table 2 is presented a new possible categorization of the content on the platform that involves solving linear equations. This content already exists but have not yet been put to full use, meaning that the amount of answers given to these tasks are yet too sparse.

Label	Form		Type	Steps	Number of tasks
A	$x + a = b,$	$a \in N, b \in Z\backslash\{0\}$	Arithmetical	One-step	910
B	$ax = b,$	$a, b \in N$	Arithmetical	One-step	467
C	$x - a = b,$	$a \in N, b \in Z\backslash\{0\}$	Arithmetical	One-step	432
D	$a - x = b,$	$a \in N, b \in Z\backslash\{0\}$	Arithmetical	One-step	493
E	$ax + b = c,$	$a, b, c \in N$	Arithmetical	Two-step	431
F	$\frac{x}{a} = b,$	$a \in N, b \in Z\backslash\{0\}$	Abstract Arith.	One-step	426
G	$x \cdot \frac{a}{b} = c,$	$a, c \in Z\backslash\{0\}, b \in N$	Abstract Arith.	One-/two-step	587
H	$\frac{a}{x+b} = c,$	$a, c \in Z\backslash\{0\}, b \in N$	Abstract Arith.	Two-/three-step	755
I	$ax + b = cx + d,$	$a, c \in N\ b, d \in Z\backslash\{0\}$	Algebraic	Two-/three-step	DBT
J	$ax + b = cx + d,$	$a, c \in Z\ b, d \in Z\backslash\{0\}$	Algebraic	Two-/three-step	DBT

Table 2 – Future categorization of linear equations on the online platform

The future goal is to use the same clustering approach but with the new, more refined variety of linear equations. Furthermore, every type of equation comes in five different difficulty levels. This gives us another opportunity for selection and distinction. This means that instead of just having the 10 categories presented in table 2 we possibly have 50 levels of distinction instead of the 5 we had for

the first iteration of the clustering attempt. We believe that having 50 levels of distinction between the different types of linear equations will be a step in the right direction in order to achieve the student clustering. As mentioned in the methodological considerations having this categorization of the types of linear equations will serve as a mean to analyse and interpret the clusters when found.

FINAL REMARKS

The overall research aim of the project is to provide Edulab with a tool to support teachers in their teaching. The utilization of the vast amount of data Edulab are able to collect shall pave the way for the development of this assessment tool. The project's vision is that Danish mathematics teachers, based on an easy accessible formative assessment source, can be given a unique opportunity to organize, plan and complete their teaching (Palm, Andersson, Boström, & Vingsle, 2017). The future goal is to set up a continuous categorization of all students using Edulab's platform in the 7^{th} grade in order to "catch" students with conceptual or strategical difficulties when working with linear equations on the platform.

Inspired by the work presented in Linsell (2009), the goal is to develop a series of online questions in order to uncover the strategies the students use or the transformational difficulties the students experience, when working with the different types of linear equations. Together with the categorization of students, the second goal is to provide teachers with information of students' strategies as well as typical errors their students make when solving different types of equations. This information should be in the form of auto-generated 'formative assessment reports' for each student that the tool has identified and verified as having conceptual difficulties working with linear equations, describing also the student's behaviour and the difficulties present based on research findings on the identified difficulties. The formative assessment report should contain procedures and guidelines for how the teacher can approach remediation of the student's identified difficulties.

An alternate way to proceed could be to provide the student suggestions for teaching materials on the online learning platform for further learning on their own. The suggestions could be video lectures addressing their concrete subject in which they have difficulties.

REFERENCES

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences*, *13*(1), 1–14.

Booth, J. L., McGinn, K. M., Barbieri, C., & Young, L. K. (2017). Misconceptions and learning algebra. In S. Stewart (Ed.), *And the rest is just algebra* (pp. 63–78). Cham: Springer.

Booth, L. (1984). *Algebra: Children's strategies and errors. A report of the strategies and errors in secondary mathematics project.* Windsor, Birkshire: NFER-NELSON Publishing company Ltd.

Filloy, E., & Rojano, T. (1989). Solving Equations: The Transition from Arithmetic to Algebra. *For the Learning of Mathematics*, *9*(2), 19–25.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485–585). New York: Springer.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics*, *27*(1), 59–78.

Kieran, C. (1981). Concepts Associated with the Equality Symbol. *Educational Studies in Mathematics*, *12*(3), 317–326.

Kieran, C. (1985). The equation-solving errors of novice and intermidiate algebra students. In L. Streefland (Ed.), *Proceedings of the Annual Conference of the International Group for the Psychology of Mathematics Education* (pp. 141–146). Noordwijkerhout, The Netherlands.

Kieran, C. (1992). The Learning and Teaching of School Algebra. In D. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 390–419). New York: Macmillan Publishing Company.

Küchemann, D. (1981). Algebra. In K. Hart (Ed.), *Children's Understanding of Mathematics: 11-16* (pp. 102–119). London: Murray.

Linsell, C. (2007). Solving equations: Students' algebraic thinking. *Findings from the New Zealand Secondary Numeracy Project 2007*, 39–44.

Linsell, C. (2009). A Hierarchy of Strategies for Solving Linear Equations. In R. Hunter, B. Bicknell, & T. Burgess (Eds.), *Crossing divides: Proceedings of the 32nd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 331–338). Palmerston North, NZ: MERGA.

Maaten, L. van der, & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Palm, T., Andersson, C., Boström, E., & Vingsle, C. (2017). A review of the impact of formative assessment on student achievement in mathematics. *Nordic Studies in Mathematics Education*, *22*(3), 25–50.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Rhine, S., Harrington, R., & Starr, C. (2018). *How Students Think When Doing Algebra*. Charlotte, NC: Information Age Publishing, Incorporated.

Vlassis, J. (2002). The balance model: Hindrance or support for the solving of linear equations with one unknown. *Educational Studies in Mathematics*, *49*(3), 341–359.

DOI: 10.17185/duepublico/70760
URN: urn:nbn:de:hbz:464-20191119-145351-8