Matching Questions with Learning Material

Blinded

Abstract

Learning analytic systems allow the amount of time spent on the system to be recorded and these times used to predict performance on subsequent assessments. Ideally it is desirable to show the relationship between time spent on specific sections of the learning material with the accuracy on associated questions. However, for many assessments there is not information showing which items rely on which parts of the material. Our approach is to examine the lexical similarity of the text in the items with the text in the learning material using Pearson's correlation. We do this with three data sets: ACT® reading test, ACT® science test, and an online university biology course. This approach was very accurate for the two ACT assessments, AUCs of .92 and .99. The diagnosticity for the biology course was lower with AUC = .72. These results show that lexical matching can be used, but cautiously, to map items to content, with the potential to provide more finely grained time-on-task analyses and more granular content-based interventions.

Keywords: matching, text analysis, personalized learning, learning analytics, higher education

1 Introduction

There is a large body of learning analytics research focused on the time spent during learning activities (Kovanovic, Gašević, Dawson, Joksimovic, & Baker, 2016; Merceron, Blikstein, & Siemens, 2016). This is particularly true in higher education, where trace data are increasingly used to assign interventions and monitor student performance (Järvelä, Malmberg, Haataja, Sobocinski, & Kirschner, 2019; Winne, 2020). Much of this research has focused on aspects of time on task, such as the use of count *vs.* continuous time measures, evidence of distraction, and various methods for validating meaningful time spent learning. Time on task analyses could be more effective if they showed the relationship between time spent on specific aspects of learning material and performance accuracy on assessment items.

A first step toward this goal is to examine the usefulness of analytic techniques that can map the lexical similarity of assessment items to learning materials. Time spent on material related to particular items can be used for prediction and to assign early interventions at a more granular level, leveraging course design to facilitate analytics (Hernández-Leo, Martinez-Maldonado, Pardo, Muñoz-Cristóbal, & Rodríguez-Triana, 2019). A fundamental principle of learning design is aligning learning outcomes, learning material, and assessments items. Examining design-based alignment between items and content using text-mining approaches can provide preliminary evidence of the extent to which lexical similarly of item text and learning material can be used to improve time on task type analyses.

This study explores the use of a bag-of-words text matching technique (Benoit et al., 2018) to provide diagnostic data on the alignment of assessments and learning materials. As an initial test of the technique, we examined the lexical similarity of assessment items with reading and science passages taken from the ACT. This allows us to examine the lexical match between items specifically developed for stand-alone passages. These results were compared with a naturalistic examination of quiz items and related content taken from a post-secondary biology course. The lexical similarity of items from end-of-module quizzes were examined in relationship to module content. The desire was to determine if items could be correctly assigned to passages/module content based on lexical similarity.

2 Materials

Two practice tests from ACT® were used. They are available at `cdn2.hubspot.net/hubfs/360031/ACT-2015-16.pdf` (accessed November, 2020). The reading and science sections were used as these have questions corresponding with particular passages. The reading test had one section corresponding with two passages from Ray Bradbury. Five questions were about one of these passages, three about the other, and two about both. This seemed a good test of our approach to see if it would accurately differentiate between passages by the same author. Some information was removed from the lexical similarity search (e.g., line numbers, question numbers, response alternative letters, the sources for the materials) and words that had been split over two lines with a hyphen (i.e., *bad breaks*) were connected.

Modular lab content across 12 weeks from a Biology course were pulled. The course was structured to offer a laboratory preparation lecture and then a quiz on that content before the start of the lab the next class. There were eight lab-sessions with full quizzes associated with content (slide content) that were pulled and converted to .txt files. The slide content were also pulled and converted to text files. Each lab quiz was designed as a knowledge check on the previous laboratory session content and as such was hypothesized to match accordingly onto the slide content. All materials are available at `https://github.com/***BLINDED`.

3 Analytic Approach

There are many approaches that can be used to analyze text data (Bécue-Bertaut, 2018; Benoit et al., 2018; Grimmer & Stewart, 2013; Jaspal, 2020; Silge & Robinson, 2016). One of the simplest types of quantitative techniques uses the bag-of-words approach, where each word is a unit for analysis. Adjoining words (e.g., word pairs) and syntax are not considered. This allows the comparison of lexical similarity between different sets of texts. We opted not to use longer strings of words because the syntax of learning materials and questions are likely to be different and this can lead to discrepancies. The software R (R Core Team, 2019) was used for data analysis. It has several packages appropriate for text analysis. Three packages are of particular note: **quanteda** (Benoit et al., 2018), **tm** (Feinerer, Hornik, & Meyer, 2008), and **tidytext** (Silge & Robinson, 2016) because they offer broad frameworks for processing text data. Other software (e.g., Python) could also be used and the descriptions given here should be sufficient for well-versed users of those systems to implement these procedures. The code used to create all the statistical analyses and plots for this paper is available at `https://github.com/***BLINDED`.

Welbers, Von Atteveldt, and Benoit (2017) describe best practice for preparing data for text analyses and their guidelines were followed. First, html code, numbers, *etc.* were removed. With bag-of-word approaches the norm is to transform the words in several ways. The data were trimmed (white space removed) and made lower case using the `stri_trim` and `stri_trans_tolower` functions from the **stringi** package (Gagolewski, 2020), respectively. The goal is to create a document-feature-matrix, or dfm, showing the frequency for each word used for each source (in this paper learning materials and test items). For a simple example, consider the dfm of three conversations that might be recorded during meals and a few of the words used shown in Table 1.

Table 1: An example document-feature-matrix (dfm).

		Food Words					
		Cereal	Milk	Coffee	Sandwich	Pizza	$\cdots$
Conversations	Breakfast	5	2	8	1	0	$\cdots$
	Lunch	0	2	6	5	3	$\cdots$
	Dinner	0	0	4	0	8	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$

The `dfm` function from the **quanteda** (Benoit et al., 2018) package is used to create dfms in this paper. The options `tolower`, `stem`, `remove_punct`, and `remove = stopwords("english")` were used so that the procedure was not case-sensitive, the stems were compared (so *pizza* and *pizzas* are treated the same), punctuation was removed, and stop words like "the" were not considered. These are standard procedures (Welbers et al., 2017). The dfms used in these analyses have the number of columns equal to the number of unique word stems used (not including stop words) and the number of rows equal to the number of sets of learning modules plus the number of questions.

The next task is estimating the similarity between each row of the dfm. There are several metrics available in R and elsewhere (Ashby & Ennis, 2007; Enflo, 2020). Many of these are described at `https://cran.r-project.org/web/packages/proxy/vignettes/`

`overview.pdf`. The ones available in the **proxy** package are for binary, nominal, and metric measures. At the time of writing there are 49 metrics that can be used plus you can write your own. To see these metrics, plus the primary publication and a brief description of each, attach the **proxy** (Meyer & Buchta, 2019) package and type: `as.data.frame(pr_DB)[,12:13]` within R. `pr_DB` is a registry in **proxy** of these similarity (and dissimilarity) metrics. Because of the value of encouraging others to adopt these proposals, the two most well known measures for similarity are considered: Pearson's product moment correlation between the variables for word use of each source with each item and the cosine of the angle between these variables. For the data in this paper these produced similar values and lead to the same conclusions (analyses with other metrics on these sources are also available from the authors).

Function	Formula	R code
Correlation	$\frac{\sum(x_i - \overline{x})\ (y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}}$	`cor(x,y)`
Cosine	$\frac{\sum x\, y}{\sqrt{\sum x^2\ \sum y^2}}$	`sum(x*y)/sqrt(sum(x^2)*sum(y^2))`

Note: $x\,y$ and x^2 are element-wise multiplication: $\sum x\,y = (x_1\,y_1) + (x_2\,y_2) + \cdots + (x_n\,y_n)$.

These correlations are used to predict whether the question is associated with each section of the learning materials. Because it is known for these stimuli which questions go with which part of the learning material, the predictive value of the similarity measures can be evaluated. We call their associate learning material their true match. After discussing the correlation matrices, the empirical receiver operating characteristics, or ROCs, are plotted. They show the diagnostic value for the similarity values to decide if an item draws information from a particular learning material passage. They plot the cumulative proportion of correct matches (the hit rate) with the cumulative proportion of making an incorrect classification (the false alarm rate) for values of the similarity measure. ROCs are often used within the context of signal detection theory, which can be formally presented as logistic (or probit) regressions (DeCarlo, 1998; Wright, Horry, & Skagerberg, 2009). Because each correlation is of a different passage with a different question, the variability both of passages and questions should be taken into account. The intent was to use a multilevel cross-classified logistic regression (Goldstein, 2011; Wright & London, 2009). These are complex models and this can lead to computational difficulties. When this occurs alternatives (e.g., using dummy variables for passages and questions) can be used which yield more reliable results.

4 Results

Document-feature-matrices (dfms) were created for the ACT Reading, the ACT Science, and the Biology course using the `dfm` function from **quanteda** (Benoit et al., 2018). From these, the correlations were calculated between each section of the learning material with each question using the `textstat_simil` function, also from **quanteda**. Tables 2 and 3 show the correlations for the ACT reading and science. Those representing accurate matches are highlighted in yellow. The biology correlation matrix is much larger and available from the authors.

Suppose that to declare an item is assigned to a section it has to have $r > .1$ and be the largest one for those materials. For the reading matrix (Table 2), ignoring the Bradbury passages, 24 of the 30 items were assigned to their correct passage and none of the remaining six were assigned incorrectly (these six were not assigned to any). Four of the five items related to the first Bradbury passage are assigned to this one, and one errantly assigned to the wrong Bradbury passage. The two items related to the second Bradbury passage are correctly assigned to this one. Of the three related to both Bradbury passages, one is assigned to the first, one to the second, and one is not assigned to any passage. Of the 40 science items, 39 are correctly assigned to the associated passage. One is incorrectly assigned to a different passage. Thus, for the ACT materials there are only two false assignments (one being a Bradbury question to the wrong Bradbury passage), and about 10% where no

Table 2: Pearson correlations for the Reading Passage. Highlighted cells show right.

	Art Deco	Sargasso	Brad A	Brad B	Trap-Jaw
Item 1	0.23	-0.04	-0.04	-0.03	-0.03
Item 2	0.29	-0.01	0.03	-0.01	-0.03
Item 3	0.26	-0.00	0.01	-0.02	-0.02
Item 4	0.06	-0.04	-0.02	-0.03	-0.03
Item 5	0.11	-0.04	0.01	-0.02	-0.04
Item 6	-0.00	-0.03	0.09	-0.03	-0.00
Item 7	0.24	-0.05	-0.01	-0.03	-0.04
Item 8	0.20	-0.04	0.01	0.00	0.00
Item 9	-0.01	-0.01	-0.01	-0.03	0.02
Item 10	0.26	0.00	0.01	-0.00	-0.03
Item 11	-0.04	-0.03	-0.01	-0.03	-0.02
Item 12	-0.01	0.30	-0.03	-0.02	-0.02
Item 13	0.01	0.25	-0.00	0.00	-0.04
Item 14	-0.05	0.20	-0.01	-0.02	-0.04
Item 15	-0.00	0.14	-0.03	-0.01	-0.04
Item 16	-0.03	0.35	-0.04	-0.02	-0.03
Item 17	-0.02	0.19	-0.03	-0.02	-0.00
Item 18	-0.02	-0.02	-0.01	-0.03	0.01
Item 19	-0.03	0.32	-0.04	-0.03	-0.02
Item 20	-0.05	0.18	-0.04	-0.00	-0.04
Item 21	-0.05	-0.07	0.13	-0.03	-0.02
Item 22	0.03	-0.01	0.21	-0.02	0.03
Item 23	-0.03	-0.03	0.26	0.02	-0.01
Item 24	-0.00	-0.02	0.13	-0.02	-0.02
Item 25	-0.02	-0.01	0.12	0.38	-0.02
Item 26	-0.01	-0.03	0.03	0.32	-0.03
Item 27	-0.02	-0.04	0.03	0.37	-0.04
Item 28	-0.02	-0.04	0.05	-0.03	-0.04
Item 29	-0.03	-0.00	0.21	0.02	-0.01
Item 30	-0.01	-0.05	0.06	0.10	-0.02
Item 31	-0.04	-0.04	-0.03	-0.03	0.55
Item 32	-0.04	-0.04	-0.02	-0.02	-0.04
Item 33	-0.04	-0.04	-0.02	-0.02	0.11
Item 34	-0.04	-0.03	-0.04	-0.00	0.46
Item 35	-0.05	-0.03	-0.03	-0.02	0.32
Item 36	-0.04	-0.02	-0.04	-0.03	0.54
Item 37	0.01	-0.03	0.10	-0.00	0.00
Item 38	-0.04	-0.04	-0.04	-0.02	0.28
Item 39	-0.05	-0.01	-0.03	-0.02	0.41
Item 40	-0.03	-0.04	-0.03	-0.02	0.50

assignment is made.

The correlations were less diagnostic for the biology course. Of the 100 questions, assign-

Table 3: Pearson correlations for the Science Passage. Highlighted cells show right.

	Passage I	Passage II	Passage III	Passage IV	Passage V	Passage VI
Item 1	0.45	-0.01	0.04	0.19	0.04	0.16
Item 2	0.13	-0.02	0.02	0.10	0.00	-0.01
Item 3	0.32	-0.01	0.02	0.13	0.05	0.11
Item 4	0.34	-0.02	-0.06	-0.03	-0.02	0.04
Item 5	0.45	-0.02	-0.04	0.00	-0.03	-0.02
Item 6	0.55	-0.03	-0.03	0.09	0.02	0.03
Item 7	0.45	-0.01	0.03	0.16	0.04	0.07
Item 8	0.02	0.35	0.07	0.16	0.03	0.08
Item 9	0.05	0.46	0.04	0.10	0.01	0.05
Item 10	-0.03	0.52	0.04	0.11	0.03	-0.00
Item 11	0.04	0.47	0.04	0.21	0.03	0.09
Item 12	-0.01	0.61	0.02	0.05	-0.01	0.02
Item 13	-0.04	0.83	-0.03	0.01	-0.02	0.00
Item 14	-0.05	0.40	-0.06	-0.03	-0.02	-0.03
Item 15	-0.02	-0.02	0.33	0.04	-0.02	0.05
Item 16	-0.03	-0.01	0.42	0.03	-0.02	0.01
Item 17	-0.01	-0.02	0.23	0.05	-0.02	0.10
Item 18	-0.01	-0.02	0.31	0.04	-0.02	0.07
Item 19	-0.02	-0.02	0.42	0.02	-0.02	0.08
Item 20	-0.05	-0.03	0.37	-0.03	-0.03	-0.03
Item 21	-0.05	-0.02	-0.03	0.30	-0.03	-0.03
Item 22	-0.01	-0.00	-0.00	0.54	-0.02	-0.00
Item 23	-0.02	-0.02	0.06	0.19	-0.02	0.03
Item 24	0.00	-0.00	0.03	0.52	0.04	0.03
Item 25	-0.02	0.02	0.10	0.38	-0.02	0.12
Item 26	0.02	0.01	0.03	0.46	-0.02	0.01
Item 27	0.03	0.00	-0.01	0.15	0.27	0.03
Item 28	-0.05	-0.01	-0.05	-0.03	0.12	-0.01
Item 29	-0.02	-0.01	0.02	0.06	0.21	-0.00
Item 30	0.03	-0.01	0.03	0.11	0.27	0.05
Item 31	0.02	-0.03	-0.00	0.11	0.25	0.01
Item 32	-0.03	-0.02	-0.03	0.05	0.19	-0.01
Item 33	0.00	-0.01	-0.02	0.09	0.23	0.03
Item 34	0.05	-0.01	0.20	0.01	-0.03	0.09
Item 35	-0.04	-0.02	-0.02	0.01	-0.03	0.52
Item 36	0.08	-0.01	-0.04	-0.01	-0.02	0.28
Item 37	0.00	-0.02	-0.04	-0.03	-0.03	0.28
Item 38	0.10	-0.02	0.03	0.03	-0.01	0.20
Item 39	-0.03	0.02	0.01	-0.02	-0.02	0.27
Item 40	0.15	-0.00	0.24	0.24	0.03	0.40

ments were made for only 78 of them if using the $r > .1$ and the largest r decision criterion. Half of these were for correct matches and half were not. With 8 passages it means that the correct matching is much greater than for an arbitrary incorrect match, but it may be desirable either to combine this with other matching procedures or to have instructors make clearer what the individual questions are asking. These were for quizzes at the end of modules, and given this context much may have been assumed.

The differences in correlations for true matches with non-matches can be presented graphical. Two methods will be used. First, Figure 1 shows the boxplots for the correlations for the three sets of materials, divided by whether they are for a true match or an non-match. For ACT materials the differences are striking. While there are a few low correlations for correct matches, almost all the non-matches had correlations near zero. For the biology courses the difference is not as evident.

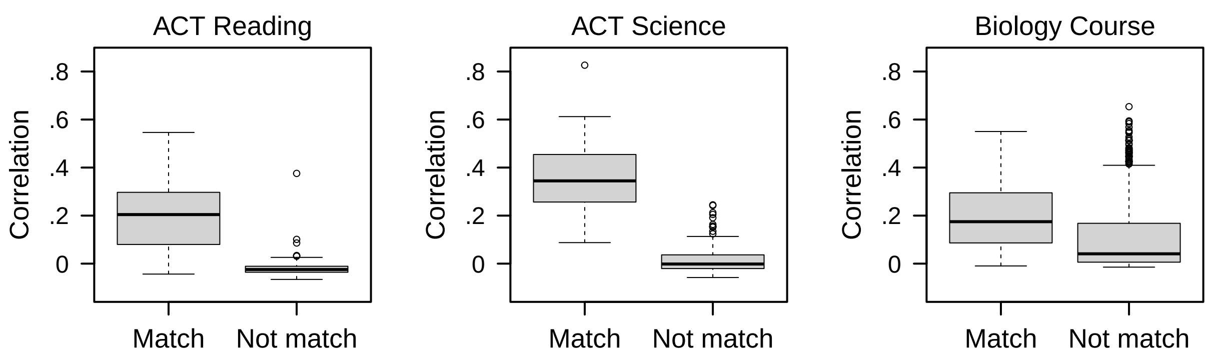

Figure 1: Boxplots for the Reading (ACT), Science (ACT), and Biology course, comparing correlations for correct matches and the incorrect matches.

The second graphical method is using receiver operating characteristics (ROCs). The area-under-the-curve, or AUC, is a common statistic for showing how diagnostic the measure is. These are shown in Figure 2. AUC values range from 0 to 1 with .5 corresponding to this hit rate and false alarm rates being equal (the diagonal line has $AUC = .5$). The values observed here are very high for the two ACT tests (for reading $AUC = 0.92$; for science $AUC = 0.99$) and lower for the biology course ($AUC = 0.72$).

Several statistical models were estimated for the relationship between a correct match and the correlation. The intent was to use a cross-classified model treating both passage and question items as random variables. These models had computed variances estimates at or near zero. When computational issues arise with multiple random variables it is often

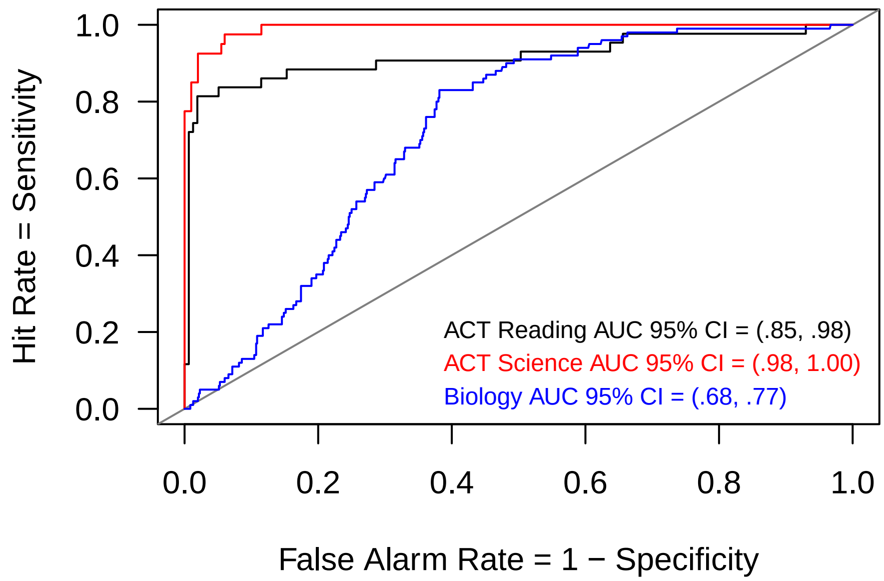

Figure 2: Receiver operating characteristics (ROCs) for the three sets using the correlation metric.

useful to treat them as fixed variables (Wright, 2017). This was done for these three sets of materials using a logistic regression with passages and items included as sets of dummy variables. The variable for the correlations was added to the model and for each set of materials the fit improved. All of the AUCs are significantly above $AUC = .5$.

$$\begin{aligned} &\text{For reading:} && \chi^2(1) = 133.18, p < .001 \\ &\text{For science:} && \chi^2(1) = 216.02, p < .001 \\ &\text{For biology:} && \chi^2(1) = 147.64, p < .001 \end{aligned}$$

5 Summary and Implications

The results showed that the bag-of-words approach performed well assigning ACT items to their corresponding passages. However, the naturalistic test of the biology items/content was less accurate than for the ACT material. The results provide proof-of-concept evidence that lexical matching can be used to map items to content, with the potential to provide more finely grained time-on-task analyses and more granular content-based interventions. In a learning situation where there is high lexical similarity between items associated with some given content (and low similarity between other reading) it is feasible to conduct inferential time on task tests with some degree of certainty that the lexical similarity of the item-to-content overlap is mutually exclusive. For example, students who spend less time on a specific passage could be assigned additional review materials with some degree of certainty that content would prepare them for improved success in a particular set of assessment items.

However, the high level of accuracy of items assignment achieved within the context of the ACT was not replicated using the course-based materials, where the likelihood of making an incorrect assignment of an item to a passage was higher. Without clear lexical overlap, there are concerns solely using this text mining approach to provide targeted learning supports based on time spent on specific content, or vice versa using item performance to refer students back to specific course content. These findings suggest that for the bag-of-words approach to work within online learning contexts, considerable attention must be paid to the learning design of course content, particularly with regard to the alignment of assessment items and content. These findings are in line with calls to leverage learning design to facilitate the use of learning analytics to personalize instruction and learning supports (Hernández-Leo et al., 2019; Lockyer & Dawson, 2011).

6 Declarations

- The ACT materials are publicly available at `cdn2.hubspot.net/hubfs/360031/ACT-2015-16.pdf`. The biology course data and other materials are available at `https://github.com/**BLINDED`.
- The authors have no competing interests.
- No funding specific to this project was received. DW and SW receive funding as part of an endowment from the Dunn Family Foundation.

- The initial idea for the project came from discussions between JH and DW. All authors planned the research. SW and DW prepared the materials for the ACT data and SW, EA, and DW prepared the biology data. Access to these data was negotiated by JH. DW did the statistical analysis and prepared the initial draft, that was then worked on by all authors.

- Daniel Wright is the Dunn Family Endowed Chair and Professor of Educational Assessment. Sarah Wells is in the Assessment and Quantitative Analysis (AQUA in Ed) PhD stream at UNLV and a graduate assistant. Jonathan Hilpert is Associate Professor of Learning Analytics. Elham Arabi is a learning consultant at the World Health Organization (WHO). She earned her PhD in Interaction & Media Sciences at UNLV.

References

Ashby, F. G., & Ennis, D. M. (2007). Similarity measures. *Scholarpedia*, *2*(12), 4116. doi: 10.4249/scholarpedia.4116

Bécue-Bertaut, M. (2018). *Textual data science with r*. Boca Raton, FL: CRC Press.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). **quanteda**: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. doi: 10.21105/joss.00774

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205. doi: 10.1037/1082-989X.3.2.186

Enflo, K. (2020). Measures of similarity. *Theoria*, *86*, 73–99. doi: https://doi.org/10.1111/theo.12222

Feinerer, I., Hornik, K., & Meyer, D. (2008, March). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54. Retrieved from https://www.jstatsoft.org/v25/i05/

Gagolewski, M. (2020). R package **stringi**: Character string processing facilities [Computer software manual]. Retrieved from http://www.gagolewski.com/software/stringi/

Goldstein, H. (2011). *Multilevel statistical models (4th ed.)*. Chichester, UK: Wiley.

Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*, 267–297.

Hernández-Leo, D., Martinez-Maldonado, R., Pardo, A., Muñoz-Cristóbal, J. A., & Rodríguez-Triana, M. J. (2019). Analytics for learning design: A layered framework and tools. *British Journal of Educational Technology*, *50*, 139–152. doi: 10.1111/bjet.12645

Järvelä, S., Malmberg, J., Haataja, E., Sobocinski, M., & Kirschner, P. A. (2019). What multimodal data can tell us about the students' regulation of their learning process? *Learning and Instruction*, **, ***-***. doi: j.learninstruc.2019.04.004

Jaspal, R. (2020). Content analysis, thematic analysis and discourse analysis. In G. M. Breakwell, D. B. Wright, & J. Barnett (Eds.), *Research methods in psychology* (5th ed., pp. 285–312). London, UK: Sage Publications.

Kovanovic, V., Gašević, D., Dawson, S., Joksimovic, S., & Baker, R. (2016). Does time-on-task estimation matter? implications on validity of learning analytics findings. *Journal of Learning Analytics*, *2*(3), 81–110. doi: 10.18608/jla.2015.23.6

Lockyer, L., & Dawson, S. (2011). Learning designs and learning analytics. In *Proceedings of the 1st international conference on learning analytics and knowledge* (pp. 153–156). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/2090116.2090140

Merceron, A., Blikstein, P., & Siemens, G. (2016). Learning analytics: From big data to meaningful data. *Journal of Learning Analytics*, *2*(3), 4–8. doi: 10.18608/jla.2015.23.2

Meyer, D., & Buchta, C. (2019). **proxy**: Distance and similarity measures [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=proxy` (R package version 0.4-23)

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Silge, J., & Robinson, D. (2016). **tidytext**: Text mining and analysis using tidy data principles in R. *Journal of Open Source Software*, *1*(3). Retrieved from http://dx.doi.org/10.21105/joss.00037 doi: 10.21105/joss.00037

Welbers, K., Von Atteveldt, W., & Benoit, K. (2017). Text analysis in R. *Communication Methods and Measures*, *11*, 245–265. doi: 10.1080/19312458.2017.1387238

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior*, *112*, 106457. doi: 10.1016/j.chb.2020.106457

Wright, D. B. (2017). Some limits using random slope models to measure academic growth. *Frontiers in Education*, *2*(58), 1–11. doi: 10.3389/feduc.2017.00058

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, *41*, 257–267. doi: 10.3758/BRM.41.2.257

Wright, D. B., & London, K. (2009). Multilevel modelling: Beyond the basic applications. *British Journal of Mathematical and Statistical Psychology*, *62*, 439–456.