Visualizing Unknown and Missing Data in the HarvardX-MITx 2013 Dataset

Introduction

In the past several years, technology-based learning has placed learning analytics at the forefront of educational research and development (Ferguson). In the sphere of learning analytics, Massive Online Open Courses stand at a vantage point, as stakeholders can easily collect a plethora of data from their users. However, large datasets often contain missing data or inconsistencies, which can lead to time consuming data cleaning procedures or, worse, inaccurate analysis or modeling derived from ignoring or being unaware of these inconsistencies. In order for useful information to be gained from large datasets, the limitations of the dataset must be understood (Broeck et. al).

The dashboard created in this projects seeks to provide stakeholders with an overview of the limitations of the HarvardX-MITx 2013 De-Identified Dataset, with specific areas displayed to highlight where gaps in the data exist. The overall goal beyond specific information gained from the particular dataset is to provide an example of supplementing public datasets with a basic dashboard visualization, so those wishing to use the data can save time and produce more accurate insights.

Learning objective

At a high level, the overall learning objective is to analyze the amount of data that is unknown or missing in the dataset. More specifically, the following questions are addressed:

(1) Analysis and comparison of courses in terms of how much education (‘LoE’), year of birth (‘YoB’), or gender (‘gender’) data is not available
(2) Analysis and comparison of courses in terms of how much data is labeled as internally inconsistent
(3) Analysis and comparison of courses in terms of how much data is labeled as geographically ‘Unknown/Other’

These questions are analyzed in the form of two dashboards, with design features and rationale explained in the ‘Design and method’ section. One dashboard features an overall view of the courses with the most unknown or missing data, while the other dashboard features a complete and detailed view of missing data over all of the courses.

User and context

This report and dashboards produced can be used by stakeholders that wish to perform analysis or build models off of this dataset. This can include data scientists who are trying to build predictive models off of the data, MOOC developers who are trying to classify certain types of users, and stakeholders that are in charge of MOOC data collection infrastructure. Instead of having to produce this preliminary analysis on their own, stakeholders can easily visualize the dataset before they go in and begin their own work. For example, a stakeholder wishing to build a predictive model off of the data will have an understanding of where the limitations to the model may be, based off of results from the dashboard, as they will see what percentage of the data is missing or inconsistent. This is helpful not only in saving time for the stakeholder, but also in increasing the accuracy of the analysis produced, because the stakeholder is less likely to make false assumptions in creating the model.

Because this data is accessible to the public, the corresponding dashboards produced were made publicly available so that anyone in the future wishing to use the dataset can begin with a higher level of understanding of the data; this means that even those who are not MOOC stakeholders but are just trying to learn from using the dataset also benefit from the dashboards. For example, students learning about learning analytics (such as future CMS.594 students!) can use these dashboards as a springboard to more complicated examination.

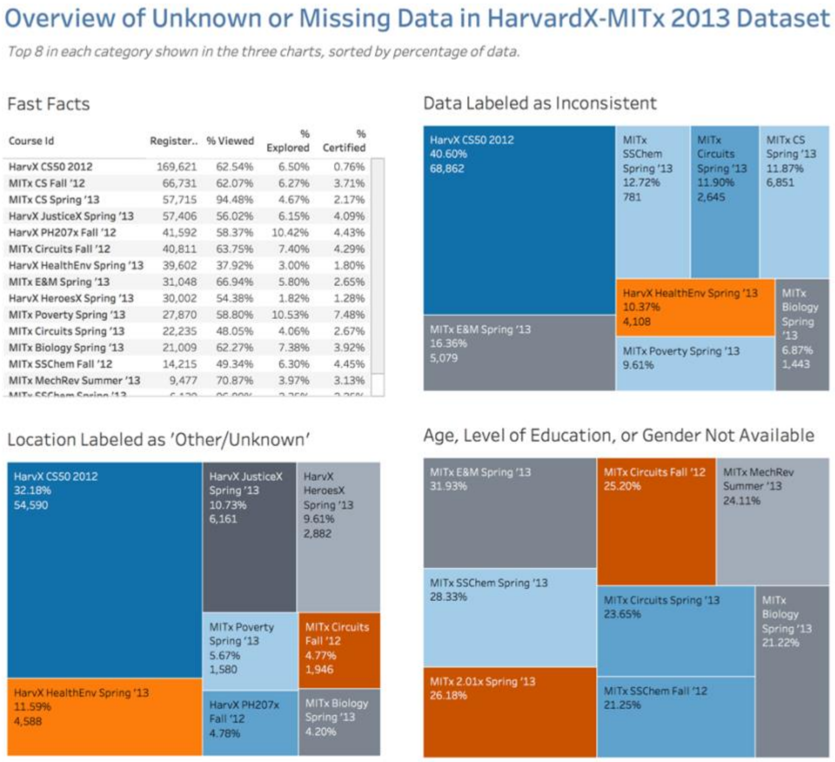
Design and method

Data Selection and Labeling:
To address the three learning objectives, the following steps were applied to the data:

- Empty cells and 'NA' cells were set as metrics in deciding if the data was unavailable
    - In the original dataset, data was labeled as 'NA' if the student created an edX account before that registration question was available. In this analysis, both empty cells and 'NA' cells are considered to be the same level of data unavailable, because they both represent information that cannot be found in the present dataset.
    - Originally, the missing/incomplete data was split into education data ('LoE'), year of birth data ('YoB'), and gender data ('gender'). However, it was found that in most instances, a row missing any one of these three metrics was likely to also be missing the other two. The final chart under 'Age, Level of Education, or Gender Not Available' display the instances where any one of these three metrics is missing, given the understanding that it is likely all three are missing at the same time.
- Course IDs were renamed from their Course Code to their Short Title, found on page 2 of the dataset documentation. For example, '14.73x' was changed to 'Poverty'. 'HarvardX' was shortened to 'HarvX", and the course dates were shortened from '_Year_Semester' format to 'Semester 'Yr' format. For example, 'MITxCircuits_2012_Fall' was changed to 'MITx Circuits Fall '12.' These changes were done for greater understanding of what courses were being displayed, while still keeping the length of the Course IDs reasonable.
- From the counts of missing/incomplete data, the data was converted to percentages. This allowed for a greater proportional understanding, and both counts and percentages were included in the dashboard.

Medium Selection:
Dashboard visualization was selected as the medium to present the data. While there are summary tables presented in the 'Results' section, these summary tables may only be helpful for a very specific application; utilizing the dynamic features of dashboarding allows for a wider range of audiences to gain insight from the data. Each visual in the dashboard is a filter, which allows the user to drill down to the specific portion of the data that they are interested in.

Two dashboards were created, one presenting an overall summary view with the top 8 courses in each category displayed (Fig. 1), and one presenting a more detailed view with bar charts and all courses displayed (Fig. 2).



Overview of Unknown or Missing Data in HarvardX-MITx 2013 Dataset

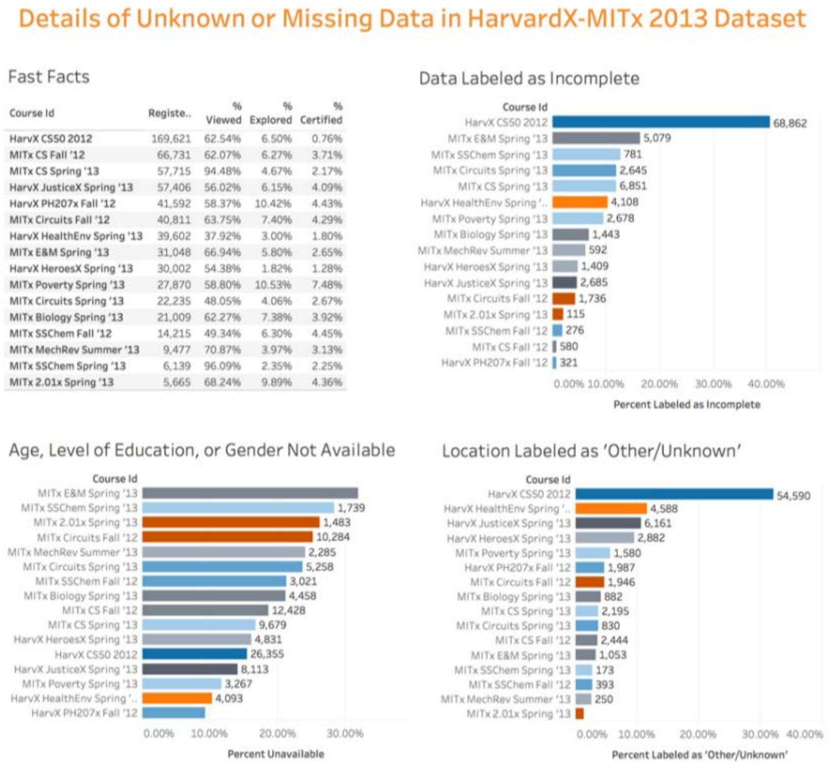
Top 8 in each category shown in the three charts, sorted by percentage of data.

Fast Facts

Course Id	Register..	% Viewed	% Explored	% Certified
HarvX CS50 2012	169,621	62.54%	6.50%	0.76%
MITx CS Fall '12	66,731	62.07%	6.27%	3.71%
MITx CS Spring '13	57,715	94.48%	4.67%	2.17%
HarvX JusticeX Spring '13	57,406	56.02%	6.15%	4.09%
HarvX PH207x Fall '12	41,592	58.37%	10.42%	4.43%
MITx Circuits Fall '12	40,811	63.75%	7.40%	4.29%
HarvX HealthEnv Spring '13	39,602	37.92%	3.00%	1.80%
MITx E&M Spring '13	31,048	66.94%	5.80%	2.65%
HarvX HeroesX Spring '13	30,002	54.38%	1.82%	1.28%
MITx Poverty Spring '13	27,870	58.80%	10.53%	7.48%
MITx Circuits Spring '13	22,235	48.05%	4.06%	2.67%
MITx Biology Spring '13	21,009	62.27%	7.38%	3.92%
MITx SSChem Fall '12	14,215	49.34%	6.30%	4.45%
MITx MechRev Summer '13	9,477	70.87%	3.97%	3.13%

Data Labeled as Inconsistent
HarvX CS50 2012 40.60% 68,862
MITx SSChem Spring '13 12.72% 781
MITx Circuits Spring '13 11.90% 2,645
MITx CS Spring '13 11.87% 6,851
HarvX HealthEnv Spring '13 10.37% 4,108
MITx Biology Spring '13 6.87% 1,443
MITx E&M Spring '13 16.36% 5,079
MITx Poverty Spring '13 9.61%
Location Labeled as 'Other/Unknown'
HarvX CS50 2012 32.18% 54,590
HarvX JusticeX Spring '13 10.73% 6,161
HarvX HeroesX Spring '13 9.61% 2,882
MITx Poverty Spring '13 5.67% 1,580
MITx Circuits Fall '12 4.77% 1,946
HarvX HealthEnv Spring '13 11.59% 4,588
HarvX PH207x Fall '12 4.78%
MITx Biology Spring '13 4.20%
Age, Level of Education, or Gender Not Available
MITx E&M Spring '13 31.93%
MITx Circuits Fall '12 25.20%
MITx MechRev Summer '13 24.11%
MITx SSChem Spring '13 28.33%
MITx Circuits Spring '13 23.65%
MITx Biology Spring '13 21.22%
MITx 2.01x Spring '13 26.18%
MITx SSChem Fall '12 21.25%

Figure 1: Summary view of dashboard, displaying top 8 courses with unknown or missing data in each category, as well as a fast facts table.



Details of Unknown or Missing Data in HarvardX-MITx 2013 Dataset

Fast Facts

Course Id	Registe..	% Viewed	% Explored	% Certified
HarvX CS50 2012	169,621	62.54%	6.50%	0.76%
MITx CS Fall '12	66,731	62.07%	6.27%	3.71%
MITx CS Spring '13	57,715	94.48%	4.67%	2.17%
HarvX JusticeX Spring '13	57,406	56.02%	6.15%	4.09%
HarvX PH207x Fall '12	41,592	58.37%	10.42%	4.43%
MITx Circuits Fall '12	40,811	63.75%	7.40%	4.29%
HarvX HealthEnv Spring '13	39,602	37.92%	3.00%	1.80%
MITx E&M Spring '13	31,048	66.94%	5.80%	2.65%
HarvX HeroesX Spring '13	30,002	54.38%	1.82%	1.28%
MITx Poverty Spring '13	27,870	58.80%	10.53%	7.48%
MITx Circuits Spring '13	22,235	48.05%	4.06%	2.67%
MITx Biology Spring '13	21,009	62.27%	7.38%	3.92%
MITx SSChem Fall '12	14,215	49.34%	6.30%	4.45%
MITx MechRev Summer '13	9,477	70.87%	3.97%	3.13%
MITx SSChem Spring '13	6,139	96.09%	2.35%	2.25%
MITx 2.01x Spring '13	5,665	68.24%	9.89%	4.36%

Data Labeled as Incomplete
Course Id
HarvX CS50 2012 68,862
MITx E&M Spring '13 5,079
MITx SSChem Spring '13 781
MITx Circuits Spring '13 2,645
MITx CS Spring '13 6,851
HarvX HealthEnv Spring '.. 4,108
MITx Poverty Spring '13 2,678
MITx Biology Spring '13 1,443
MITx MechRev Summer '13 592
HarvX HeroesX Spring '13 1,409
HarvX JusticeX Spring '13 2,685
MITx Circuits Fall '12 1,736
MITx 2.01x Spring '13 115
MITx SSChem Fall '12 276
MITx CS Fall '12 580
HarvX PH207x Fall '12 321
0.00% 10.00% 20.00% 30.00% 40.00%
Percent Labeled as Incomplete
Age, Level of Education, or Gender Not Available
Course Id
MITx E&M Spring '13
MITx SSChem Spring '13 1,739
MITx 2.01x Spring '13 1,483
MITx Circuits Fall '12 10,284
MITx MechRev Summer '13 2,285
MITx Circuits Spring '13 5,258
MITx SSChem Fall '12 3,021
MITx Biology Spring '13 4,458
MITx CS Fall '12 12,428
MITx CS Spring '13 9,679
HarvX HeroesX Spring '13 4,831
HarvX CS50 2012 26,355
HarvX JusticeX Spring '13 8,113
MITx Poverty Spring '13 3,267
HarvX HealthEnv Spring '.. 4,093
HarvX PH207x Fall '12
0.00% 10.00% 20.00% 30.00%
Percent Unavailable
Location Labeled as 'Other/Unknown'
Course Id
HarvX CS50 2012 54,590
HarvX HealthEnv Spring '.. 4,588
HarvX JusticeX Spring '13 6,161
HarvX HeroesX Spring '13 2,882
MITx Poverty Spring '13 1,580
HarvX PH207x Fall '12 1,987
MITx Circuits Fall '12 1,946
MITx Biology Spring '13 882
MITx CS Spring '13 2,195
MITx Circuits Spring '13 830
MITx CS Fall '12 2,444
MITx E&M Spring '13 1,053
MITx SSChem Spring '13 173
MITx SSChem Fall '12 393
MITx MechRev Summer '13 250
MITx 2.01x Spring '13
0.00% 10.00% 20.00% 30.00% 40.00%
Percent Labeled as 'Other/Unknown'

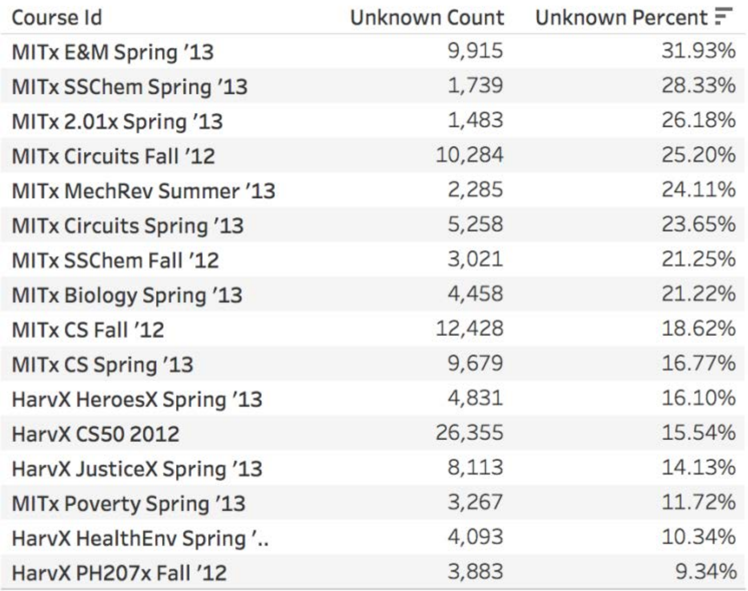
Figure 2: Detailed view of dashboard, displaying all courses in bar chart form with all charts and tables serving as filters for the dashboard.

The intended workflow for a stakeholder would be to first view the overview dashboard for a global understanding of the missing or incomplete data, and then to use the detailed dashboard to drilldown and gain specific information. When mousing over the bars in the dashboard, both the counts and specific value of the percentage show up, for those wishing to acquire more detailed numerical information. The color palettes selected are friendly to those who are color blind, and each course ID has a specific color that stays consistent within the two dashboards for ease of understanding and switching between the two views.

Results

Below we present three results, with Table 1 addressing RQ1, Table 2 addressing RQ2, and Table 3 addressing RQ 3.

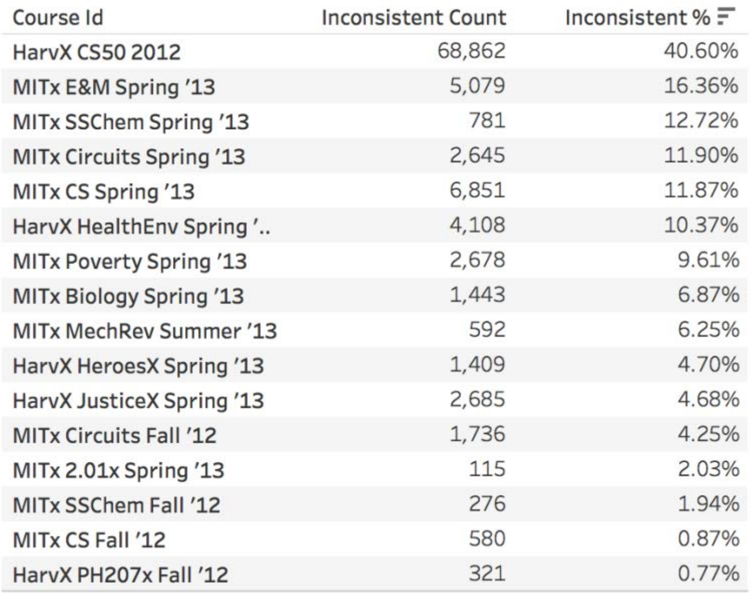
We find that for RQ 1, the MITx courses tend to have a higher percentage of data that is unavailable or unknown. A takeaway from this insight that a stakeholder could have is to depend less on the gender, education, or year of birth data when building a model for MITx courses, as the MITx course data tends to have more missing.



Course Id	Unknown Count	Unknown Percent
MITx E&M Spring '13	9,915	31.93%
MITx SSChem Spring '13	1,739	28.33%
MITx 2.01x Spring '13	1,483	26.18%
MITx Circuits Fall '12	10,284	25.20%
MITx MechRev Summer '13	2,285	24.11%
MITx Circuits Spring '13	5,258	23.65%
MITx SSChem Fall '12	3,021	21.25%
MITx Biology Spring '13	4,458	21.22%
MITx CS Fall '12	12,428	18.62%
MITx CS Spring '13	9,679	16.77%
HarvX HeroesX Spring '13	4,831	16.10%
HarvX CS50 2012	26,355	15.54%
HarvX JusticeX Spring '13	8,113	14.13%
MITx Poverty Spring '13	3,267	11.72%
HarvX HealthEnv Spring '..	4,093	10.34%
HarvX PH207x Fall '12	3,883	9.34%

Table 1. Table of courses with percent and amount of unknown education, year of birth, or gender data shown, sorted by percent unknown.

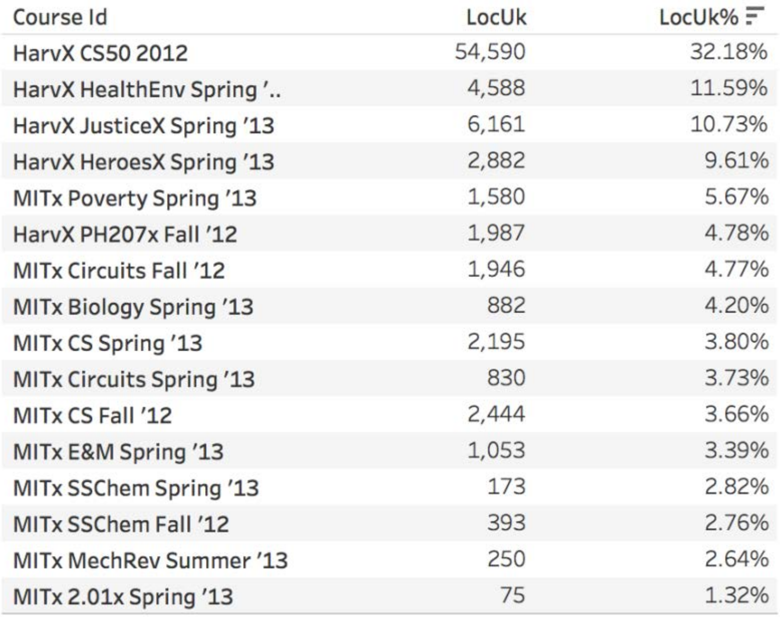
We find that for RQ 2, the Harvard 2012 Computer Science Class ('HarvX CS50 2012') has, by far, proportionally the most data that is labeled as internally inconsistent. The data documentation pointed out that this class in particular had a lot of internal inconsistencies, so this result was expected. Another takeaway for RQ 2 is that institution does not seem to affect internal inconsistencies.



Course Id	Inconsistent Count	Inconsistent %
HarvX CS50 2012	68,862	40.60%
MITx E&M Spring ’13	5,079	16.36%
MITx SSChem Spring ’13	781	12.72%
MITx Circuits Spring ’13	2,645	11.90%
MITx CS Spring ’13	6,851	11.87%
HarvX HealthEnv Spring ’..	4,108	10.37%
MITx Poverty Spring ’13	2,678	9.61%
MITx Biology Spring ’13	1,443	6.87%
MITx MechRev Summer ’13	592	6.25%
HarvX HeroesX Spring ’13	1,409	4.70%
HarvX JusticeX Spring ’13	2,685	4.68%
MITx Circuits Fall ’12	1,736	4.25%
MITx 2.01x Spring ’13	115	2.03%
MITx SSChem Fall ’12	276	1.94%
MITx CS Fall ’12	580	0.87%
HarvX PH207x Fall ’12	321	0.77%

Table 2. Table of courses with percent and amount of data labeled as internally inconsistent, sorted by percent unknown.

We find that for RQ 3, the Harvard 2012 Computer Science Class (‘HarvX CS50 2012’) has proportionally the most data that is geographically labeled ‘Unknown/Other’. This is important to those who are trying to consider geographical information when deriving insight on the dataset; it’s not possible to tell if the data is labeled as ‘Unknown/Other’ because the user’s location is actually not known, or because the user’s location is in an area deemed “other.”



Course Id	LocUk	LocUk%
HarvX CS50 2012	54,590	32.18%
HarvX HealthEnv Spring ’..	4,588	11.59%
HarvX JusticeX Spring ’13	6,161	10.73%
HarvX HeroesX Spring ’13	2,882	9.61%
MITx Poverty Spring ’13	1,580	5.67%
HarvX PH207x Fall ’12	1,987	4.78%
MITx Circuits Fall ’12	1,946	4.77%
MITx Biology Spring ’13	882	4.20%
MITx CS Spring ’13	2,195	3.80%
MITx Circuits Spring ’13	830	3.73%
MITx CS Fall ’12	2,444	3.66%
MITx E&M Spring ’13	1,053	3.39%
MITx SSChem Spring ’13	173	2.82%
MITx SSChem Fall ’12	393	2.76%
MITx MechRev Summer ’13	250	2.64%
MITx 2.01x Spring ’13	75	1.32%

Table 3. Table of courses with percent and amount of ‘Unknown/Other’ geographical data shown, sorted by percent unknown.

Reproducing your work

The dashboard was produced on Tableau Public, a software that is free for anyone to use. With these basic parameters set in place, one could imagine a similar analysis being performed on other public data sources, such that it becomes a norm for these starting point dashboards to accompany public data sources.

References

Broeck, J. V., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities. PLoS Medicine, 2(10). doi:10.1371/journal.pmed.0020267

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. International Journal of Technology Enhanced Learning, 4(5/6), 304-317. doi:10.1504/ijtel.2012.051816

Collinge, R. (2017, January 17). How to Design for Color Blindness. Retrieved from https://usabilla.com/blog/how-to-design-for-color-blindness/