Diversity in Children's Books 2018-2022

Mylinh Hamlington

2022-12-19

Case Study

An established publishing company is planning on starting a new imprint that will publish children's literature. The company focuses on publishing books by and about Black, Indigenous, and People of Color, as well as other diverse aspects like disability, LGBTQIA+ and religion. They have hired a Data Analyst to explore diverse books currently being published for children. Their inquiries are below:

1. **What is the current environment of diverse children's books published in the US?**
2. **What diverse subjects should we focus on for our first few titles?**
3. **How can we establish ourselves quickly as a diverse book publisher?**

The Data

The Cooperative Children's Book Center (CCBC) is a research center that is part of the University of Wisconson-Madison School of Education. They support teaching, learning, and research related to children and receives children's books for review and inclusion from publishers in the US (and one publisher in Canada.

Beginning in 2018, they started documenting the content and authors of each book receive as well as additional aspects of identity including disability, LGBTQIA+, and religion. Both test and illustrations are used to establish elements of diversity.

This data was pulled from the CCBC Book Search database, found here: https://ccbc.education.wisc.edu/recommended-books/

Citation: "Data on books by and about Black, Indigenous and People of Color compiled by the Cooperative Children's Book Center (CCBC), School of Education, University of Wisconsin-Madison, based on its work analyzing the content of books published for children and teens received by the CCBC annually." https://ccbc.education.wisc.edu/literature-resources/ccbc-diversity-statistics/books-by-about-poc-fnn/ Accessed December 9, 2022

The 21 diversity subjects that CCBC document are:

Arab, Asian, Black/African, Brown Skin Unspecified, Indigenous, Latinx, Middle East, Multicultural General, Pacific Islander, Christian, Jewish, Muslim, Other Religion, Gender Nonconformity, LGBTQ Character/Topic, LGBTQ Family, LGBTQ Innuendo, LGBTQ Non-Fiction, Cognitive/Neurological Disability/Condition, Physical Disability/Condition, Psychiatric Disability/Condition

Data Cleaning

Building the Environment

Add data

```
DiversityRpt <- read_csv("CCBC Diversity Report.csv")
```

Explore the data

```
dim(DiversityRpt)
```

```
## [1] 17223   106
```

```
head(DiversityRpt)
```

```
## # A tibble: 6 x 106
##      Id Title          Isbn  Year CallN~1 CcbcC~2 Genres NewEd~3 NonUs~4 Publi~5
##   <dbl> <chr>         <dbl> <dbl> <chr>   <chr>   <chr>  <chr>   <chr>   <chr>
## 1    11 Good Night~ 9.78e12  2018 Rey     Pictur~ Anima~ No      No      Hought~
## 2    12 So Light, ~ 9.78e12  2018 Strass~ Pictur~ Anima~ No      No      Charle~
## 3    13 My Rainbow~ 9.78e12  2018 Sklans~ Pictur~ Board~ No      No      Cartwh~
## 4    14 Wiggles     9.78e12  2018 Zucche~ Pictur~ Board~ No      No      Chroni~
## 5    15 Little Tru~ 9.78e12  2018 Gomi    Pictur~ Board~ No      No      Chroni~
## 6    16 Be Kind, B~ 9.78e12  2018 Schulz  Pictur~ Anima~ No      No      Simon ~
## # ... with 96 more variables: OriginallyPublishedIn <dbl>, Translated <chr>,
## #   TranslatedNotes <chr>, Subject <chr>, ‘Subject Disability Notes‘ <chr>,
## #   ‘Subject Heritage/Region Notes‘ <chr>, ‘Subject Lgbtq Notes‘ <chr>,
## #   ‘Subject Religion Notes‘ <chr>, ‘Primary Character 1‘ <chr>,
## #   ‘PrimaryCharacterDisabilityNotes 1‘ <chr>,
## #   ‘PrimaryCharacterHeritageNotes 1‘ <chr>,
## #   ‘PrimaryCharacterSexualityNotes 1‘ <chr>, ...
```

Because there are so many columns and we will only be using a few, we are going to pull out a selection to make the dataframe easier to handle.

```
DRSimple <- select(DiversityRpt, c('Title', 'Year', 'Isbn', 'CcbcCollection', 'Genres', 'Subject', 'Publ

glimpse(DRSimple)
```

```
## Rows: 17,223
## Columns: 7
## $ Title          <chr> "Good Night, Curious George", "So Light, So Heavy", "My~
## $ Year           <dbl> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2~
## $ Isbn           <dbl> 9.781329e+12, 9.781581e+12, 9.781338e+12, 9.781452e+12,~
## $ CcbcCollection <chr> "Picture Book", "Picture Book", "Picture Book", "Pictur~
## $ Genres         <chr> "Animal Fantasy, Board Book", "Animal Fantasy, Board Bo~
## $ Subject        <chr> NA, NA, "Multicultural General", NA, NA, NA, NA, NA, NA~
## $ Publisher      <chr> "Houghton Mifflin Harcourt", "Charlesbridge", "Cartwhee~
```

Clean the data

```
## duplicate values
distinct(DRSimple)
```

The number of distinct titles is the same as those in the original data (17,223), so there are no duplicates.

```
## missing values
apply(is.na(DRSimple), 2, sum)
```

```
##          Title           Year           Isbn CcbcCollection         Genres
##              0              0              9              0           1189
##        Subject      Publisher
##           8787             38
```

Missing ISBN are likely due to titles being added to the database before as an Advanced Reader's Copy before being officially published. This information as well as missing Genres could be found online and added to the data. Because we will not be using these elements for our analysis at this time, we will not spend time on completing these missing details.

Missing Subject elements indicate books that do not include any of the documented diversity subjects, so we are not concerned about having them in the data.

Missing Publishers will be found online and the data will be updated.

The elements that we will be using in this analysis seem to be clean. This is likely because they were entered by the CCBC research group.

Organize the data

Original Column Names

```
head(DRSimple)
```

```
## # A tibble: 6 x 7
##   Title                        Year          Isbn CcbcCo~1 Genres Subject Publi~2
##   <chr>                       <dbl>         <dbl> <chr>    <chr>  <chr>   <chr>
## 1 Good Night, Curious George   2018 9781328795915 Picture~ Anima~ <NA>    Hought~
## 2 So Light, So Heavy           2018 9781580898492 Picture~ Anima~ <NA>    Charle~
## 3 My Rainbow Surprise          2018 9781338110982 Picture~ Board~ Multic~ Cartwh~
## 4 Wiggles                      2018 9781452164755 Picture~ Board~ <NA>    Chroni~
## 5 Little Truck                 2018 9781452163000 Picture~ Board~ <NA>    Chroni~
## 6 Be Kind, Be Brave, Be You!   2018 9781534412514 Picture~ Anima~ <NA>    Simon ~
## # ... with abbreviated variable names 1: CcbcCollection, 2: Publisher
```

```
## Genre, Publisher have multiple entries, but unable to tell Subject by first 6 rows

head(DRSimple$Subject, n = 15)
```

```
##  [1] NA                                   NA
##  [3] "Multicultural General"              NA
##  [5] NA                                   NA
##  [7] NA                                  NA
##  [9] NA                                  "Brown Skin Unspecified"
## [11] NA                                  NA
## [13] NA                                  "Arab, Asian, Middle East, Muslim"
## [15] "Brown Skin Unspecified"
```

Genre, Publisher, and Subject have multiple values, so we are going to separate them to only have one a single value per observation.

Separating Columns

```
## max number of genre and subject columns

Number_of_Genres <- str_count(DRSimple$Genres, ",") + 1
DRSimple['NumberOfGenres'] <- Number_of_Genres


Number_of_Subject <- str_count(DRSimple$Subject, ",") + 1
DRSimple['NumberOfSubjects'] <- Number_of_Subject

## Separate publisher and imprint

DRSimple <- DRSimple %>%
  rename('PublisherImprint' = 'Publisher') %>%
  separate(PublisherImprint, c('Publisher', 'Imprint'))

## Separate Genres and Subjects

DRSimple$Subject <- strsplit(DRSimple$Subject, split = ', ')
DRSimple <- unnest(DRSimple, Subject)

DRSimple$Genres <- strsplit(DRSimple$Genres, split = ', ')
DRSimple <- unnest(DRSimple, Genres)
```

New Data Frame

Here is how our data frame looks now:

```
head(DRSimple)
```

```
## # A tibble: 6 x 10
##   Title      Year    Isbn CcbcC~1 Genres Subject Publi~2 Imprint Numbe~3 Numbe~4
##   <chr>     <dbl>   <dbl> <chr>   <chr>  <chr>   <chr>   <chr>     <dbl>   <dbl>
## 1 Good Nig~  2018 9.78e12 Pictur~ Anima~ <NA>    Hought~ Mifflin       2      NA
## 2 Good Nig~  2018 9.78e12 Pictur~ Board~ <NA>    Hought~ Mifflin       2      NA
## 3 So Light~  2018 9.78e12 Pictur~ Anima~ <NA>    Charle~ <NA>          3      NA
## 4 So Light~  2018 9.78e12 Pictur~ Board~ <NA>    Charle~ <NA>          3      NA
## 5 So Light~  2018 9.78e12 Pictur~ Conce~ <NA>    Charle~ <NA>          3      NA
## 6 My Rainb~  2018 9.78e12 Pictur~ Board~ Multic~ Cartwh~ Schola~       2       1
## # ... with abbreviated variable names 1: CcbcCollection, 2: Publisher,
## #   3: NumberOfGenres, 4: NumberOfSubjects
```

Analysis

1. What is the current environment of diverse children's books published in the US?

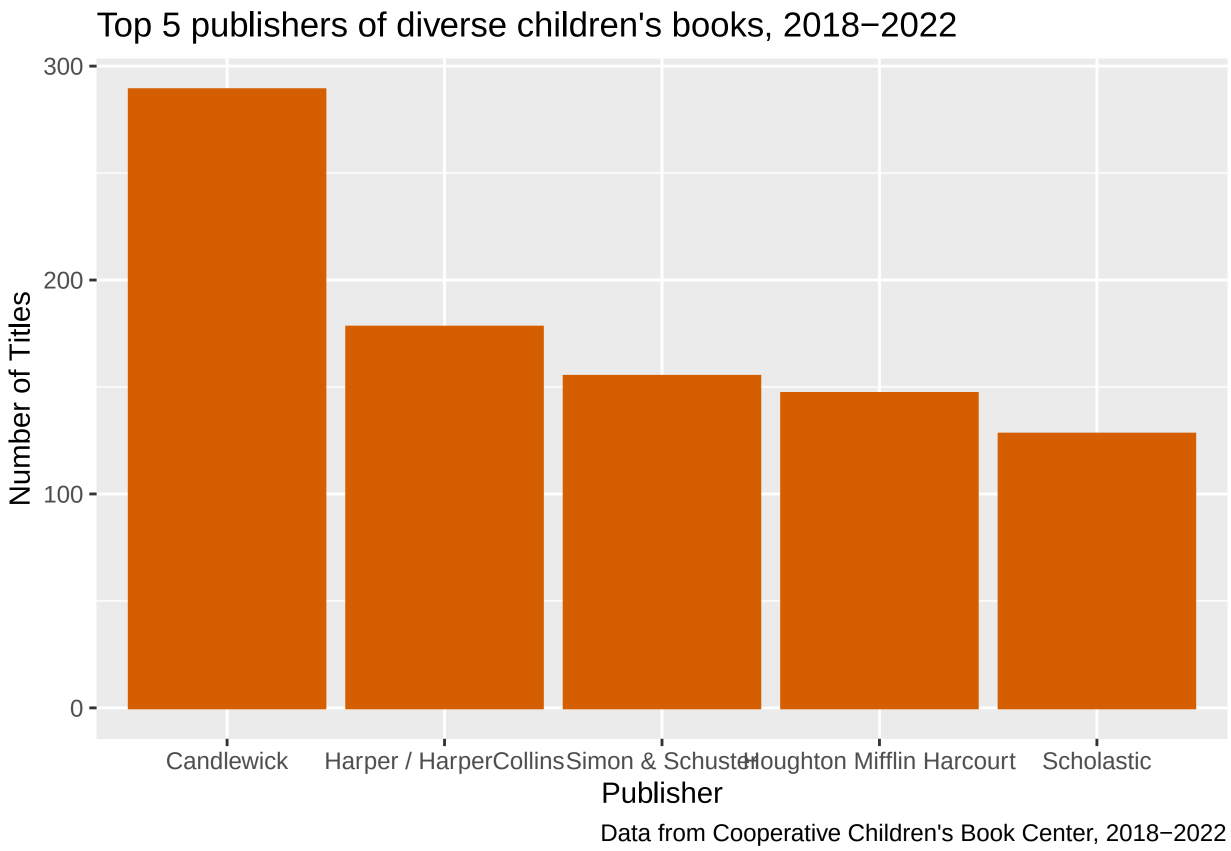
Current companies publishing diverse books

Find the total number of publishers and top 5 producers. Data is pulled from full DiversityRpt dataframe to view publisher and imprint together.

```
n_distinct(DiversityRpt$Publisher)
```

```
## [1] 867
```

```
## remove titles with N/A subjects (non-diverse)
TopPublishers <- DiversityRpt %>%
  select(Publisher, Subject) %>%
  na.omit()

TopPublishers <- data.frame(sort(table(TopPublishers$Publisher), decreasing = TRUE)[1:5])
TopPublishers <- TopPublishers %>%
  rename("Publisher" = "Var1",
         "Titles" = "Freq")
ggplot(TopPublishers, aes(Publisher, Titles))+
  geom_bar(stat = "identity", color = "#D55E00", fill = "#D55E00")+
  labs(title = "Top 5 publishers of diverse children's books, 2018-2022", y = "Number of Titles", capti
```



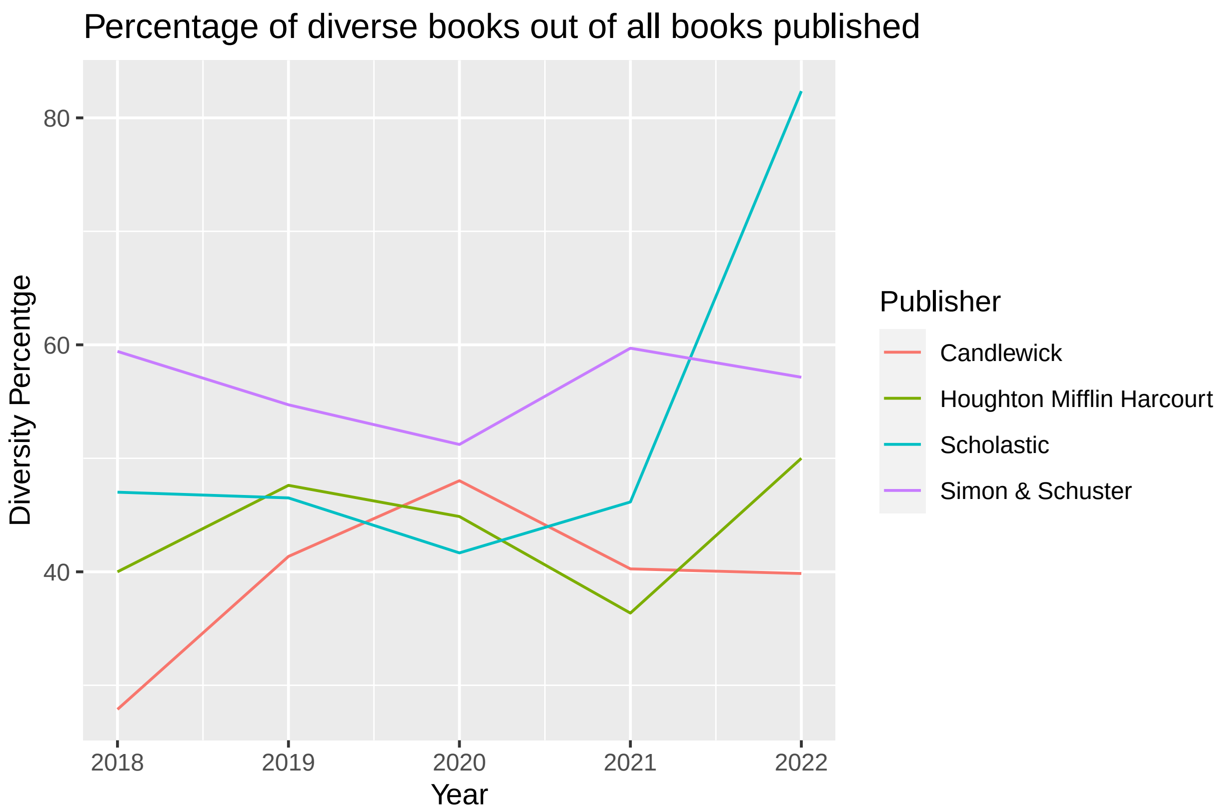
There are 867 total publishers and imprints which have published diverse books in the past 5 years. The top 5 publishers are:

- Candlewick
- Harper/HarperCollins
- Simon & Schuster
- Houghton Mifflin Harcourt
- Scholastic

Candlewick is far and away the most prolific publisher of diverse children's books.

Publisher's ratio of diverse books to non-diverse books

```
TitleCount <- DiversityRpt %>%
  select(Year, Publisher, Subject) %>%
  filter(Publisher %in% c('Candlewick', 'Harper/Harper Collins', 'Simon & Schuster', 'Houghton Mifflin

TitleRatio <- TitleCount %>%
  count(Year, Publisher) %>%
  rename("AllTitles" = "n")

TitleDiv <- TitleCount %>%
  filter(!is.na(Subject)) %>%
  count(Year,Publisher) %>%
  rename("DiverseTitles" = "n")

TitleRatio <- merge(TitleRatio, TitleDiv, by = c("Year", "Publisher"))

TitleRatio$Ratio <- 100 * (TitleRatio$DiverseTitles / TitleRatio$AllTitles)

ggplot(TitleRatio, aes(x = Year, y = Ratio, group = Publisher, color = Publisher)) +
  geom_line()+
  labs(title = "Percentage of diverse books out of all books published", y = "Diversity Percentge", cap
```

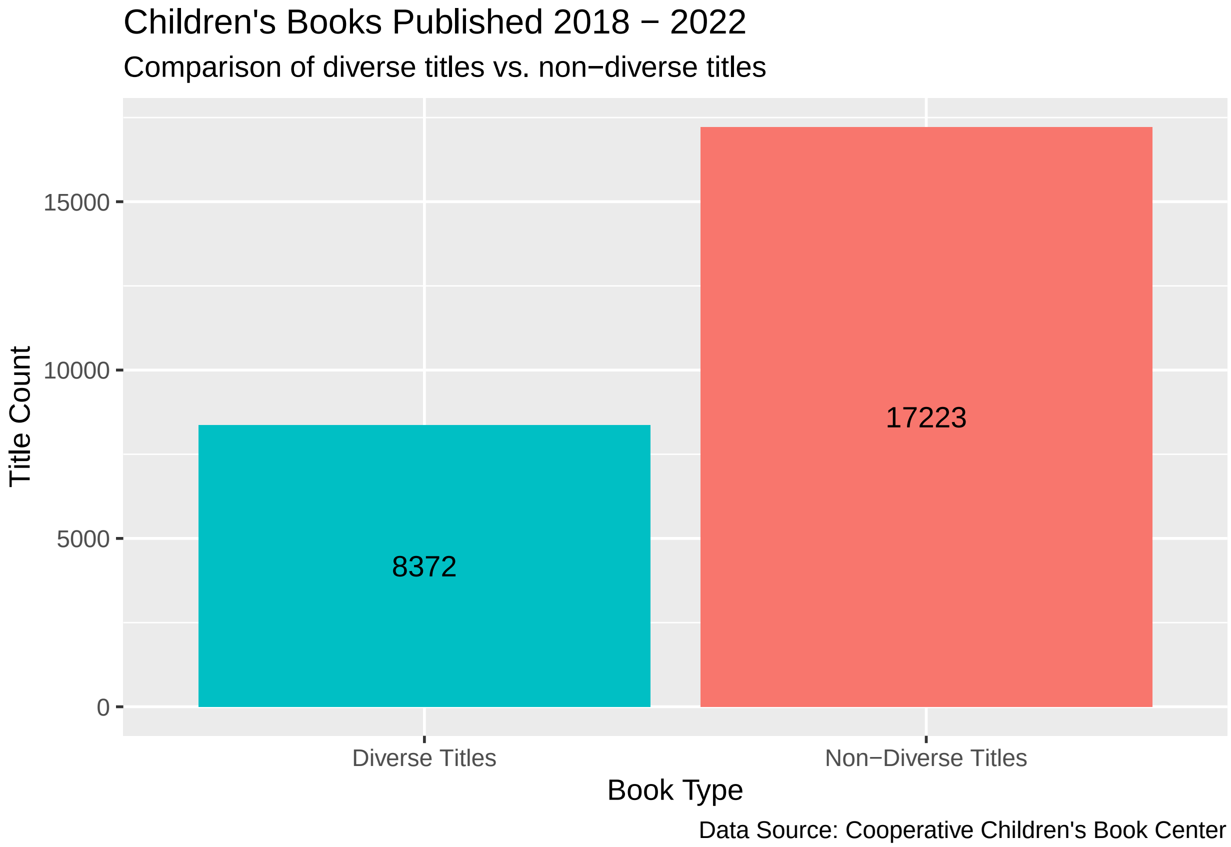


Data from Cooperative Children's Book Center

Although Candlewick has published the largest total number of diverse titles, Scholastic currently has the highest percentage of diverse titles to total titles published.

Number of Diverse Titles

```
DiverseTitles <- DRSimple %>%
  distinct(Title, .keep_all = T) %>%
  filter(!is.na(Subject))

TitleDiversity <- data.frame(group = c('Diverse Titles', 'Non-Diverse Titles'),
                             value = c(nrow(DiverseTitles), nrow(DiversityRpt)))

ggplot(TitleDiversity, aes(x=group, y=value, fill=group)) +
  geom_bar(stat="identity") +
  scale_fill_manual(values=c("#00BFC4", "#F8766D")) +
  geom_text(aes(label = value ),
            position = position_stack(vjust = 0.5)) +
  theme(legend.position = "none")+
  labs(x ="Book Type", y = "Title Count",
       title = "Children's Books Published 2018 - 2022",
       subtitle = "Comparison of diverse titles vs. non-diverse titles",
       caption = "Data Source: Cooperative Children's Book Center")
```


Children's Books Published 2018 − 2022

Comparison of diverse titles vs. non−diverse titles

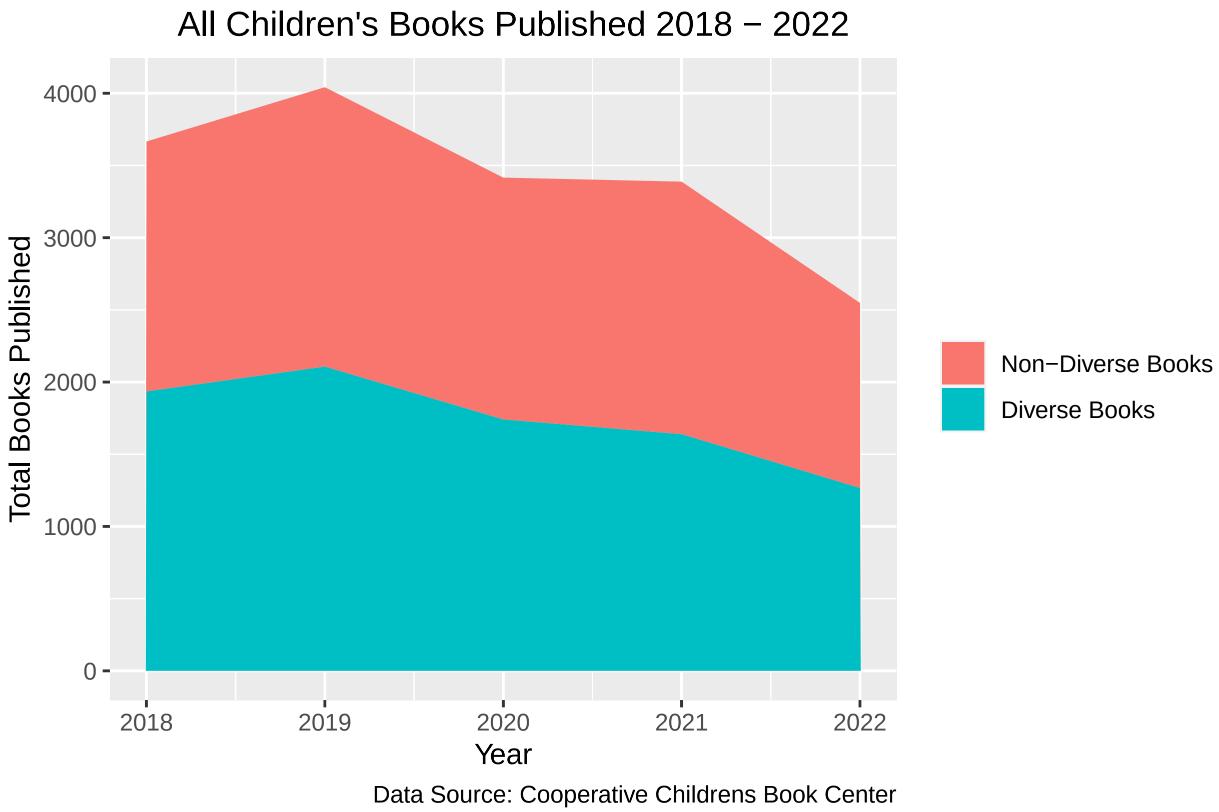
Data Source: Cooperative Children's Book Center

Less than 50% of all children's books published over the past 5 years have had some element of diversity. This means that more than 50% of the books published over the past 5 years appear to have been written by White, able-bodied, heterosexual, cisgendered authors and feature either an animal, non-living personified item, or White main character.

Annual totals of published diverse and non-diverse books

```
DRSimple$Diverse <- DRSimple$Title %in% DiverseTitles$Title

DiverseByYear <- DRSimple %>%
  distinct(Title, .keep_all = T) %>%
  select(Year,Diverse)

TotalByYear <- DiverseByYear %>%
  filter(Diverse == "FALSE") %>%
  group_by(Year, Diverse) %>%
  summarise('OtherBooks' = n())
TotalByYear <- TotalByYear[-6,]

DiverseByYear <- DiverseByYear %>%
  filter(Diverse == "TRUE") %>%
  group_by(Year, Diverse) %>%
  summarise('DiverseTotal' = n())
```

```
DiverseByYear$TotalBooks = TotalByYear$OtherBooks

DiverseByYear %>%
  pivot_longer(cols = c('DiverseTotal', 'TotalBooks'),
               names_to='group',
               values_to='total') %>%
  ggplot(aes(x=Year, y=total, fill=group)) +
  geom_area() +
  labs(title ="
       All Children's Books Published 2018 - 2022",
       y = 'Total Books Published', caption = "Data Source: Cooperative Childrens Book Center") +
  scale_fill_discrete(labels = c("Non-Diverse Books", "Diverse Books"))+
  theme(legend.title= element_blank())
```

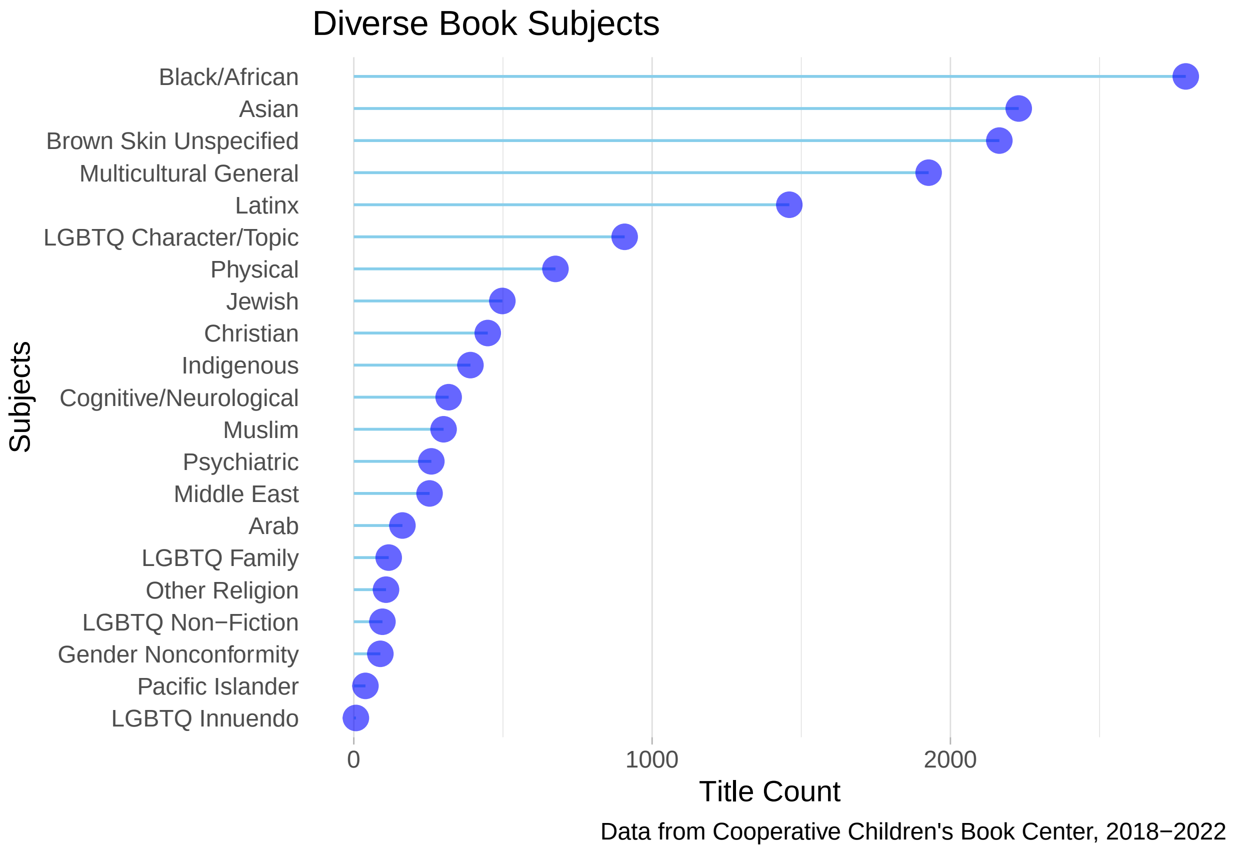


The ratio of non-diverse to diverse books published each year continues to be around 50%. The dip in total books published starting in 2020 could be attributed to the beginning of the COVID-19 pandemic, but further analysis into the publishing industry would need to be done before making that conclusion.

2. What diverse subjects should we focus on for our first few titles?

Most published subjects

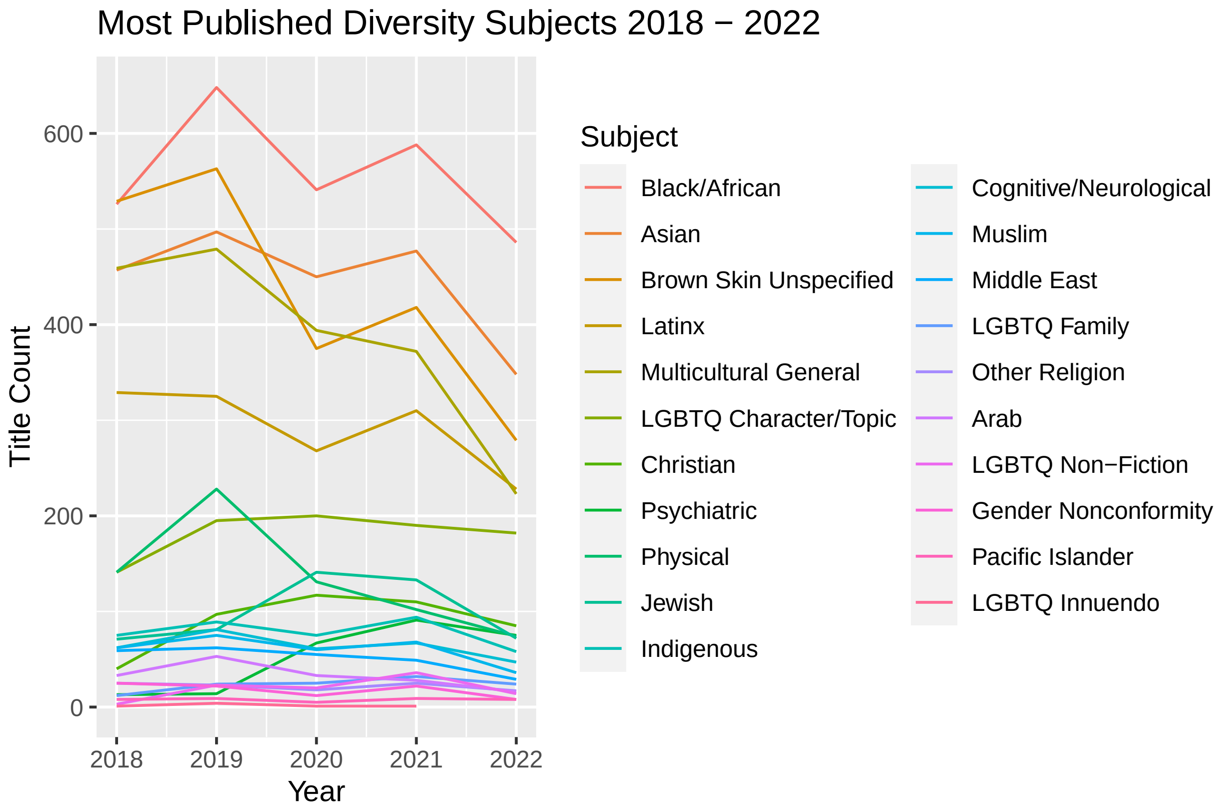
```r
DRSCount <- DRSimple %>%
  group_by(Subject) %>%
  drop_na(Subject) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

DRSCount %>%
  ggplot(aes(x = reorder(Subject, count), y = count)) +
  geom_segment(aes
             (x = reorder(Subject,count), xend =reorder(Subject,count),
              y = 0, yend = count), color = "skyblue") +
  geom_point(color = "blue", size = 4, alpha = 0.6)+
  coord_flip()+
  theme_light()+
  theme(
    panel.grid.major.y = element_blank(),
    panel.border = element_blank(),
    axis.ticks.y = element_blank())+
  labs(title = 'Diverse Book Subjects', x = 'Subjects', y = 'Title Count',
  caption = "Data from Cooperative Children's Book Center, 2018-2022")
```



This plot is measuring every subject in every title for the past 5 years. This means that some titles are counted more than once if they feature multiple subjects. As we can see, the top subject is Black/African. After the top 5 race or country of origin based subjects, the subject title count from LGBTQIA+ characters onwards declines rapidly. This tells us that diverse books featuring general multicultral issues, Black/African, Asian, and Latinx characters or authors are the most widely published. Books featuring other diverse aspects like gender, sexual orientation, physical and psychological challenges, religion, and Indigenous, Arab, and Pacific Islander backgrounds are published much less often.

Most popular subject over time

```
Pop <- DRSimple %>%
  select(Year, Subject) %>%
  drop_na(Subject) %>%
  group_by(Year, Subject) %>%
  summarize(count = n()) %>%
  arrange(desc(count))

Pop %>%
  ggplot(aes(x=Year, y=count, color=fct_reorder2(Subject, Year, count))) +
  geom_line() +
  labs(title = 'Most Published Diversity Subjects 2018 - 2022',
       y = 'Title Count', color = 'Subject',
       caption = "Data Source: Cooperative Children's Book Center")
```



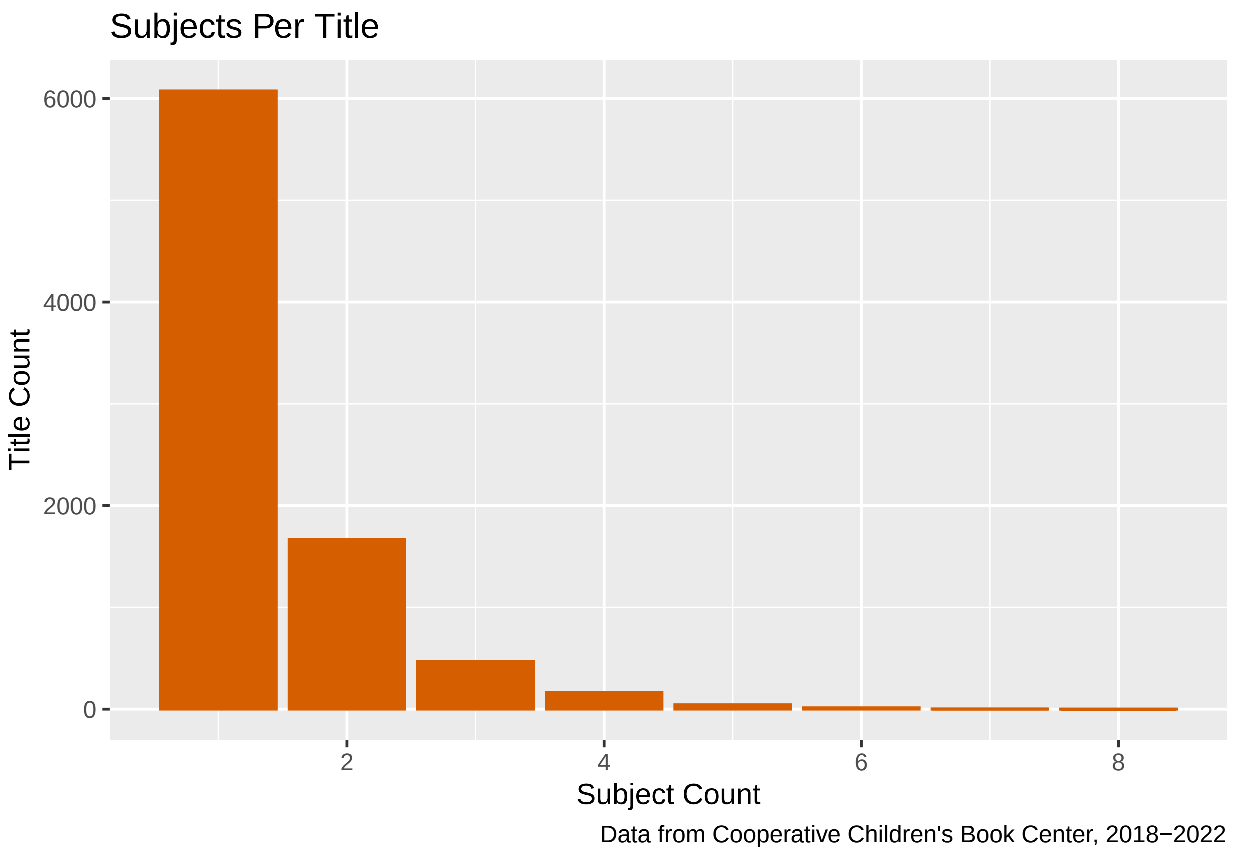
Data Source: Cooperative Children's Book Center

Here again we see a dip in titles published in 2020 which could possibly be attributed to the COVID-19 pandemic. Interestingly while the title count for the top 5 subjects has recently gone down, the title count of LGBTQIA+ Character/Topic has remained relatively steady. This indicates that LGBTQIA+ titles were not as affected by the change in 2020 and that the ratio of LGBTQIA+ titles published has increased when compared to the top 5 subjects.

We can also see in the top 5 subjects that the more generalized subjects of “Brown Skin Unspecified” and “Multicultural General” are becoming less used while the more specific subjects of “Black/African”, “Asian”, and “Latinx” have remained more steady. This could indicate a shift from general diversity to more focused writing on specific diverse aspects.

Number of subjects per title

```
TitleSubjectCount<- DiversityRpt %>%
  select(Title, Subject)

TitleSubjectCount$Count <- str_count(TitleSubjectCount$Subject, ', ')+1


TitleSubjectCount <- TitleSubjectCount %>%
  drop_na()


ggplot(TitleSubjectCount, aes(x=Count))+
  geom_bar(color = "#D55E00", fill = "#D55E00")+
  labs(title = "Subjects Per Title", x = "Subject Count", y = "Title Count",
       caption="Data from Cooperative Children's Book Center, 2018-2022")
```

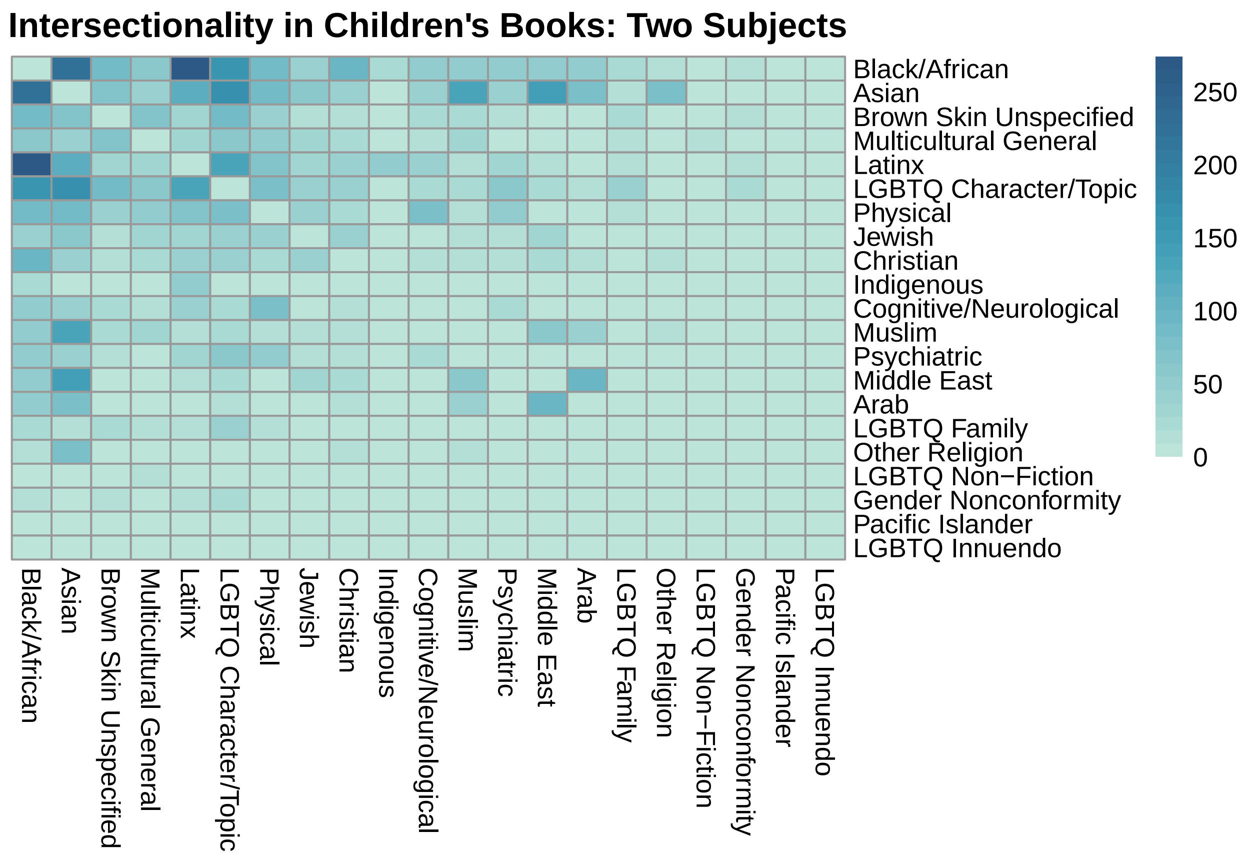


After noticing that some titles had more than one subject, we were interested in seeing what the most number of subjects in a single book was as well as the number of subjects per title. The maximum number of subjects for a single title is 8, but most titles feature only one subject.

Title intersectionality with two subjects

```
SubjectDF <- select(DRSimple, c('Title', 'Subject'))

SubjectDF <- SubjectDF %>%
  distinct() %>%
  drop_na()


SubjectDFMatrix <- SubjectDF %>%
  mutate(n = 1) %>%
  spread(Subject, n, fill=0) %>%
  select(-Title) %>%
  {crossprod(as.matrix(.))} %>%
  `diag<-`(0)

SubjectDFMatrix <- SubjectDFMatrix[DRSCount$Subject, DRSCount$Subject]
```

```r
pheatmap(SubjectDFMatrix, treeheight_row = 0, treeheight_col = 0,
         cluster_rows = FALSE, cluster_cols = FALSE,
         color = paletteer_c("ggthemes::Blue-Teal", 30),
         main = "Intersectionality in Children's Books: Two Subjects")
```



Because there are titles that feature more than one subject, we decided to look at the intersectionality of two subject pairs in diverse titles. In this plot, we are looking at all titles with two or more subjects and seeing how often a subject is written about with another subject. Interestingly, Black/African and Latinx are the most concurrent subjects with Black/African and Asian being behind that. This could be occurring because these are the top racial/country of origin subjects and they are written about with other topics besides those two pairs. This plot is most useful when looking at the least concurrent subjects, where we notice that LGBTQ Family, LGBTQ Non-Fiction, Gender Nonconformity, and LGBTQ Innuendo fall in the bottom 6 categories of concurrent subjects. This may indicate that there is a need for both these subjects on their own and titles featuring these subjects along with others.

Recommendations

3. How can we establish ourselves quickly as a diverse book publisher?

More than 850 publishers and imprints have published diverse books in the past 5 years. As a new imprint, it may be difficult to compete with the top publishers of diverse books by publishing upwards of 50 diverse titles in a single year. We would recommend that the new imprint focus on subject rather than quantity of books. As shown in our analysis, the number of books featuring LGBTQIA+ Character/Topic has remained steady over the past 5 years, even when the top subjects have trended downwards. Additionally, LGBTQ

Family, LGBTQ Non-Fiction, Gender Nonconformity, and LGBTQ Innuendo are some of the least published and least concurrent subjects.

We recommend that the new imprint utilize many of their resources on finding and publishing books featuring LGBTQ Characters and Authors. Particularly, looking at elements outside of just having an LGBTQ main character and looking at other aspects like family and gender nonconformity.

Additionally, because the top 5 subjects published feature race or country of origin, other subjects that we recommend publishing are the less published subjects of Indigenous, Arab, and Pacific Islander.

The analysis showed the majority of titles do not have more than one subject and those that do usually feature race/country of origin as a second subject. There is no need to publish titles with overly dense number of subjects, but it is recommended to add race and country of origin diversity as a second subject when able.

By publishing broader LGBTQ+ subjects and adding additional race and country of origin diversity when appropriate, as well as publishing Indigenous, Arab, and Pacific Islander subjects, the new imprint would meet important needs in the publishing industry. These subjects both follow the publishing trends of diverse books and provide unique subjects that are not currently being offered.

Focusing on publishing some configuration of this unique set of subjects, the new imprint will quickly stand out to readers and buyers when compared to other publishers in the industry.

Further Analysis

Analyzing the spread of diverse subjects published compared to the diversity in the country could indicate what subjects need more titles in order to mirror the diversity in the US. We also recommend looking at the downturn of publishing in 2020 and examining whether this was due to the COVID-19 pandemic and supply chain issues or if it was due to reader and buyer changes.