



# Accountability in Research

## Policies and Quality Assurance

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/gacr20>

## Text recycling in STEM: A text-analytic study of recently published research articles

Ian G. Anson & Cary Moskovitz

To cite this article: Ian G. Anson & Cary Moskovitz (2020): Text recycling in STEM: A text-analytic study of recently published research articles, *Accountability in Research*, DOI: [10.1080/08989621.2020.1850284](https://doi.org/10.1080/08989621.2020.1850284)

To link to this article: <https://doi.org/10.1080/08989621.2020.1850284>



Published online: 24 Nov 2020.



Submit your article to this journal [↗](#)



Article views: 25




View related articles [↗](#)



View Crossmark data [↗](#)



# Text recycling in STEM: A text-analytic study of recently published research articles

Ian G. Anson <sup>a</sup> and Cary Moskowitz<sup>b</sup>

<sup>a</sup>Department of Political Science, University of Maryland Baltimore County, Baltimore, MD, USA;

<sup>b</sup>Thompson Writing Program, Duke University, Durham, NC, USA

## ABSTRACT

Text recycling, sometimes called “self-plagiarism,” is the reuse of material from one’s own existing documents in a newly created work. Over the past decade, text recycling has become an increasingly debated practice in research ethics, especially in science and technology fields. Little is known, however, about researchers’ actual text recycling practices. We report here on a computational analysis of text recycling in published research articles in STEM disciplines. Using a tool we created in R, we analyze a corpus of 400 published articles from 80 federally funded research projects across eight disciplinary clusters. According to our analysis, STEM research groups frequently recycle some material from their previously published articles. On average, papers in our corpus contained about three recycled sentences per article, though a minority of research teams (around 15%) recycled substantially more content. These findings were generally consistent across STEM disciplines. We also find evidence that researchers superficially alter recycled prose much more often than recycling it verbatim. Based on our findings, which suggest that recycling some amount of material is normative in STEM research writing, researchers and editors would benefit from more appropriate and explicit guidance about what constitutes legitimate practice and how authors should report the presence of recycled material.

## ARTICLE HISTORY



Integra16 November 2020

## KEYWORDS

text recycling;  
self-plagiarism; plagiarism;  
research ethics; publication  
ethics

## Introduction

In STEM (Science, Technology, Engineering, and Mathematics) fields, successful researchers routinely write new papers that build directly on their prior work. Although each new paper is expected to be intellectually and substantively distinct from that prior work, STEM researchers frequently need to perform some of the same discursive tasks across multiple papers. For example, researchers who use the same experimental apparatus or data analysis method in successive papers will need to describe that apparatus or

**CONTACT** Cary Moskowitz  cmosk@duke.edu  Thompson Writing Program, Duke University, Durham, NC, USA

method again. Similarly, authors will often need to discuss some of the same research literature or background information. It should come as no surprise that scientists sometimes choose to reuse some material from their prior papers in their new ones, a practice known as text recycling (or, problematically, “self-plagiarism”).

Over the past decade, dozens of scientific journals have published editorials on text recycling, many of which state outright that recycling material from one’s previously published work is inherently unethical and unacceptable (Moskovitz, 2019). In contrast, organizations as diverse as the American Psychological Association (2020), John Wiley & Sons, (2014), and the Committee on Publication Ethics (2013) have issued written guidelines clearly stating that the practice is legitimate in some contexts.

Aside from any substantive policy differences, guidelines on text recycling are often written in ways that make it difficult for authors, editors, and other stakeholders to determine whether the guidelines pertain only to verbatim recycling (exact duplication of words) or also apply to recycled material that has been altered. For example, the Society for Industrial and Applied Mathematics’s “Authorial Integrity in Scientific Publication” (n.d.) says this [emphasis added]:

A related form of authorial misconduct is duplicate publication, meaning *unacceptably close replication* of the author’s own previously published text or results without acknowledgment of the source. This is sometimes called “self-plagiarism”.

Another passage in this same document uses different language [emphasis added]:

If a few *identical* sentences previously published by the current author appear in a subsequent work by the same author, this is unlikely to be regarded as duplicate publication. In contrast, it is unacceptable for an author to include *significant verbatim or near-verbatim portions* of his/her own work, or to depict his/her previously published results as new, without acknowledging the source.

The Council of Science Editors (2018) White Paper, in comparison, mentions only verbatim reuse, instructing authors to “avoid duplicate publication, which is reproducing *verbatim content* from their other publications” [emphasis added].

The proliferation of such policies in recent years demonstrates that text recycling has become an increasingly important issue in publishing and research ethics. Nevertheless, little is known about researchers’ actual recycling practices, such as how common the practice is and how often researchers reuse material in verbatim versus altered forms.

To help fill this gap, we report here on a detailed computational analysis of text recycling in published STEM research articles. Using a tool we created in the R programming environment (Anson, Moskovitz & Anson, 2019), we

analyze a corpus of 400 published research reports, performing pairwise comparisons of articles from 80 federally funded research projects across eight disciplinary clusters.<sup>1</sup> Our exploratory analysis revealed meaningful patterns in contemporary text recycling. STEM research groups frequently recycle some material from their previously published articles, regardless of discipline. We also distinguish between verbatim and partial or “patch-written” forms of text recycling. Our results show that many authors superficially manipulate their recycled prose, perhaps in an attempt to mask the recycled nature of the material; instances of verbatim recycling were rare in comparison. Based on our descriptive findings, text recycling guidelines should address both legitimate and inappropriate text recycling practice – including non-verbatim reuse.

## Background

### *What is text recycling?*

Scientists use non-original material in their research reports in different ways. A widely accepted type of reuse is the “commonplace” – a generic expression such as “It can be shown that ...” or “Data are reported at a significance level of  $p < 0.05$ .” A widely condemned use of non-original material is plagiarism, the unattributed reuse of material composed by others. Text recycling is distinct from both commonplaces and plagiarism in that the reused material is the author’s *own original* material. The matter of what constitutes one’s “own” work, however, is complicated in scientific fields, as scientific research papers tend to have multiple authors, and related papers produced by a given lab may have overlapping but not identical authors.

Adding to the complexities of authorship, the amount and distribution of text recycling in a scientific article can vary markedly. In one instance, authors might recycle only a single, verbatim passage of material; in another, they may copy and paste multiple sentences from one of their prior papers and then edit those sentences – deleting and/or adding words, phrases, or sentences to adapt the material to the new context, perhaps inserting the recycled material into newly composed paragraphs. Authors may also “rewrite” recycled material just to make it different – replacing some words with synonyms, reordering clauses and sentences, and so on – either because they were instructed to do so by editors or because they believe others will judge verbatim reuse as inappropriate. Recycled material may, therefore, be distributed in complex patterns. An example of such complexity is shown in [Figure 1](#), where recycled material in the first two pages of a published research report is highlighted in gray.

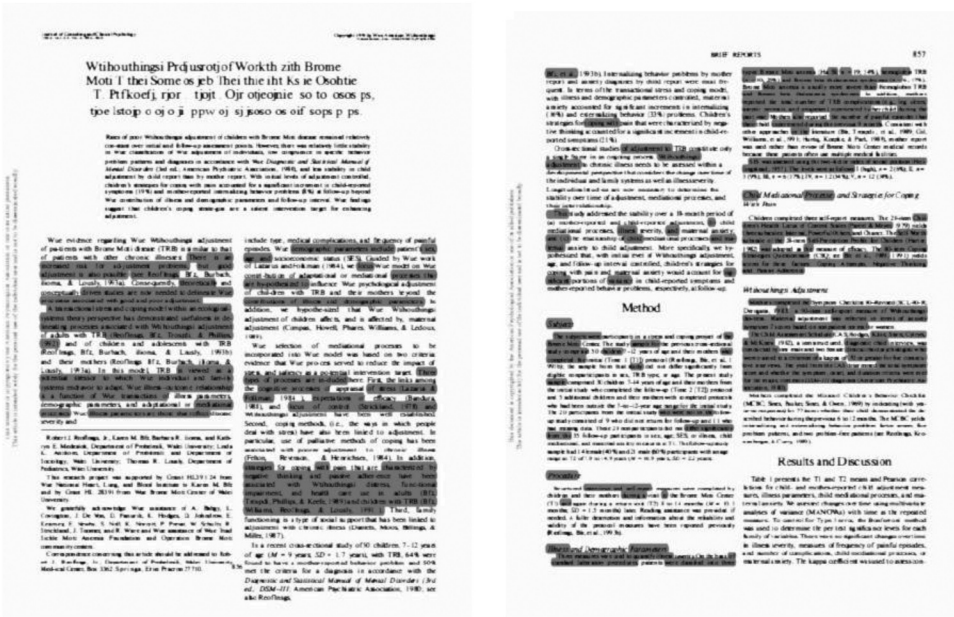


Figure 1. Example of distribution of recycled material in a scientific report. (Text has been altered for anonymity.)

Text recycling is clearly a complex discursive practice. Given the various ways in which recycled text can appear in a manuscript, examining researchers’ actual recycling practices computationally is challenging.

### Prior research on text recycling practices

Scholars from a variety of disciplines have studied the reuse of textual materials empirically. Some of this work has been performed to answer discipline-specific research questions. Clough et al. (2002), for example, developed an analytic tool to study journalists’ reuse of “copy” (material from newswire services); Lee (2007) developed a tool for identifying text reuse in ancient literary texts. A considerable body of empirical research has focused on the broader and more practical matter of identifying plagiarism in academic works. For a review of such approaches to text reuse detection, see Wilks (2004).

Few scholars, in contrast, have undertaken analytical studies of text recycling. Eaton and Crossman’s recent review of self-plagiarism in the social sciences (2018) includes no references to studies employing text analytic methods, and in a recent review of computational methods to detect academic plagiarism by Foltýnek, Meuschke, and Gipp (2019), text recycling again goes unmentioned.

The few analytic studies that have examined authors’ reuse of their own prior work have focused almost exclusively on occurrences deemed unethical or

otherwise inappropriate – which the authors typically label as “self-plagiarism.” Collberg and Kobourov (2005) conducted an informal analytic study to determine whether computer scientists ever published papers of “questionable originality.” Working with publications listed on websites of fifty computer science departments, they compared authors’ listed publications using their own Web spider and text-similarity analyzer, SPLaT. After culling “acceptable forms of republication,” they found an unspecified number of “questionable cases” involving substantial overlap combined with a lack of citation. Bretag and Carapiet (2007) used the plagiarism detection tool Turnitin to study “self-plagiarism” in a corpus of social science and humanities papers written by Australian academics; they included only papers which had overlapping, continuous passages of at least 10% and also excluded all papers that cited the source paper regardless of amount of overlap. Horbach and Halffman (2019), who analyzed published papers from four domains (biochemistry & molecular biology, economics, history, and psychology) by authors affiliated with Dutch universities, limited their study of to “unacceptable” text recycling – that is, cases which they believe would be considered as misconduct. Like Bretag and Carepiet, they excluded passages for which a citation was provided as well as papers with less than 10% identical overlap (among other factors).

To our knowledge, only a few published papers use text-analytic methods to study text recycling *not* restricted to “inappropriate” cases, and these tend to focus narrowly on specific academic disciplines. In an early study, Roig (2005) compared nine “target” articles from a single issue of a psychology journal with references in those papers by the article’s authors. Using a Microsoft Word macro routine, Roig compared every 6-word string in the target paper with each such string in the reference papers, highlighting matches. Five pairs of papers yielded “a substantial number” of string matches – nearly all of which were found in the papers’ methods sections. Sun and Yang (2015) also analyzed a corpus of discipline-specific journal articles – from the language learning and education disciplines published in 2009. Using the Turnitin tool, the authors identified 2298 “paraphrasing attempts”, of which 67% were recycled from the authors’ own work rather than from the work of others.

In a notable study, Citron and Ginsparg (2015) computed the amount of “reused” text in papers included in the arXiv.org preprint repository from 1991 to 2012. The arXiv repository consists of unpublished papers that generally belong to three fields: mathematics, physics, and earth and space sciences (Larivière et al. 2014). Citron and Ginsparg searched almost 800,000 of these preprint manuscripts for recycled “7-grams” (seven-word-long strings of text). They report that around 100,000 of these papers had more than 100 7-grams in common due to overlapping authorship. This means that of the papers analyzed, roughly one in eight contained more than 10% recycled content (based on an average article length of 7,000 words).

García-Romero and Estrada-Lorenzo (2014) performed a study that was broader in terms of discipline, using a corpus of biomedical papers from the Medline database included in the De'ja'vu database. Their analysis, however, focused on patterns of citation and bibliometric indicators (e.g., journal rankings) in 247 selected papers. The authors determined the amount of overlap between these articles through a manual qualitative content analysis. They found that for article pairs sharing at least one author, papers that did not cite the prior paper averaged 25% more text overlap than those that referenced the sources of reused material.

### ***Our study's contribution***

Our study differs from existing work in several respects. First, the aim of our work is to better understand the nature and scope of normative text recycling practices in high-impact scholarly writing, independent of *a priori* definitions of acceptability. This means that unlike earlier work, we study both verbatim and altered (or “patch-written”) forms of recycling, and we do so without considering whether or how the source document was cited in the new work. To accomplish this detailed view of text recycling in practice, we naively identify instances of text recycling using a specialized computer-assisted classifier which can identify both forms of text recycling and distinguish between them. Existing software such as Turnitin is less effective for our purposes than a tool designed expressly for this task.

Our study also focuses on journal articles in STEM disciplines, the setting in which text recycling has been most discussed and debated. We chose to analyze articles formally linked to NSF grants through the grant identifier that forms part of the publication metadata. Grant-funded, peer-reviewed papers are representative of the highest echelon of academic publishing, and selecting papers by grant numbers allowed us to build a corpus of successive papers produced in a line of research under a common principal investigator rather than by an identical group of authors. Compared to previous work such as that of Citron and Ginsparg (2015), this structured, intentional sampling design affords us a greater degree of internal validity in identifying author linkage (while substantially limiting the degree of power we might otherwise possess had we opted for a large-N data collection strategy).

## **Methods**

### ***Corpus construction***

In contrast to studying plagiarism, which ideally involves comparing texts against the set of all existing material, studying text recycling involves comparing texts only to other texts produced by the same author(s). In STEM fields, successive papers produced in a line of research often have

overlapping but not identical authors. Thus, while we usually think of “authors” as specific individuals, it is more useful in the context of STEM research writing to frame authorship in relation to research teams, or “labs.” (For a discussion of authorship in relationship to text recycling, see Moskovitz 2019). This can, however, complicate the process of case selection.

To identify sets of published papers that were produced by the same research teams on related research projects, we relied on U.S. National Science Foundation (NSF) grants as an identifying device. Each grant is assigned a number by NSF, which must be reported as a funding source in any eventual publication. Because research teams are awarded grants to study specific subjects, searching for shared grant numbers across papers allows us to find sets of papers written by the same research group on the same or closely related topics – a “most-likely” case selection design. NSF grants are also awarded across a wide range of STEM disciplines, which allowed us to investigate disciplinary trends in text recycling by stratifying our search for papers across different NSF research areas.

Because we wanted to study current text recycling practices, we selected publications from grants that ended in 2015. Earlier grants would have likely resulted in older papers, jeopardizing the relevance of our findings; later grants may not yet have yielded enough publications for comparison. To stratify our corpus across a broad range of STEM disciplines, we selected two program areas from each of four NSF directorates<sup>2</sup>: Biology; Engineering; Mathematical and Physical Sciences; and Social, Behavioral, and Economic Sciences.

### ***Article selection***

Within each of these program areas, we used the online NSF grant search tool to identify grants with the desired end date, limiting the search by award amount and grant type as described in the supplementary materials. Working from this output, we selected twenty grants within each disciplinary area that had yielded at least five published papers to date (since we needed enough publications from each grant for comparison). To find these articles, we searched the Web of Science (WOS) database by NSF grant number. When we were not able to locate a sufficient number of papers for any grant using WOS, we supplemented our search using the [redacted] University Library online search and Google Scholar. We performed this search method until we had accumulated 400 total papers across the four academic disciplines under study.

### ***Analytical method***

In order to assess the nature and extent of text recycling in our dataset, it was necessary to create a specialized tool. We constructed an algorithm in the R programming environment that used Levenshtein distance to calculate the similarity of texts at the sentence level (e.g., Yujian and Bo 2007).

Levenshtein distance measures the degree of similarity of words using a simple calculation, assigning lower scores to words with increasingly overlapping letters. This word-level scoring was computed at the sentence level, allowing us to measure sentence similarity through the creation of “Levenshtein distance matrices.” See [Figure 2](#) for an example.

For each of the 80 grants, we performed pairwise comparisons of all five research articles, for a total of ten article comparisons per grant (i.e., paper 1 vs. paper 2, paper 1 vs. paper 3, . . . paper 4 vs. paper 5). For each comparison, we calculated Levenshtein distance matrices for all sentence pairs (i.e., paper 1, sentence 1 vs. paper 2, sentence 1; paper 1, sentence 1 vs. paper 2, sentence 2; etc.). The diagonals of these matrices report the Levenshtein distance of each word in both sentences. In the clearest case of text recycling (each word is identical in both sentences), the diagonal would record a set of 0’s (zero non-overlapping characters in each word). In our example sentences in [Figure 2](#), the first and third words are identical (yielding zeros in matrix positions [1,1] and [3,3]), but the rest are not identical (non-zero values).

Because text recycling does not always begin and end where sentences begin and end, we sought to measure *partial* sentence-level text recycling by using the off-diagonal areas of the matrices as well. We instructed the algorithm to report three summary statistics from each Levenshtein matrix;

	["the", "lazy", "brown", "dog", "jumps", "under"]					
["the",	0	5	5	5	4	5
"quick",	4	5	5	5	5	5
"brown",	4	5	0	5	5	5
"fox",	4	5	5	2	4	5
"jump",	4	5	4	5	1	5
"over"]	4	5	4	3	4	3

**Figure 2.** Example of Levenshtein distance matrix for a pair of sentences: “The quick brown fox jump over” and “The lazy brown dog jumps under.” Note that “The” and “Brown” are identical words, yielding scores of 0 in position [1,1] and [3,3].

we “counted” the amount of recycled content in each sentence by combining three different metrics: verbatim sequences, exact word matches, and partial matches (as described below).

### *Verbatim sequences*

First, we instructed the algorithm to report the longest string of consecutive zeroes (pieces of sentences that matched exactly) in any diagonal of the matrix. For instance, if we count 10 consecutive zeroes in a matrix constructed from two 10-word sentences, those sentences are fully identical. If we counted seven consecutive zeroes in any diagonal for these sentences, the algorithm would report an exact match of seven words. Thus, regardless of *where* in a sentence verbatim recycling occurs, we obtain a count of that overlapping N-gram’s length. This yields a more robust account of verbatim reuse than Citron and Ginsparg (2015) count of 7-grams, as we count matching N-grams of any length.

### *Exact word matches*

Our second score is the overall number of identical Levenshtein pairs in the matrix. This score allows us to identify instances of “patch-writing,” in which authors obfuscate sequences of recycled words by making superficial alterations such as substituting synonyms or rearranging clauses. By counting identical words in any position of the sentence, we can “see through” such efforts to obscure text recycling. For example, a pair of sentences might differ because an author has shifted a three-word clause from the end of the original sentence to the beginning. This pair of sentences would still yield a high Exact Word Match score.

### *Partial matches*

Finally, we also capture “partial matches” – the number of words (excluding stopwords<sup>3</sup>) which had very similar content according to Levenshtein distances. We counted any word longer than three characters as a “partial match” if it had a Levenshtein score of 1. This means that we can identify patchwriting efforts employing change verb tense, plurality, and other superficial edits to the words of a manuscript.

### *A threshold for cases of text recycling*

Because we performed this protocol for all sentence pairs across 800 paper-to-paper comparisons, the output is too large to analyze meaningfully in its entirety. Instead, we instructed the algorithm to only report cases of “suspected” text recycling that matched specific criteria.

First, we combined our three text recycling measures together into a single additive score. Then, we engaged in a qualitative human coding exercise in order to evaluate the effectiveness of this combined score at identifying text

recycling. We performed this validity check using a hand-coded training set of texts ( $N = 303$ ) that allowed us to evaluate the performance of the algorithm. We used this exercise to adjust the weighting of each of the three measures and determine the scoring threshold that minimized the instance of text recycling false positives. For further details on this evaluation strategy, see Anson, Moskovitz & Anson, 2019..

We also tested our scoring method on simulated data. Our initial test case was Darwin's *Origin of Species* (1859), for which we compared a selection of the original text against a second copy in order to verify that scoring was being accurately registered. In a second test case, we interspersed passage from Darwin's *Descent of Man* with passages of varying length from Darwin's *Origin of the Species*. We examined how the various recycling patterns we had artificially created were scored and measured by the algorithm and then used these findings to calibrate the algorithm. In additional tests, we developed filters to exclude certain kinds of verbatim content, such as block quotes from earlier papers and long strings of parenthetical citations, using regular expressions. Our tests revealed a considerable number of false positives resulting from two sources: sentences with many short words (acronyms, chemical symbols, variables, and abbreviations), and tables that had been inadvertently treated as sentences. So prior to analysis, we instructed the algorithm to ignore "hits" with large numbers of two-letter words and a large number of fully capitalized words in the original text. A final round of hand-coded validation showed the number of false positives to be within an acceptable range: our scoring method achieved a precision score of 92.77, meaning that we were relatively successful at screening out instances of erroneous text recycling identification.

### **Procedure**

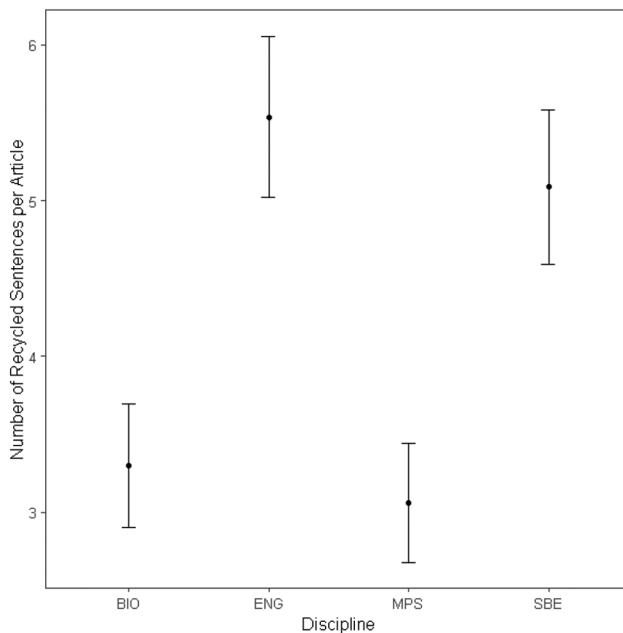
Texts were first ingested by the algorithm by grant cluster (five papers in each grant cluster). Once all five files in a grant were ingested, text was extracted from .pdf, .txt, and .html filetypes. After sentence breaks were identified using a specialized parser,<sup>4</sup> the sentences from the chronologically earliest document were compared with all four other documents. Sentences were preprocessed using regular expressions and lowercasing. We repeated this process to compare the second document with the remaining three documents, and so on, until all documents had been compared in a pairwise fashion (ten sets of comparisons in each grant cluster). Sentence scores were calculated using the Levenshtein distance approach described above, and "hits" that exceeded our validated text recycling threshold were recorded in a database. This database contained metadata about the grant, article, and sentence pair. This resulting data set of all "hits" was stored and used for subsequent analysis.

Based on our initial dataset of 400 articles from four fields of study, we computed 800 pairwise article comparisons and found a total of  $N = 1,359$  likely instances of text recycling at the sentence level. These data are the subject of our detailed examination of text recycling below.

## Results

Our results demonstrate that text recycling across the NSF disciplinary areas is widespread, though not necessarily extensive. First, we present an overall summary of the text recycling algorithm's findings across the four disciplinary clusters. [Figure 3](#), below, shows the average number of recycled sentences per article identified by our text matching strategy. We note that our key results are presented here in terms of instances of text recycling *per paper*. While we recognize that variation in typical paper length across disciplines might pose a threat to validity, robustness checks show that weighting each paper's results by paper length yields very similar conclusions. For the purposes of simplicity, we present the unweighted findings below.

[Figure 3](#) shows that on average, all four disciplinary areas exceeded 3 recycled sentences per paper, corresponding to around 1.2% of the average 250-sentence paper. Engineering (ENG) and Social, Behavioral, and Economic Sciences (SBE) exhibited more text recycling per article, with roughly 5.2 and 5.5 sentences recycled in an average paper (roughly 2% of



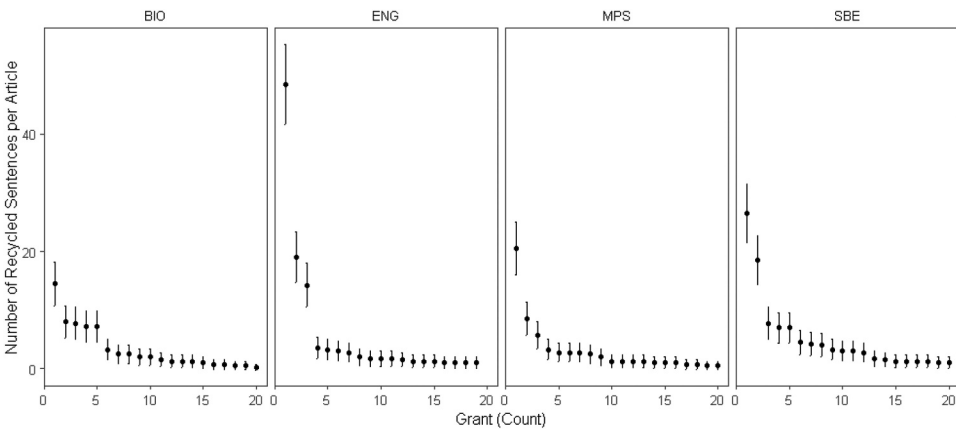
**Figure 3.** Average Number of Recycled Sentences per Article, Four Disciplinary Areas. (Error bars denote 95% Confidence Intervals based on Poisson Distribution.).

all content). However, as seen in [Figure 4](#), these averages are highly influenced by extreme outliers.

[Figure 4](#) plots the average number of recycled sentences per article across the 20 grants in each of the four disciplinary clusters. For each disciplinary cluster, outliers clearly drive the findings seen above in [Figure 4](#). For example, the grant in our dataset containing the most instances of recycling is captured in the leftmost bar of the ENG panel, revealing that this group of authors recycled roughly 50 sentences per paper across their publications. Assuming an average text length of around 250 sentences, this means that approximately 20% of this research group’s articles were duplicated from one paper to the next. Clearly, these findings stand in contrast to most other grants in our corpus, in which articles showed much less recycling. In fact, while nearly all grants had at least one paper containing some text recycling, 19.5% of papers in our corpus appeared to be free of text recycling. These findings suggest that a small proportion of articles contain substantial recycled content, whereas the plurality of others exhibit a modest – but nonzero – amount of text recycling.

### **Altered vs. verbatim text recycling**

Next, we consider the extent to which recycled material identified in these papers is verbatim and/or “altered.” We differentiate between these two forms of text recycling using a fairly blunt heuristic: We separate the identified instances of text recycling in our dataset according to our “Verbatim Sequences” measure described earlier. If we identified a Verbatim Sequence that extended across 80% or more of the total length of a given sentence, we label that instance of text recycling as verbatim. If the algorithm identified a text recycling instance in which there was no Verbatim Sequence exceeding



**Figure 4.** Number of “hits” per grant by disciplinary area

80% of the total sentence length, we assume that instance has been subject to some authorial manipulation, via patch-writing or other strategies.<sup>5</sup> While we cannot confirm the authors' intent in altering the text, we still arrive at a useful comparison of different “forms” of text recycling as they appear in the manuscripts under study.

Figure 5, above, shows the average count of altered and verbatim text recycling sentences in our dataset, across the four disciplinary areas under study. The average lengths of papers (in number of sentences) in our corpus by discipline are as follows: BIO, 214; ENG, 204; MPS, 230; SBE, 319. We see that while the volume of overall text recycling differs for each field (much like in Figure 3), the four fields exhibit a similar overall pattern: in each case, we see that altered recycling is more prevalent than verbatim recycling. In fact, we see that verbatim text recycling in each area is only found in around 1 or 2 sentences per article on average. This finding suggests that STEM researchers frequently see it necessary to make changes to recycled material – whether to adapt that material to the new context or to obfuscate their reuse.

Next, we consider the proportion of text recycling instances in each grant cluster that fit our definition of verbatim text recycling. As shown in Figure 6, of all grants which included some text recycling, about one-fourth contained

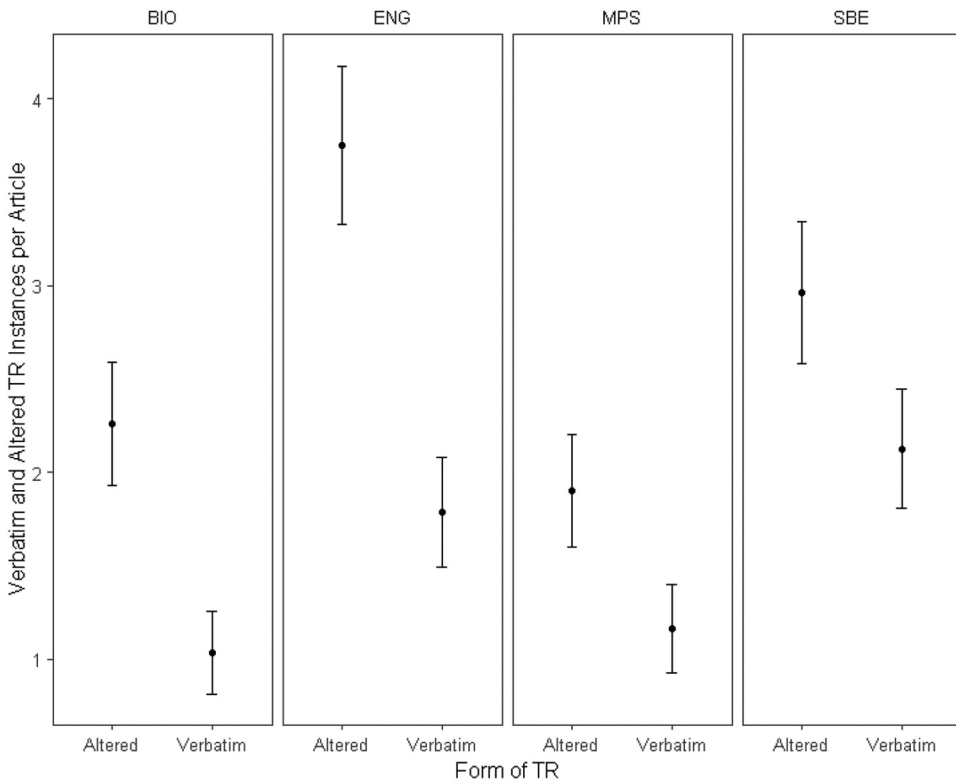
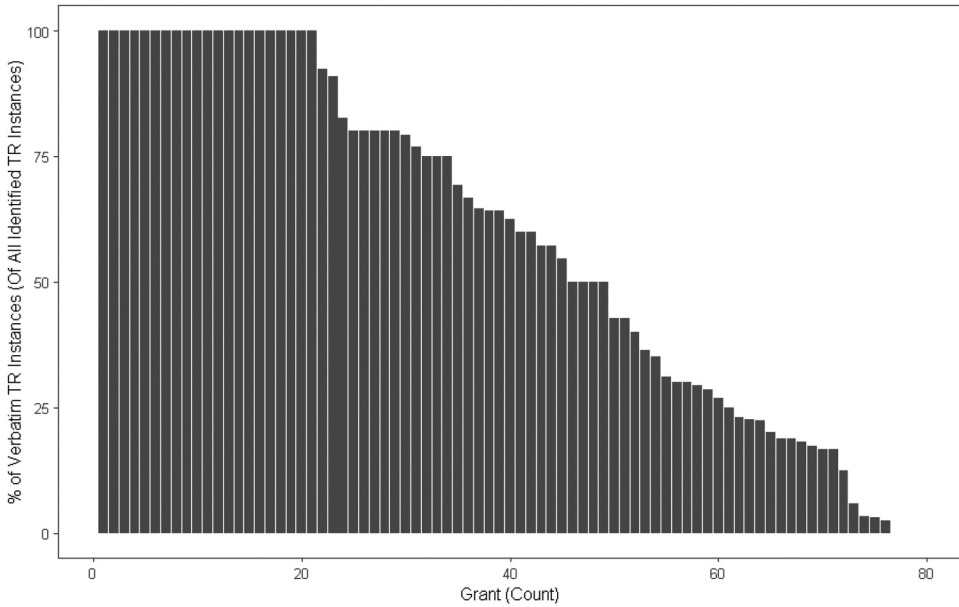


Figure 5. Average amount of verbatim and altered text recycling per article by disciplinary area.



**Figure 6.** Proportion of sentences recycled verbatim among all recycled sentences.

verbatim reuse exclusively. For the other grants, the proportion of verbatim text recycling declines in a fairly linear fashion, showing that verbatim reuse is practiced at varying levels by other research teams.

Taken together, [Figures 5 and 6](#) depict a situation in which a minority of practitioners engage in a small amount of verbatim duplication, while a larger group of practitioners mixes verbatim and altered reuse.

### ***Categorizing text recycling***

To better understand these findings, we conducted an informal qualitative examination and coding of a sample of sentences identified by our code as recycled. Sun and Yang (2015) identified a wide variety of ways in which authors reuse material from sources including their own: verbatim copying, reordering words/phrases, using synonyms, and so on. We identified four general categories in our sample hits: verbatim recycling, altered recycling, borderline cases, and false positives. [Table 1](#) shows examples of verbatim recycling. Note that the passages are not only identical, but idiosyncratic and thus likely show reuse of the researchers' own language rather than phrasings that are widely used (commonplaces). (Regarding this table and the others below, readers are reminded that sentences were stripped of punctuation prior to analysis.)

Examples of what we are calling altered recycling are shown in [Table 2](#). While researchers may have many reasons for altering recycled text, two are likely. One involves making changes needed to adapt the recycled material to

**Table 1.** Examples of verbatim recycling.

Paper A	Paper B
Although adolescents on average reported fairly high scores on the four indicators of family belonging moderate variation existed in this measure	Although adolescents on average reported fairly high scores on the four indicators of family belonging moderate variation existed in this measure
In the deadband mode no reheating is performed ie TSA TCA and supply air flow rate is set to the minimum allowed value	In the deadband mode no reheating is performed ie TSA TCA and supply air flow rate is set to the minimum allowed value
The entire element is taken to vanish when three of the eight integration points in the element have reached this stage	The entire element is taken to vanish when three of the eight integration points in the element have reached this stage

**Table 2.** Examples of altered recycling. Bold, underlined, and italicized text shows differences between source and destination papers.

Paper A	Paper B
Mother child and <b>stepfather</b> child relationship quality <u>was</u> defined in this study as adolescents perceptions of ...	Mother child and <b>father</b> child relationship quality <u>are</u> defined in this study as adolescents perceptions of ...
In the showup conditions participants viewed single photograph and were <b>instructed</b> to indicate <b>if</b> the perpetrator was present or absent	In the showup condition participants viewed single photograph and were <b>asked</b> to indicate <b>whether</b> the perpetrator was present or absent
The experiments also employed <b>platinum wire counter electrode</b> and <i>nonaqueous AgAgCl pseudoreference electrode that was separated from the solution by frit</i>	The experiments also employed <i>nonaqueous AgAgCl pseudoreference electrode which was separated from the solution by frit</i> and <b>platinumwire counter electrode</b>

the context of the new work. A clear instance of such adaption is shown in the first row of [Table 2](#), where *stepfather* is replaced by *father* – reflecting the different population being studied. In contrast, the second row includes synonym substitutions that seem rhetorically and stylistically inconsequential and thus obfuscatory: “instructed” for “asked” and “if” for “whether.” The third row also shows unproductive alterations, although the changes here are structural: the phrases “platinum wire counter electrode” (bolded) and “nonaqueous AgAgCl pseudoreference electrode which was separated from the solution by frit” (underlined) have swapped places.

In [Table 3](#) we see clusters of words and phrases interspersed with new material. These cases were scored by our code at levels just high enough to meet our threshold for text recycling. One might reasonably argue whether the overlap should be considered recycling. Even so, the presence of seemingly idiosyncratic, identifiable strings suggests that the similarity is not happenstance.

Finally, [Table 4](#) shows examples of false positives from our sample. The examples in Rows 1 and 2 might be categorized as “boilerplate” – specific phrasings that are routinely and openly reused by different authors within an organization. In some cases, such as in row 1, institutions encourage their researchers to use such phrasing exactly for regulatory compliance. Row 3, in

**Table 3.** Examples of Borderline reuse.

Paper A	Paper B
The NMR spectrum <b>of</b> in CD <b>consists of very broad resonance at ppm assignable to the tertbutyl groups of the ketimide ligand while</b> broad resonance at ppm is assignable to the methyl <b>groups of the acac ligand</b>	Its NMR spectrum in CD <b>Cl reveals the presence of single</b> broad resonance at ppm assignable to the methyl <b>protons of the TEMPO moiety</b>
<b>Here we describe how</b> DCPM <b>can be applied</b> to measure colloidal interactions <u>with</u> the surface of live cells	<b>We have expanded this capability to simultaneously track particles and cells which we implemented in conjunction with new analytical and interpretative methods to demonstrate proofofprinciple capabilities of</b> DCPM to measure colloidal interactions <u>at</u> the surface of live cells

**Table 4.** Examples of false positives. Italics show differences.

Paper A	Paper B
All procedures were performed in accordance with the guidelines of the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Universitys Animal Care Committee	All procedures were performed in accordance with the guidelines of the National Institutes of Health Guide for the Care and Use of Laboratory Animals and approved by the Universitys Animal Care Committee
An electronic companion to this paper is available as part of the online version <i>that can be found at <a href="http://journalinformatics.org/proxylibdukeedu">http://journalinformatics.org/proxylibdukeedu</a></i>	An electronic companion to this paper is available as part of the online version <i>at <a href="http://dxdoio.org">http://dxdoio.org</a></i>
<i>Verapamil hydrochloride and acetic acid</i> were purchased from Sigma Aldrich St Louis MO USA and stock solutions were prepared in <i>HPLCgrade water</i>	<i>Verapamil bradykinin and reserpine</i> were purchased from Sigma Aldrich St Louis MO and stock solutions were prepared in <i>methanol</i>

contrast, might better be described as “commonplace” – a phrasing widely used by members of the discourse community. Frequent use of both types of language reuse in STEM writing makes such false positives nearly unavoidable computationally.

### Authorship

Our corpus provided the data to investigate one other empirical dimension of text recycling: author overlap. Fundamental to the definition of text recycling is that authors are reusing “their own” material rather than that of others. In the humanities, scholars usually publish as solo authors or perhaps in duos or other small ad-hoc groups. But in contemporary STEM authorship, solo authorship is rare; instead, most journal articles have multiple authors – frequently in the double digits (and, in some disciplines, more than that). How to defines “one’s own” prior work is therefore both critical and problematic. To get a sense of author overlap in our corpus, we selected a sample of 80 paper pairs, the first and second (chronologically) papers for each grant. Within this sample, only five pairs (6.25%) had identical authors; 35% had more than two authors on the second paper who were not authors on the

first, and 10% had more than five authors on the second paper not on the first. Even with this small sample, it seems clear that text recycling in STEM usually involves papers with non-identical authors.

## Discussion

Unlike writing practices that are considered inherently unethical such as data fabrication and plagiarism, text recycling may be appropriate or inappropriate depending on details of the situation. Understanding the ways in which researchers recycle textual materials across disciplines can illuminate normative practices and allow us to determine how these practices align with existing policies and expectations.

Our findings support several basic facts about text recycling practices in STEM research. First, reusing some amount of material from one's prior published articles appears to be fairly common. That said, most STEM researchers tend to recycle very limited quantities of material.

Second, verbatim reuse of entire sentences is not ubiquitous. Instead, we find that STEM researchers frequently alter the material in some way when importing it into the new work. In some cases, these alterations are necessary for adapting the material to the context of the new research. In other cases, the alterations seem intended to disguise the act of recycling. While our study does not provide data regarding author's motivations for disguising their recycling, two seem probable based on recent studies. First, from their survey of STEM researchers, Hall et al. (2020) found that 42% of experts and 56% of novice researchers (graduate students and post docs) believed recycling text from their prior published papers to be inappropriate. Faced with the need to repeat some content included in their previously published papers but believing recycling to be unethical, some authors may "massage" those passages in the new paper to avoid obvious recycling. In fact, some plagiarism detection companies such as iThenticate encourage authors to use their services for exactly this purpose. (It is worth noting that such rewording will likely make it more difficult for editors to detect the presence of recycled material.) Second, Pemberton et al. (2019) reported that some interviewed editors instruct authors to "rewrite" or "reword" all recycled passages prior to publication, often as a result of concern about possible copyright infringement.<sup>6</sup>

Third, our findings appear to hold across STEM disciplines, including the quantitative social sciences. While we saw some variation in the amount of recycled material across the field clusters, the number of grants having non-negligible amounts of recycling was surprisingly similar. These findings align with recent survey studies reporting that opinions on the acceptability of text recycling were largely independent of discipline (Hall, Moskovitz, and Pemberton 2018; Moskovitz and Hall 2020). Thus, norms for text recycling

across fields appear to be sufficiently consistent to allow for common rather than discipline-specific guidelines within STEM.

### ***Policy implications***

The results of our study suggest that the wording of extant policies on text recycling do not adequately account for actual practice. We find that STEM researchers frequently recycle material from their prior articles, but also that they often alter that material in some way. Because altered recycling is as least as common as verbatim recycling, policies should explicitly address both.

Our findings also highlight the lack of policy coordination when it comes to verbatim vs. non-verbatim reuse. Regardless of motivation, our findings may suggest that STEM authors often rewrite recycled material merely to make it *look* different, raising important questions about science ethics and communication. Is disguising recycled material more or less ethical than the transparency of verbatim recycling? Is it better or worse for readers following that line of research? Current policies appear to come to very different conclusions in response to these foundational questions. Given that both verbatim and non-verbatim forms of recycling are practiced by authors across scientific fields, journal editors and publishers should attempt to arrive at a reasonable consensus about what text recycling practices are desirable, which discouraged, and which forbidden.

Finally, our dataset also reveals that authorship itself is more complex than what is accounted for by current policy language. While text recycling is commonly visible across paper pairs in our dataset, very few of these pairs were written by identical authors. This is not surprising: authorship routinely varies on STEM publications as researchers join or leave projects or when specialists are recruited as coauthors for their particular expertise. Text recycling policies need to address this real authorship situation, rather than just referring vaguely to “the author’s own writing.”

While prior research has shown that text recycling practices are common, the present findings have shed light on *how* text recycling is practiced in STEM disciplines. Together, those results show that existing policies do not address text recycling as typically practiced, and thus are in need of revision. Stakeholder organizations should commit to developing new guidelines addressing both verbatim and non-verbatim recycling and specifics of authorship. We note that accomplishing this task will also require stakeholders to agree on some basic terminology, since key terms such as *text recycling*, *self-plagiarism*, and *duplicate publication* are routinely used by different organizations with substantively different meanings.

## **Limitations**

Given the design of our corpus, our findings are subject to some scope conditions. First, our data represent the practices of researchers who were successful in obtaining N.S.F. grants: P.I.s who were U.S. residents and who demonstrated sophisticated scientific capability. While our sample is perhaps not representative of all STEM researchers, it does provide insights into the practices of successful STEM researchers – challenging assumptions that recycling is typically practiced by authors insufficiently expert to know the norms of their fields. Our study is also limited to a single genre – the journal article; recycling practices for other genres such as review articles or commentaries may be different. We also excluded disciplines that tend to produce articles consisting largely of non-prose material; thus, our results may not hold for mathematics, computer science, and related disciplines. And since our aim was to study the practice of textual reuse itself rather than in relation to considerations of appropriate or inappropriate practice, we did not consider the presence or absence of citations to the prior work. Practices of STEM researchers in other countries may vary as a result of different cultural norms or less fluency in English-language writing, which is the standard in STEM fields. Our study did not investigate language proficiency nor any of a number of other possible causative factors.

We also note that this study only investigated recycling of prose. STEM researchers may also recycle visual materials (such as diagrams and photographs) as well as equations. We make no claims about the extent or patterns of such recycling in general or by discipline.

Finally, we note that our analysis was conservative, likely underestimating the frequency of text recycling. Our methodology was capable only of detecting recycled prose, not equations or visuals which may also have been recycled. Also, our code likely missed words that were hyphenated for column formatting; such hyphenations are common in scientific research articles.

## **Recommendations for future research**

Our study and its limitations suggest a number of areas for further investigation:

- The aim of the present study was to learn about current text recycling practices. Given the considerable rise in the use of text similarity tools for plagiarism detection in recent years, it would be interesting to know whether and how text recycling has changed over recent decades and the role that these tools may have had in driving these changes.
- Our study analyzed text recycling only at the level of the individual sentence. Policies and editorials disallowing text recycling often instruct authors to instead place recycled material in quotation

marks. Analysis of structural patterns of recycling (i.e., do researchers tend to recycle as a single, contiguous block of text or in multiple, dispersed text strings) would help determine the practicality of these expectations.

- We did not collect data on the structural location (Introduction, Methods, etc.) of occurrences. Given that guidelines often allow for some (or a greater quantity) of recycling in some sections or for some rhetorical purposes, future studies might investigate these parameters.
- Policies for text recycling often require a citation to the prior work and sometimes also ask authors to make more explicit announcements of the presence of recycled material for readers. Future studies might examine whether and how authors attribute the recycled material to its source and the relation between such attributions and the nature of the recycled material.
- One of the potential limitations of the present study is that text recycling as defined can be more difficult to accurately identify than simply overlapping or identical text. This is because some recycling could take the form of common expressions, useful linguistic conventions, or “turns of phrase” that the author has (unintentionally) internalized over their career. We certainly recognize the potential for the “thought recycling” that might inadvertently occur when a researcher has written on the same topic for many years! Future studies are poised to develop new tools for scoring and identifying common linguistic conventions and instances of intentional text recycling, despite the inability of our scoring system to do so. Perhaps future studies could work to establish a “baseline” of unintentionally recycled material that occurs naturally in language by studying a kind of “placebo group” of papers with no shared authorship and no plagiarized content. While outside the scope of our study, these techniques would more precisely identify text recycling in practice.<sup>7</sup>

Overall, our findings suggest the need for clear and consistent guidelines on text recycling. Since we have empirical evidence that this practice is widespread at the highest levels of scientific communication, the call for a clear ethical description of the practice, within and across disciplines, is increasingly urgent.

## Notes

1. Discerning readers might wonder why such a specialized tool was necessary when proprietary software (such as Turnitin) is available that can perform text matching. While we discuss these reasons in greater detail in the sections that follow, the most

important reason for developing our own algorithm is that it allows us to examine various forms of text recycling (rather than the mere presence or absence of recycled content). While Turnitin's core algorithm is likely based on similar methods to our own (string pattern matching and scoring), the source code is unavailable for public use. Our algorithm allows us to explore authorial practice in a more robust and fine-grained mode.

2. Funding for science and engineering research at the NSF is done through seven directorates: Biological Sciences, Computer and Information Science and Engineering, Engineering, Geosciences, Mathematical and Physical Sciences, Social, Behavioral and Economic Sciences, and Education and Human Resources. To keep the scope of our study manageable, we eliminated some directorates from consideration: We eliminated Education and Human Resources because we were focused on scientific writing. We also excluded fields that often produce papers with little prose. While text recycling practices in these fields is certainly of interest, the tools we needed to develop to investigate recycling of prose would likely not be appropriate for analyzing papers consisting largely of equations or code. We thus excluded the NSF directorate "Computer and Information Science and Engineering." We also excluded Geosciences because of its interdisciplinarity.
3. Common English "stopwords" were excluded from this part of the analysis using the Snowball stopword dictionary. See <https://cran.r-project.org/web/packages/stopwords/stopwords.pdf> for more information.
4. Sentence parsing was performed using the sentence TokenParse command in the lexRankr package for R.
5. We use 80% rather than 100% for two reasons: first, there are sometimes trivial editorial or layout edits that result in very minor alterations; second, a sentence may be recycled verbatim from the source but then have additional material added to it.
6. Pemberton et al. (2019) also report that these editors tended to have little understanding of copyright law. A recent legal analysis of text recycling in STEM research conducted by members of our research group (not yet published) suggests that typical recycling practices in STEM research articles are, in fact, legal – at least under U.S. law.
7. We thank an anonymous reviewer for this interesting idea.

## Acknowledgments

We thank our colleague Chris Anson for his valuable conversations about this work as it developed and undergraduate students Dennis Nguyen, Juliana Hoover, and Evelyn Scarrow for data mining and cleaning. A special thank you to Brooke Harmon, who's exceptional thoughtfulness and diligence in collecting, cleaning, and double-checking these data was essential to this study. We also thank audience members at these meetings for their feedback: the 8th International Conference on Writing Analytics, September 2019, Winterthur, Switzerland; the 7th International Conference on Writing Analytics, 2019, St. Petersburg, FL, and the Annual International Conference of The Association for Practical and Professional Ethics, 2020. This research is supported by NSF grant SES-1737093.

## Data availability statement

Data are available through Duke University on an individual basis. Please contact Cary Moskovitz at [cmosk@duke.edu](mailto:cmosk@duke.edu) for details.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

This work was supported by the National Science Foundation [SES-1737093].

## ORCID

Ian G. Anson  <http://orcid.org/0000-0002-1545-4270>

## References

- American Psychological Association. 2020. *Publication Manual of the American Psychological Association*. 7th ed. Washington, D.C.: American Psychological Association.
- Anson, I. G., C. Moskovitz, and C. M. Anson. 2019. "A Text-Analytic Method for Identifying Text Recycling in STEM Research Reports." *Writing Analytics* 3: 125–150.
- Bretag, T., and S. Carapiet. 2007. "A Preliminary Study to Identify the Extent of Self-plagiarism in Australian Academic Research." *Plagiary* 2 (5): 1–12.
- Citron, D. T., and P. Ginsparg. 2015. "Patterns of Text Reuse in a Scientific Corpus." *Proceedings of the National Academy of Sciences* 112 (1): 25–30. doi:10.1073/pnas.1415135111.
- Clough, P., R. Gaizauskas, S. S. Piao, and Y. Wilks. 2002. "Meter: Measuring Text Reuse." In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 152–159. Association for Computational Linguistics.
- Collberg, C., and S. Kobourov. 2005. "Self-plagiarism in Computer Science." *Communications of the ACM* 48 (4): 88–94. doi:10.1145/1053291.1053293.
- Committee on Publication Ethics. 2013. *Text Recycling Guidelines*. <https://publicationethics.org/text-recycling-guidelines>
- Council of Science Editors. 2018. *CSE's White Paper on Promoting Integrity in Scientific Journal Publications*. Editorial Policy Committee. [https://www.councilscienceeditors.org/wp-content/uploads/CSE-White-Paper\\_2018-update-050618.pdf](https://www.councilscienceeditors.org/wp-content/uploads/CSE-White-Paper_2018-update-050618.pdf)
- Eaton, S. E., and K. Crossman. 2018. "Self-plagiarism Research Literature in the Social Sciences: A Scoping Review." *Interchange* 49 (3): 285–311. doi:10.1007/s10780-018-9333-6.
- Foltýnek, T., N. Meuschke, and B. Gipp. 2019. "Academic Plagiarism Detection: A Systematic Literature Review." *ACM Computing Surveys (CSUR)* 52 (6): 1–42. doi:10.1145/3345317.
- García-Romero, A., and J. M. Estrada-Lorenzo. 2014. "A Bibliometric Analysis of Plagiarism and Self-plagiarism through Déjà Vu." *Scientometrics* 101 (1): 381–396. doi:10.1007/s11192-014-1387-3.
- Hall, S., C. Moskovitz, and M. A. Pemberton. 2018. "Attitudes toward Text Recycling in Academic Writing across Disciplines." *Accountability in Research* 25 (3): 142–169. doi:10.1080/08989621.2018.1434622.
- Horbach, S. P. J. M. S., and W. W. Halfman. 2019. "The Extent and Causes of Academic Text Recycling or 'Self-plagiarism'." *Research Policy* 48 (2): 492–502. doi:10.1016/j.respol.2017.09.004.

- John Wiley & Sons. 2014. *Best Practice Guidelines on Publishing Ethics: A Publisher's Perspective*. Hoboken: John Wiley & Sons. <https://authorservices.wiley.com/asset/Best-Practice-Guidelines-on-Publishing-Ethics-2ed.pdf>
- Larivière, V., C. R. Sugimoto, B. Macaluso, S. Milojević, B. Cronin, and M. Thelwall. 2014. "arXiv E-prints and the Journal of Record: An Analysis of Roles and Relationships." *Journal of the Association for Information Science and Technology* 65 (6): 1157–1169. doi:10.1002/asi.23044.
- Lee, J. 2007. "A Computational Model of Text Reuse in Ancient Literary Texts." *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*: 472–479.
- Moskovitz, C. 2019. "Text Recycling in Scientific Writing." *Science and Engineering Ethics* 25 (3): 813–851. doi:10.1007/s11948-017-0008-y.
- Moskovitz, C., and S. Hall. 2020. "Text Recycling in STEM Research: An Exploratory Investigation of Expert and Novice Beliefs and Attitudes." *Journal of Technical Writing and Communication* 004728162091543. doi:10.1177/0047281620915434.
- Pemberton, M., S. Hall, C. Moskovitz, and C. M. Anson. 2019. "Text Recycling: Views of North American Journal Editors from an Interview-based Study." *Learned Publishing* 32 (4): 355–366. doi:10.1002/leap.1259.
- Roig, M. 2005. "Re-using Text from One's Own Previously Published Papers: An Exploratory Study of Potential Self-plagiarism." *Psychological Reports* 97 (1): 43–49.
- Society for Industrial and Applied Mathematics. Authorial Integrity in Scientific Publication. n.d. Accessed 9 September 2020. <https://www.siam.org/publications/books/about-siam-books/for-authors/detail/authorial-integrity-in-scientific-publication>
- Sun, Y. C., and F. Y. Yang. 2015. "Uncovering Published Authors' Text-borrowing Practices: Paraphrasing Strategies, Sources, and Self-plagiarism." *Journal of English for Academic Purposes* 20: 224–236. doi:10.1016/j.jeap.2015.05.003.
- Wilks, Y. 2004. "On the Ownership of Text." *Computers and the Humanities* 38 (2): 115–127. doi:10.1023/B:CHUM.0000031184.28781.47.
- Yujian, L., and L. Bo. 2007. "A Normalized Levenshtein Distance Metric." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29 (6): 1091–1095. doi:10.1109/TPAMI.2007.1078.