Duplicate references in the 'Introduction' and 'Discussion' sections of scientific articles on physical education and sports

Leonid Podrigalo¹ABCDE, Sergii Iermakov²ABCDE
¹ Kharkiv State Academy of Physical Culture, Ukraine
² Kharkiv State Academy of Design and Arts, Ukraine

Authors’ Contribution: A – Study design; B – Data collection; C – Statistical analysis; D – Manuscript Preparation; E – Funds Collection

Abstract

Background and Study Aim

In the context of the increasing volume of scientific publications, the adequacy and effectiveness of source use in key sections of articles become crucial tasks. This requires detailed consideration, especially in light of potential duplication of references, which can affect the perception of originality and depth of research. The purpose of this article is to investigate the nature and impact of duplicate references in the ‘Introduction’ and ‘Discussion’ sections of scientific articles on physical education and sports.

Material and Methods

The authoritative Web of Science Core Collection database was selected as the information source. The search was conducted under the categories “Hospitality, Leisure, Sport & Tourism and Sport Sciences”. Publications available in full version and open access were extracted (n=21). The selection criteria included indicators such as a high level of citation, including the “Citing Items Breakdown / Background” indicator (WoS). In this study, the PyCharm CE development environment and Python libraries were used to extract and analyze duplicate references in the «Introduction» and «Discussion» sections of scientific articles. The Shapiro-Wilk test was utilized to assess the normality of data distribution, and the Spearman correlation coefficient was applied to analyze the relationship between citation frequency and the frequency of duplicate references. To mitigate the impact of outliers, the robust RANSAC method was also employed.

Results

Key trends in the use of sources and citation strategies in highly cited articles on physical education and sports were identified. Articles categorized as ‘Highly cited threshold ≤1%’ and ‘≥1%’ exhibit different aspects of their impact in the academic environment. A statistically significant relationship was established between the number of citations of articles and the frequency of duplicate references within them, with a Spearman correlation coefficient of 0.645 (p= 0.0016).

Conclusions

The duplication of key sources in the «Introduction» and «Discussion» sections can be an indicator of the depth of research and the quality of the scientific argument. The results highlight the importance of conscious and strategic use of literature in scientific articles, which can enhance their impact and visibility in the academic community.

Keywords: scholarly publications, academic citations, literature analysis, scholarly impact, research methodology

Introduction

In the context of the increasing volume of scientific publications, the adequacy and effectiveness of source usage in key sections of articles become an important task. This requires detailed consideration, especially in light of potential duplication of references, which can affect the perception of originality and depth of research. In this context, the repeated use of cited sources in key sections affects the structure and content of scientific discourse, as well as the assessment of their contribution to ensuring consistency and depth of the research argument.

Research on various aspects of preparing the “Discussion” section presents specific recommendations on structure, language, and sequence of material presentation [1, 2, 3, 4, 5, 6]. Many studies support the idea of a logical interconnection between sections of a scientific article. For instance, Jawaid et al [7] assert that the “Introduction” and “Discussion” sections are the most important parts of an article. Therefore, most references will be found in these sections.

In the study by Vieira et al. [8], it is recommended to start the discussion by recalling important information included in the “Introduction” and/or “Material and Methods” sections. This approach allows for contextualizing the research. Day et al. [9] suggest considering “Introduction” and “Discussion” in close interaction as pairs of important elements of an article. The recommendations presented above accurately reflect the structure and content of the considered sections of a scientific article. The results of these studies indicate a close interrelation between the “Introduction” and “Discussion” sections.
Goulston provides a detailed explanation of the interrelation between the content of the “Introduction” and “Discussion” sections [10]. The author notes that if theories or studies mentioned in the introduction can explain the conclusions, then these sources should be used as a basis for discussing the results. This is corroborated by the research of Bavdekar [11], who points out that the “Discussion” section should be written purposefully and directly address the research question raised in the “Introduction” section.

It is also recommended to compare the research results with those of recent, relevant high-quality studies. Wordvice [12] advises that in the “Discussion” section, authors should demonstrate how their research findings fit into the broader scientific literature reviewed in the “Introduction” section. Enago Academy [13] notes that studies mentioned in the literature review should only be used as support and evidence for the research.

Major publishers, journals, and universities also emphasize various aspects and stages of formatting scientific reports and manuscripts [10, 13, 14, 15]. For instance, Nundy [15] recommends using the same key concept in the discussion that was introduced in the introduction. The authors believe that there is no need to repeat citations already used in the introduction.

These and other research findings highlight the close interrelation between the “Discussion” and “Introduction” sections. They also emphasize the need to use recent, relevant, high-quality publications in the discussion. This approach enables the preparation of a paper at a higher quality level.

Hypothesis: The quality of a paper is determined, among other factors, by the repeated use (duplication) of the most significant cited sources from the “Introduction” section in the “Discussion” section.

The aim of this paper is to investigate the nature and impact of duplicated references in the ‘Introduction’ and ‘Discussion’ sections of scientific papers on physical education and sports.

**Material and Methods**

**Information sources**

The information source selected for this study was the authoritative Web of Science Core Collection database. A search in the categories “Hospitality, Leisure, Sport & Tourism and Sport Sciences” and using the keywords “physical education” yielded 1099 documents. From these, 21 highly cited sources were selected as of July/August 2023. These articles are included in the top 1% of academic fields in categories in the Web of Science Core Collection. Additionally, several articles with a high number of citations were also selected.

The procedure for extracting references from the “Introduction” and “Discussion” sections

The procedure for extracting text from scientific articles in PDF format is demonstrated in various publications [16, 17, 18, 19]. However, these tools have limited capabilities for extracting data from PDF files in accordance with the objectives of the current study. Therefore, the text extraction procedure was simplified. A flowchart for visualizing the sequence of actions based on the code is shown in Figure 1. The flowchart describes the process from loading the PDF file to saving the results and visualizations (Figure 2).

The inclusion criteria for selecting documents included the following: the publication must have sections and formatting that are used in most journals according to templates and in the field of physical education and sports. Many such articles adhere to the formatting styles recommended for biomedical journals. The sorting of documents was carried out based on the «Citation Class/Citing Items Breakdown/Background» indicator. Publications from the last 5 years were selected. The Web of Science Core Collection citation classification «What is Citing Items Breakdown?» was used. This analysis helps describe why an article has been cited by using citation classifications assigned to newly published articles to aggregate mentions to this article based on the author’s intent. Background – previously published research that orients the current study within a scholarly area [20]. For comparison, the category «CLINICAL MEDICINE» was selected as the most cited category in the field of «Physical Culture and Sports» («InCites Essential Science Indicators»). In the selection field «Top Papers by Research Fields/Research Fields,» the most cited documents in «CLINICAL MEDICINE» were chosen.

**Data analysis**

To extract references from documents, the PyCharm CE development environment was used in combination with various Python programming language libraries, such as PyPDF2. A code was developed that identifies duplicate references between the «Introduction» and «Discussion» sections. This code also performs analysis and visualization of matching data between these sections, saving the results in CSV format and creating graphical representations for clarity. For statistical calculations, the PyCharm CE development environment was also used in combination with various Python programming language libraries.

The Shapiro-Wilk test was used to assess the normality of data distribution. Correlation analysis between two quantitative variables was conducted using the Spearman’s correlation coefficient. Additionally, to reduce the influence of outliers and check the robustness of the results, the robust
RANSAC (RANdom Sample Consensus) method was applied. Mean and Standard Deviation indicators were calculated for all categories. The significance level was set at 0.05.

**Results**

Table 1 presents the characteristics of highly cited articles in the «CLINICAL MEDICINE» category. Analysis of the table’s data reveals the following:

- The article by Collins et al. [22] has a significant number of citations in the «Background» section, indicating deep integration of the article into existing academic literature.

- The article by Madley-Dowd et al. [23] has a relatively smaller number of citations in «Background» but more in «Discuss», suggesting a greater focus on critical analysis and discussion of existing data.

- In articles where the number of references in «Discuss» exceeds the number in «Background», it can be assumed there is more active involvement in current discussions or critical analysis.

Overall, the data in the table indicate important trends in the use of sources in highly cited articles in clinical medicine.

Similar key trends in the use of sources and citation strategies in highly cited articles in physical education and sports are also presented in Table 1. Articles belonging to the categories 'Highly cited threshold ≤1%' and '>1%' demonstrate different aspects of their impact in the academic environment. A deep integration of some works into existing academic literature is observed, reflecting their significance in the fundamental understanding of this field. The ratio between the total number of references and duplicated references indicates the widespread use of key sources, testifying to their significance in this area.

The analysis of mean values and standard deviations across all three categories indicates the following (Table 1):
The ratio in the three categories between Total Citations and Breakdown (Background) suggests that, on average, about 22.5% of all citations in articles are attributed to the 'Background' section. The Standard Deviation indicates greater variability in the total number of citations compared to the number of citations in the 'Background' section.

The ratio in the three categories between Total Citations and Breakdown (Discuss) suggests that approximately 8.9% of all citations in articles are attributed to the 'Discuss' section. The Standard Deviation indicates greater variability in the total number of citations compared to the number of citations in the 'Discuss' section.

In articles, on average, a larger portion of citations is attributed to 'Background' than to 'Discuss'. Both ratios show significant variability in the total number of citations, which may be related to the diversity of research topics, depth, and scope of the discussed material.

The results of the Shapiro-Wilk test (W-statistic: 0.508; P-value: 0.000) indicated that the distribution of data significantly deviates from normal distribution. Within the scope of the study, an analysis was conducted to examine the relationship between the total number of citations (TC) and the Duplicate Reference Index (DRI). The Spearman's correlation coefficient was used to assess the correlation between these two quantitative variables (Figure 3). The results showed that the Spearman's correlation coefficient between TC and DRI was 0.645 (p = 0.0016). These findings indicate a statistically significant relationship between the number of citations of articles and the frequency of duplicate references within them.

The presence of TC values such as 4726 (Table 1), which significantly exceed other values, may indicate outliers. Therefore, the robust RANSAC (RANdom SAmple Consensus) method was used for additional verification of the results. This approach showed that the relationship between TC and DRI remains even after reducing the influence of outliers (Figure 3). This indicates that the initial conclusions about the significance of duplicate references for the citability of articles remain valid even under more stringent statistical analysis. The application of the robust RANSAC method as an additional analytical tool underscores the strength and reliability of the results.

Discussion

In our study, we analyzed the nature and impact of duplicate references in the 'Introduction' and 'Discussion' sections of scientific articles on physical education and sports. Based on our hypothesis that the quality of an article is partially determined by the repeated use of key cited sources, our results confirm a statistically significant relationship between the frequency of citations of an article and the number of duplicate references. However, it is important to note that duplication of references should be done with consideration for academic ethics.

Our results largely align with previous studies.
Table 1. Characteristics of highly cited articles in Web of Science

<table>
<thead>
<tr>
<th>Articles (n=21)</th>
<th>Highly cited threshold (%)</th>
<th>Total Citations (TC)</th>
<th>Breakdown (Background)</th>
<th>Breakdown (Discuss)</th>
<th>Total References</th>
<th>Total Duplicate references</th>
<th>Duplicate References Index (DRI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category &quot;Clinical Medicine&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collins et al. [22]</td>
<td>≤1%</td>
<td>4726</td>
<td>1333</td>
<td>335</td>
<td>56</td>
<td>2</td>
<td>3.57</td>
</tr>
<tr>
<td>Madley-Dowd et al. [23]</td>
<td>≤1%</td>
<td>427</td>
<td>22</td>
<td>91</td>
<td>44</td>
<td>7</td>
<td>15.91</td>
</tr>
<tr>
<td>Garrity et al. [24]</td>
<td>≤1%</td>
<td>372</td>
<td>14</td>
<td>12</td>
<td>32</td>
<td>1</td>
<td>3.15</td>
</tr>
<tr>
<td>Santesso et al. [25]</td>
<td>≤1%</td>
<td>329</td>
<td>3</td>
<td>6</td>
<td>26</td>
<td>9</td>
<td>34.62</td>
</tr>
<tr>
<td>Quan et al. [26]</td>
<td>≤1%</td>
<td>298</td>
<td>3</td>
<td>4</td>
<td>50</td>
<td>7</td>
<td>14.00</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>1230.4</td>
<td>275.0</td>
<td>89.6</td>
<td>41.6</td>
<td>5.2</td>
<td>14.25</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td>1954.7</td>
<td>591.49</td>
<td>141.92</td>
<td>12.44</td>
<td>3.49</td>
<td>12.8</td>
</tr>
<tr>
<td>Category &quot;Hospitality, Leisure, Sport &amp; Tourism&quot; and &quot;Sport Sciences&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edwards et al. [27]</td>
<td>≤1%</td>
<td>241</td>
<td>81</td>
<td>50</td>
<td>94</td>
<td>15</td>
<td>0.16</td>
</tr>
<tr>
<td>Donnelly et al. [28]</td>
<td>≤1%</td>
<td>1002</td>
<td>229</td>
<td>87</td>
<td>168</td>
<td>138</td>
<td>0.82</td>
</tr>
<tr>
<td>Jayanthi et al. [29]</td>
<td>≤1%</td>
<td>328</td>
<td>48</td>
<td>42</td>
<td>20</td>
<td>33</td>
<td>1.65</td>
</tr>
<tr>
<td>Lloyd et al. [30]</td>
<td>≤1%</td>
<td>277</td>
<td>42</td>
<td>32</td>
<td>243</td>
<td>72</td>
<td>0.50</td>
</tr>
<tr>
<td>Martins et al. [21]</td>
<td>≤1%</td>
<td>180</td>
<td>31</td>
<td>27</td>
<td>50</td>
<td>23</td>
<td>0.46</td>
</tr>
<tr>
<td>Fan et al. [31]</td>
<td>&gt;1%</td>
<td>96</td>
<td>18</td>
<td>13</td>
<td>33</td>
<td>17</td>
<td>0.52</td>
</tr>
<tr>
<td>Mountjoy et al. [32]</td>
<td>&gt;1%</td>
<td>217</td>
<td>59</td>
<td>33</td>
<td>32</td>
<td>49</td>
<td>1.53</td>
</tr>
<tr>
<td>Casamichana et al. [33]</td>
<td>&gt;1%</td>
<td>235</td>
<td>36</td>
<td>25</td>
<td>38</td>
<td>8</td>
<td>0.21</td>
</tr>
<tr>
<td>Kalajas-Tilga et al. [34]</td>
<td>&gt;1%</td>
<td>87</td>
<td>37</td>
<td>18</td>
<td>52</td>
<td>74</td>
<td>1.42</td>
</tr>
<tr>
<td>Lander et al. [35]</td>
<td>&gt;1%</td>
<td>103</td>
<td>31</td>
<td>17</td>
<td>111</td>
<td>90</td>
<td>0.81</td>
</tr>
<tr>
<td>An et al. [36]</td>
<td>&gt;1%</td>
<td>73</td>
<td>25</td>
<td>14</td>
<td>71</td>
<td>7</td>
<td>0.10</td>
</tr>
<tr>
<td>Kang et al. [37]</td>
<td>&gt;1%</td>
<td>62</td>
<td>24</td>
<td>5</td>
<td>59</td>
<td>12</td>
<td>0.20</td>
</tr>
<tr>
<td>Ang et al. [38]</td>
<td>&gt;1%</td>
<td>105</td>
<td>24</td>
<td>23</td>
<td>45</td>
<td>10</td>
<td>0.22</td>
</tr>
<tr>
<td>Giblin et al. [39]</td>
<td>&gt;1%</td>
<td>120</td>
<td>23</td>
<td>6</td>
<td>39</td>
<td>4</td>
<td>0.10</td>
</tr>
<tr>
<td>Costigan et al. [40]</td>
<td>&gt;1%</td>
<td>111</td>
<td>22</td>
<td>10</td>
<td>44</td>
<td>3</td>
<td>0.07</td>
</tr>
<tr>
<td>Okano et al. [41]</td>
<td>&gt;1%</td>
<td>102</td>
<td>30</td>
<td>19</td>
<td>48</td>
<td>7</td>
<td>0.15</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>208.69</td>
<td>47.5</td>
<td>24.94</td>
<td>71.69</td>
<td>35.12</td>
<td>0.54</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td></td>
<td>226.42</td>
<td>51.02</td>
<td>19.46</td>
<td>58.69</td>
<td>39.18</td>
<td>0.54</td>
</tr>
</tbody>
</table>

Note. Articles: Title of the Article and Authors. Highly Cited Threshold: High Citation Threshold (≤1%). Total Citations (TC): Total Number of Citations. Breakdown (Background): Citations in the 'Background' Section (Background Research). Breakdown (Discuss): Citations in the 'Discuss' Section (Discussion). Total References: Total Number of References. Total Duplicate References: Total Number of Duplicate References. Duplicate References Index (DRI): Duplicate References Index (a metric reflecting the ratio of duplicate references to the total number of references in References). Background - previously published research that orients the current study within a scholarly area. Discuss—references mentioned because the current study is going into a more detailed discussion.

[8, 9, 11, 15], which also emphasized the importance of a careful and thoughtful approach to citing in scientific articles. However, our analysis expands on these insights by providing specific data on the impact of duplicate references on the citability of articles in the field of physical education and sports. This text deepens the discussion by connecting your results with the broader context of academic writing and research practices, and also highlights the importance of balance and academic ethics in the use of citations.

In our study, we confirm the idea proposed by Jawaid et al. [7] about the importance of the relationship between the «Introduction» and...
«Discussion» sections in scientific articles. Our results show that the duplication of references between these sections correlates with an increase in the citation rate of the article, supporting the assumption about the significance of these sections for the quality of the article.

According to Vieira et al. [8], starting the discussion with a reminder of key information from the «Introduction» section helps to contextualize the research. In our case, the duplication of references serves not just as repetition, but as a reinforcement of the main ideas and support for the depth of argumentation. Day et al. [9] and Goulston [10] emphasize the importance of interaction between the «Introduction» and «Discussion» sections. We found that using key sources from the «Introduction» in the «Discussion» helps to create a more cohesive and convincing presentation of the material, addressing the research questions posed at the beginning of the work. In another study Bavdekar [11], it is pointed out that the «Discussion» should be directed towards answering the research questions raised in the «Introduction.» Our results show that duplicating references can facilitate this process, making the discussion more focused and substantive. In line with the recommendations of Wordvice [12] and Enago Academy [13], we aimed to demonstrate how our results fit into the broader scientific literature reviewed in the «Introduction.» This helped to highlight the contribution of our research to the existing field of knowledge. In accordance with the recommendations of Nundy [15], we used the same key concepts in the «Discussion» that were presented in the «Introduction,» avoiding unnecessary repetition but reinforcing the connection between the main themes and research findings.

It should be noted that the analysis of highly cited articles confirms our assumptions about the importance of duplicating references for citation frequency. Our study shows that articles with a higher Duplicate References Index (DRI) often have a higher level of citations. For example, works such as Collins et al. [22] and Madley-Dowd et al. [23], with DRIs of 3.57 and 15.91 respectively, demonstrate a significant number of citations. This indicates a positive correlation between the duplication index and citation frequency.

Articles with a relatively low duplication index, such as Andermo et al. [36] and Giblin et al. [39] with a DRI of 0.10, can also achieve a high level of citations, indicating the complexity and multifaceted nature of factors influencing citation frequency. This highlights that while duplicating references is an important aspect, it is not the sole factor determining an article’s citability.

It is important to emphasize that our study has several limitations that should be considered when interpreting the results. One of the key limitations is related to the availability of the scientific articles analyzed. Many highly cited publications are behind paywalls or presented in formats that limit analysis capabilities, such as in protected PDF files. This led to the exclusion of a significant portion of potentially relevant works from our analysis, which in turn limited the scope and diversity of our data.
Additionally, the limited sample of articles could have impacted the generalizability of our results. While we aimed to cover a broad spectrum of articles in the field of physical education and sports, the small sample size and the specific nature of the analyzed publications may not fully reflect trends across the entire field. Therefore, the results of our study should be viewed as a research contribution that requires further verification and refinement in future work.

It should be noted that despite these limitations, our study provides valuable insights into the relationship between the use of duplicate references in scientific articles and their citation frequency. We recommend that future research expand on this work by including a more extensive sample of articles and applying a more diverse range of methodological approaches.

Conclusions
Our data show a statistically significant relationship between the frequency of duplicate references in the «Introduction» and «Discussion» sections and the citation level of articles on physical education and sports. The duplication of key sources in these sections can be an indicator of the depth of research and the quality of the scientific argument, which, in turn, contributes to increasing the citation frequency of the article.

The results also highlight the importance of conscious and strategic use of literature in scientific articles, which can improve their impact and visibility in the academic community. Despite the identified relationship, the results also point to the complexity and multifaceted nature of factors influencing the citation frequency of an article, including the quality of content and the context of publication.

References
12. Wordvice HJ. How to Write a Discussion Section for a Research Paper. [Internet] 2022 [updated 2022 Oct 12; cited 2023 Jun 26]; Available from: https://blog.wordvice.com/research-writing-tips-editing-manuscript/discussion/


**Information about the authors:**

**Leonid Podrigalo;** (Corresponding Author); https://orcid.org/0000-0002-7893-524X; leonid.podrigalo@gmail.com; Department of Physical Culture and Sports Rehabilitation, Kharkiv State Academy of Physical Culture; Kharkiv, Ukraine.

**Sergii Iermakov;** https://orcid.org/0000-0002-5039-4517; sportart@gmail.com; Department of Methodologies of Cross-Cultural Practices, Kharkiv State Academy of Design and Arts; Kharkiv, Ukraine.

Cite this article as:
https://doi.org/10.15561/health.2023.0203

This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (http://creativecommons.org/licenses/by/4.0/deed.en).

Received: 28.11.2023
Accepted: 29.12.2023; Published: 30.12.2023