Exploring Trends in Education Program Evaluation Models in High Schools: Bibliometric Analysis

1755 – 1574

Mariano Dos Santos[1], Ahmad[2]*

[1] Instituto Católico para a Formação de Professores (ICFP) Baucau, Timor - Leste,

[2] Bumigora University Mataram, Indonesia.

Email Correspondence: ahmad@universitasbumigora.ac.id*

Abstract: This study aims to analyze the program evaluation model applied in high school using a bibliometric approach integrated with interpretive analysis. The main problem identified was the importance of understanding the various evaluation models that are often applied in education and their influence on improving the quality of the program. The method used in this study is a bibliometric analysis by selecting 120 articles published between 1985 and 2024 from the Scopus database related to the evaluation of educational programs in secondary schools. Data was collected through a search for articles with relevant keywords, and an analysis was conducted to find thematic trends and collaborations between authors in this field. Data were analyzed using the prism method and *the R Studio application* with the Bibliometrix::biblioshiny() package to explore thematic trends, author collaboration networks, and conceptual structures. The results show that evaluation models, such as the Outcome-Mediation Cascade (OMC), TADIPHE, and EP_PISTdu, are increasingly relevant in complex educational contexts. The study also found the importance of gender elements in program evaluation, with a focus on equality between women and men. In addition, international collaborations between authors from different countries, such as the United States and Canada, are instrumental in enriching the quality of program evaluation research. This study suggests that the development of more innovative and relevant evaluation models should be carried out to increase the effectiveness of educational program evaluations in the future.

INTRODUCTION

Evaluation is the process of assessing the strengths and weaknesses of something that is needed every day in a variety of different contexts and purposes (D'Inverno et al., 2021; Garden & Tang, 2019; Yasin et al., 2023). In everyday life, many laypeople unconsciously apply evaluation in decision-making or assessment (Amouzou et al., 2022; Blizniuk et al., 2014). The ever-changing dynamics of life and different needs encourage the application of diverse evaluations, as the objects, programs, and projects of evaluation are highly dependent on the context, needs, and goals to be achieved. Therefore, evaluation methods and approaches develop into various models adapted to these parameters (Watt, 2017). This shows that evaluation is an inherent and inseparable aspect of everyday human life.

In the context of program evaluation, an effective evaluation must be able to comprehensively assess the object or evaluation and uncover various problems that may arise (Amahoroe, 2023). This evaluation acts as a data-driven scientific paper that serves to guide and recommend policy or overall decision-making. The main purpose of program evaluation is to determine the success of program implementation, using quality measures that have become classic standards, namely effectiveness, efficiency, adequacy, fairness, responsiveness, suitability (as needed), and sustainability (Dolor et al., 2010; Mustofa et al., 2023).

It must also be understood that a program does not just emerge, but is based on a policy that ultimately depends on political will. For those who lack a good understanding of program evaluation, the close relationship between policies, evaluations, and projects is often not well realized (Guyadeen & Season, 2018; Nouraey et al., 2020). Policies are based on a philosophical foundation with clear long-term goals, while programs are a more specific set of activities with concrete objectives that must be achieved once the program is implemented (Browne et al., 2021) . Projects, on the other hand, are more focused activities and are designed to respond to specific problems in society, support policy implementation, or actualize specific individual goals (Al-Alawi & Alexander, 2020; Chanthalangsy et al., 2024; Saeed et al., 2023).

In the context of education, specifically the evaluation of an educational program managed by a single individual or team, a program consists of a set of resources and activities mobilized to achieve a specific goal (Hidayah & Romelah, 2022; Vieira et al., 2019). In addition to activities and objectives, educational programs also involve resources such as budget, staff (human resources), and limited time. Evaluation of educational programs has an important role, including providing information for decision-making and policy development, assessing student learning outcomes, evaluating curriculum, increasing school trust, monitoring the use of funds, and improving educational materials and programs (Tang & Hai, 2021; Vaiopoulou et al., 2023; Yangari & Inga, 2021). This evaluation process includes selecting types and sources of information, data collection, and analysis to produce informative reports that support decision-making and selection of appropriate alternatives. After evaluation, a decision can

be made to discontinue, upgrade, or disseminate the program (Asrul et al., 2014; Kania et al., 2023).

Research related to program evaluation is carried out in various contexts, including systematic review, meta-analysis, bibliometric analysis, application of certain models, or reporting of program evaluation results. However, systematic reviews are often criticized for their high likelihood of subjectivity, making them vulnerable to coming to less convincing conclusions. Meta-analysis, which is quantitative in nature, is less suitable for non-quantitative scientists, while bibliometric analysis is criticized for its limited contribution to the development of theories (Fu et al., 2023; Khan, 2022). Recently Liu et al., (2023) suggests an integrative approach to enhancing the value of bibliometric studies by combining various analytical techniques. In addition to this study, there is also relevant previous research that examined the evaluation model of education programs in secondary schools. For example, a study by Hidayah & Romelah, (2022)Evaluate early childhood education programs with a focus on the effectiveness of resource use and the achievement of learning objectives. While Chanthalangsy et al., (2024) applying the CIPP model to evaluate graduate education programs in Laos, demonstrating the importance of local context in evaluation design. In addition, research by Mustofa et al., (2023) adopts the Kirkpatrick model to assess teacher training, emphasizing behavior change as an indicator of program success.

However, of the three papers, none specifically examined trends in the use of evaluation models at the secondary school level over a wide span of time with a bibliometric approach integrated with interpretive analysis. This section is important to study because it can identify dominant evaluation models, new models that are rarely used, and long-term trends and researcher collaborations. Thus, this study offers an update by presenting a thematic map and conceptual network of evaluation of educational programs in secondary schools based on 120 articles from 1985 to 2024.

The integration of systematic reviews such as interpretive analysis with quantitative bibliometric analysis can enrich the results of the analysis, while also opening up opportunities to leverage advanced technologies such as machine learning and artificial intelligence to mine and analyze big data more effectively (Donthu et al., 2021). This integrated interpretive-bibliometric approach is expected to reduce reasoning errors and make a more significant epistemological contribution to the field of program evaluation. The development of interest in program evaluation, including trends and mapping of the literature, is often driven by a variety of personal and academic motivations. Experienced researchers can use tools like Google Alerts to monitor the latest developments online, while novice researchers, such as doctoral students, often use large databases like Scopus to mine relevant literature. Bibliometric studies like these help them map trending topics to avoid duplication of research and follow the direction of research development (Ellegaard & Wallin, 2015; Guo et al., 2019; Tigre et al., 2023).

As one of the methods that bridge quantitative and qualitative insights, bibliometric studies serve as a 'preview lens' to expand the horizons of the research world. Integrated interpretive-bibliometric analysis is able to reveal the author's social

network, conceptual relationships, confiscation patterns, and dominant themes in the field of educational program evaluation, especially at the high school level. Thus, the main objective of this study is to comprehensively analyze the social interactions between authors, conceptual networks and quotations, as well as dominant themes in publications that discuss the evaluation of educational programs in secondary schools during the period 1985 to 2024.

METHOD

This study uses a quantitative approach with a literature study method to analyze relevant articles regarding program evaluation in secondary schools. The quantitative approach with the literature study method was chosen because it allows a systematic analysis of large amounts of publication data, especially to identify trends, dominant evaluation models, and patterns of author collaboration in the field of evaluation of educational programs in secondary schools. The location of this research was carried out through the Scopus database which provides access to scientific articles related to the topics discussed. The study population consisted of articles published between 1985 and 2024, and the sample of this study totaled 120 articles (n=120). The main criteria for articles to be 'included' are those that contain mentions of "program evaluation" and "high school", both in the title and in the abstract.

Data or articles are pulled from the Scopus database using primary search queries, i.e. "high school" AND "program evaluation", as well as variations such as "evaluation model" AND school AND education AND program. Applying the entire filter yields the following advanced search queries: TITLE-ABS-KEY (high AND school AND "program evaluation") AND PUBYEAR > 1985 AND PUBYEAR < 2025 AND (LIMIT-TO(SUBJAREA, "SOCI") OR LIMIT-TO(SUBJAREA, "ARTS")) AND (LIMIT-TO(EXACTKEYWORD, "Education") OR LIMIT-TO(EXACTKEYWORD, "PROGRAM EVALUATION") OR LIMIT-TO(EXACTKEYWORD, "EVALUATION STUDY") AND (LIMIT-TO(LANGUAGE, "English")).

The sample of articles considered in this study was selected through the process of article identification, screening, eligibility, and inclusion steps described in the following prism method flowchart (January, 2024).
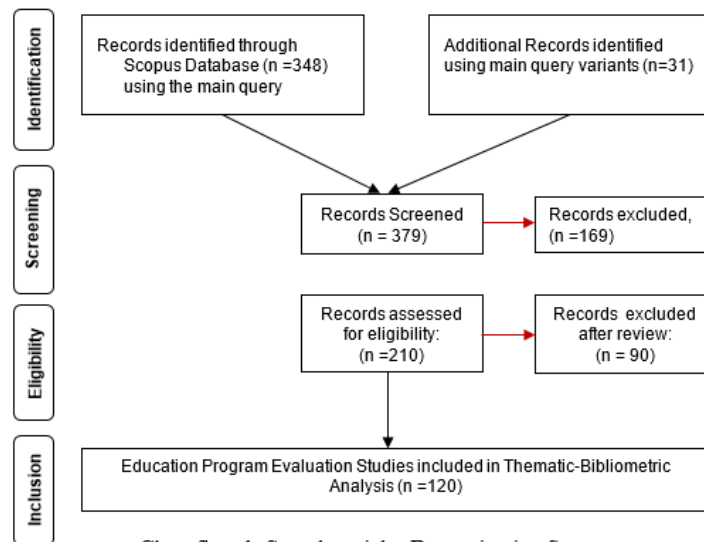
Chart flow 1: Sample articles Determination Steps

Figure 1. Chart Flow

The flowchart in Figure 1 illustrates the stages in determining the sample of articles used in thematic-bibliometric analysis. The process begins with the identification of articles through the Scopus database, which generates 348 articles using the main query, plus an additional 31 articles found through the main query variation. Then, out of a total of 379 articles screened, 169 articles were removed because they did not meet the relevance criteria. The remaining 210 articles were then deemed eligible, and after evaluation, 90 articles were removed for not meeting the eligibility standards. Finally, 120 articles that met all the inclusion criteria were included in the thematical-bibliometric analysis, which became a valid research sample.

Data analysis in this study was carried out using the prism method and *the R Studio application* with the Bibliometrix:: biblioshiny() package which is specially designed for bibliometric analysis. The application is used to extract, visualize, and analyze publication data from the Scopus database, including keyword trend analysis, author collaboration networks, thematic maps, and conceptual structures. The role *of Biblioshiny* is very important in facilitating the process of analyzing bibliometric data in an interactive and systematic manner, resulting in more valid, structured, and easily interpretable findings.
.

RESULTS AND DISCUSSION
Result
1. Evaluation Models and Keywords: Interpretive Synthesis

To determine the number of mentions of the phrase "evaluation model" in the abstract and article title, the search feature in MS Word is used by typing the phrase in the search box. As a result, 47 articles were found that included the keyword "program evaluation". However, we cannot conclude that only 47 of the 120 articles included the keyword "program evaluation" in their titles or abstracts. This is due to the fact that in some articles, there is more than one mention or number for a keyword in a single

abstract. In fact, in five cases more than one mention was found, with one of the abstracts listing "program evaluation" five times.

2. Risk of Error

In line with these findings, we note that the number of evaluation models and the frequency of their mention are the main concerns in this study. The searches and filters used may cause variations in the number of keyword mentions, which affects the accuracy of our findings. In addition, these findings provide insight into how the keyword "program evaluation" is often used in diverse contexts in the articles studied, reflecting the diversity of evaluation models adopted. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 1 below.

Table 1. Number of Evaluation Models

Yes	Kind	Important
1	Kirkpatrick	13
2	Face	1
3	CIPPO	1
4	Formative Writer	4
5	UCLA CSE	3
6	Goal-oriented	1
7	Driven by theory	3
8	Outcome Mediation Cascade (WTO)	1
9	TADIPHE	1
10	4C	1
11	CIPP	10
12	Relational System Evaluation (CSR)	1
13	EP_PISTdu	4

For starters in table 1, the number or mention of keywords in bibliometric analysis can be misleading if calculated on a frequency basis only. Keyword frequency should not be calculated twice in the same modality. Table 1 shows the case with the mention of duplicate keywords. In some abstracts, the name of the evaluation model is mentioned more than once: twice for the formative model, three times for the UCLA CSE, four times for the 4C model, and five times for the EP_PISTdu model. Given the abstract duplication of keywords, there is a concern that arguments based on this number of mentions may be invalid. Bibliometric analysis computer programs that do not take duplicate keywords into account can generate a *cloud of incorrect words*, which can cause researchers to make invalid arguments, leading to incorrect thinking.

Similarly, many abstracts that meet the criteria have a clear evaluation and include the phrase "high school". However, canonically, one would question how the evaluation of educational programs can be conducted conclusively without being supported by a clear evaluation model. In other words, a strict selection of articles that include the names

of the evaluation models in the abstract will lead to a reduction in the sample of this study to only 42 articles (n=42).

This condition poses challenges in the world of research, especially regarding the use of words such as "evaluation" or "assessment", as well as their related variants, by those who do not understand the evaluation of educational programs. This can lead to errors in bibliometric studies. Bibliometric analysis researchers who analyze hundreds or even thousands of articles may not be aware of the existence of keyword duplication, which can lead to incorrect conclusions because they are based on the wrong premise. Researchers who are limited in time and do not conduct systematic reviews can ignore this duplication, so the resulting scientific mapping becomes flawed. Mentioning keywords more than once in the abstract can lead to errors in the mapping of interests, trends, and the field or subfields of research being studied, ultimately leading to unfounded conclusions.

a. Uncovering Rarely Used or New Evaluation Models

Popularity in research and development encourages the creation of educational products in the academic realm. Most universities, especially in Indonesia, have regulations that require all doctoral students to develop educational products, including evaluation models. Students are not only required to develop a program evaluation model, but also test the effectiveness of the model. This may also be true at other universities, which may explain the emergence of new or rarely used evaluation models.

The articles exemplified in this study contain models that are rarely found in the existing literature. For example, between 1985 and 2024, scientific papers related to the evaluation of educational programs indexed by Scopus recorded a number of unusual models, as will be described below. This can be seen in the results of the analysis using the *R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 2 below.

Table 2. Rarely Used or New Program Evaluation Models

Yes	Kind	Number of Articles
1	Driven by theory	15
2	Outcome Mediation Cascade (WTO)	2
3	TADIPHE	1
4	4C	3
5	Relational System Evaluation (CSR)	6
6	EP_PISTdu	4

Of the six evaluation models that are relatively new or rarely used as shown in Table 2, namely Theory-driven, Outcome-Mediation Cascade (OMC), TADIPHE, 4C, Relational System Evaluation (RSE), and EP_PISTdu—three of them, namely TADIPHE, 4C, and EP_PISTdu, require additional explanation because their meaning cannot be understood directly from the name of the model alone. The TADIPHE model, for example, consists of seven main components: target, assessment, design, installation, process,

outcome, and effectiveness. The 4C model stands for Creativity, Critical Thinking, Communication, and Collaboration, which emphasizes the evaluation of 21st century skills. Meanwhile, EP_PISTdu refers to the Evaluation of the Integrated Social Science Learning Program at the junior high school level, where the initials EP can be interpreted as "Program Evaluation" in the context of Indonesian.

This shows that the dominance of established evaluation models opens up space for the emergence of new or rarely used evaluation models. The presence of these models in Scopus-indexed journals, both in Quartiles 1, 2, 3, and 4, is expected to encourage their dissemination. We carefully argue that researchers who want to use trending or newly developed evaluation models now have more diverse alternatives.

b. Bibliometric Analysis Results (Quantitative)

Figure 1 below provides an overview or important information regarding the data collected. If we visualize the data on the first line, we can see the time span covered which is from 1985 to 2024. A total of 492 authors have published 120 articles in 87 different journals. In this article, 380 keywords were found mentioned and a total of 4609 references were recorded in 120 documents. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in figure 2 below.

Figure 2. Main Information

c. Social Structure

Information related to social structure in examples of scientific papers can be obtained from units such as institutions and countries. The interaction between writers can be seen in co-authorship. With 9.167% of scholarly papers involving international co-authorship, it shows that at least 45 out of 492 authors partner with their international peers in writing scientific papers related to program evaluation in high school. The average number of co-authors per document is 4.4.

Although not explicitly, Figure 1 shows the social structure in the published work. This information can be deduced from the pattern of co-authorship. With 9.167% international authorship and an average of 4.4 co-authors per document, it can be assumed that around 45 researchers are involved in co-authoring articles. It reflects the significant social interaction between authors in work related to the evaluation of high school programs. Co-authorship also has the potential to improve the quality of published articles. When associated with an institution, elements of social structure can be visually

visible through collaborative networks, as illustrated in Figure 2 using *the analysis of the R Studio application* with the Bibliometrix::biblioshiny() package below.
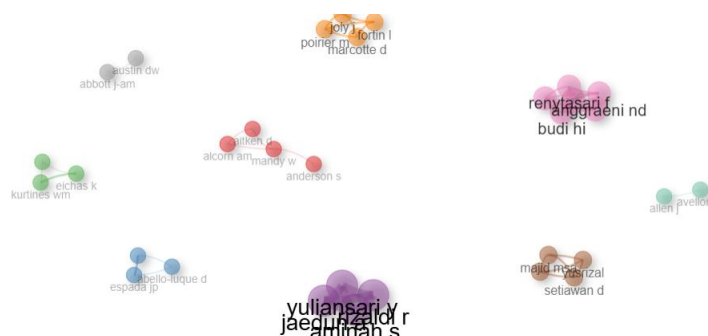

Figure 3. Collaboration Network

Figure 3 illustrates the network of scientific collaboration among the 492 authors in the sample article. There are nine clusters or groups of nodes that are closely interconnected. However, three of them showed a thicker edge, which signified that these three groups collaborated frequently, resulting in a stronger level of collaboration in program evaluation. Given that the nodes in Table 3 refer to authors, it is important to deepen our understanding of the nature of collaboration based on the degree of interconnectedness and closeness between them. Further explanation of clusters and nodes will be provided in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 3 below.

Table 3. Author Groups, Proximity, and Between

Knot	Group	Between	Closeness	Page Rank
Mandy W	1	2	0.33	0.04
Aitken D	1	0	0.25	0.03
Alcorn am	1	0	0.25	0.031
Anderson s	1	0	0.2	0.018
JP Sword	2	0	0.5	0.035
Morales has	2	0	0.5	0.035
Abello-Luque D	2	0	0.5	0.024
Eichas K	3	0	0.5	0.035
WM Curtain	3	0	0.5	0.035
Albrecht Re	3	0	0.5	0.024

Table 3 complements Figure 3 by showing the three main groups and members of the collaboration group, and illustrates the degree of closeness and intermediation between the authors. In cluster 1, authors had a stronger degree of intermediation than their proximity. On the other hand, the intermediation index for clusters 2 and 3 is 0, whereas these two clusters have the same degree of proximity among the authors. This

can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 4 below.

Table 4. State Collaboration Network

From	You	Frequency
Brazil	Portugal	1
Canada	India	1
Canada	Norway	1
Colombia	Spain	1
Italy	France	1
English	Australia	1
English	Spain	1
United States	Canada	2
United States	Colombia	1
United States	Finland	1

Table 4 shows the author's country of origin and shows how the countries collaborate in the evaluation of programs related to secondary school. The most dominant collaboration occurred between the United States and Canada, which was recorded more than once. The United States and Canada are the countries with which authors collaborate the most, while Canada shows a higher level of collaboration, as seen in the 'from' and 'to' columns.

It can be concluded that there are nine communities of implicit writers who contribute to the production of scientific papers on program evaluations carried out in high schools. Three of these groups are more collaborative, with authors from the United States most often collaborating with international peers. Meanwhile, Canadian authors show the highest level of collaboration as they not only seek out international co-authors, but also receive contributions from several international peers.

d. Conceptual Structure

To get information about this structure, keyword analysis is used. Through this analysis, we can track the interactions or relationships between concepts, keywords, or topics that provide an idea of the dominant concept or topic discussed in the sample. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in figure 4 below.
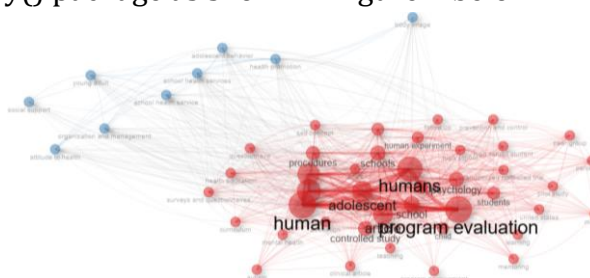

Figure 4. Co-Emergence Networks

Given the "evaluation model" of the basic query AND the school AND the school AND the program AND education AND and the associated variants, co-occurrence is the frequency of occurrence or how often it happens that a particular term appears along with that base query. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 5 below.

Table 5. Among the Words that Appeared at the Same Time

Knot	Group	Between
Program Evaluation	1	62.88903861
woman	1	46.84233641
Man	1	34.89268626
Adolescent	1	27.46319257
Stuart	1	67.38625231
Human	1	57.98120445
Article	1	33.56791644
child	1	4.142791694
School	1	13.39252799
School	1	11.72865977

Table 5 shows that 'program evaluation', 'male', and 'female' had the highest intermediate levels related to the basic question or theme of the high school program evaluation. The concept of 'male', with intermediate levels of 67.39 and 57.98 respectively, suggests that secondary school education and related interventions/programmes are conducted for and with men. The 'male' element includes students, teachers, stakeholders, and the evaluation team members themselves. The pedagogical approach, student welfare, teacher training, and other aspects are aimed at the benefit of the beneficiaries, who can be identified as women, men, and adolescents.

The node or concept of 'program evaluation', with an intermediate rate of 62.89, has a big role. An example article shows that the researcher evaluated the program implemented in the school. Aspects that can be assessed through evaluation include monitoring/formative, effectiveness, and impact of interventions on participants. The 'women' node (with an average rate of 46.84) indicates that the programs evaluated may be related to interventions to address gender gaps, educational opportunities, and/or policies. These keywords connect discussions about gender-specific programs, equality, and inclusivity. Some papers in the sample reported evaluation of interventions related to gambling, sexual behavior, well-being, or the impact of low income on female students compared to male students.

e. Intellectual Structure

To obtain information about the intellectual structure around the basic questions or main themes of this study, the analysis of scientific communication patterns, citations, and the simultaneous occurrence of keywords can be seen in the results of the analysis

using *the R Studio application* with the Bibliometrix:: biblioshiny() package as shown in table 6 below.

Table 6. Overtime Dominant Program

Paper	Year	LCS	GCS	Group
Poirier M, 2013, Educ Res Eval	2013	1	8	1
Poirier M, 2017, Health Sciences Quebec	2017	0	3	1
Doumas Dm, 2015, J Subst Abuse Treatment	2015	1	13	2
Doumas Dm, 2017, Addictive Behavior	2017	0	13	2
Cavell Ta, 2018, Teenage Children Serv Rev	2018	1	5	3
Glass, 2023, Educ Urban Soc	2023	0	11	3

Given the citations from time to time, Table 6 shows that six authors have made major contributions to the following dominant programs: intervention programs aimed at preventing depression in adolescents at risk of dropping out, including prevention of depression in adolescents, intervention programs to reduce alcohol consumption among high school students, intervention programs to reduce cognitive risk factors associated with alcohol consumption, as well as improving protective behavior strategies, evaluation of mentoring programs implemented in high schools, and the effectiveness of mentoring programs in high schools for first-generation student enrollment.

In addition, there are two other indexes that describe the intellectual structure of the sample article: the most frequently cited keywords and the trending topics. Regarding the most frequent keywords, there is evidence of a significant trend in the frequency of keyword use related to the theme of program evaluation conducted in high school. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in table 7 below.

Table 7. Most Frequent Keywords

Word	Event
Program Evaluation	116
Woman	93
Man	81
Adolescent	77
Stuttgar	64
Human	56
Article	39
Child	38
School	30
School	28

First, the term 'program evaluation' emerged as the dominant keyword, recorded 116 times across the sample article. This shows how important this concept is in the

discourse of program evaluation in high school. As can be seen, Figure 4 complements Table 7 with a more in-depth visualization. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in figure 5 below.
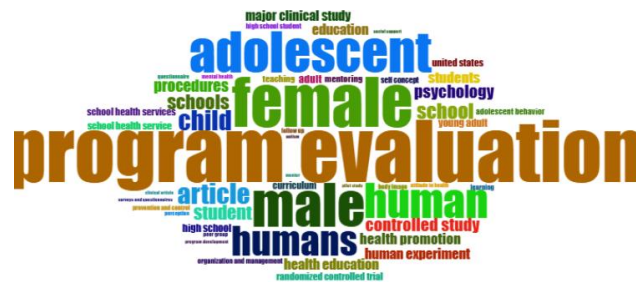

Figure 5. Word Cloud

In addition, gender-related keywords, particularly 'female' (93 incidents) and 'male' (81 incidents), highlighted a focus on gender-specific outcomes, such as interventions for women related to gender or equality, versus issues related to male students, such as alcohol use or substance abuse, or program beneficiaries who are female or male in secondary school programs. These terms collectively contribute to the core intellectual structure of this bibliometric study.

While trending topics and keywords have similarities, they both have differences. The topic is broader, although there are some examples of keywords that could also be considered topics.

Table 8. Trending Keywords

Items	Freq	year_q1	year_med	year_q3
Education	20	2016	2017	2020
Mental Health	8	2016	2017	2019
Perception	8	2017	2017	2019
woman	93	2017	2018	2020
Man	81	2017	2018	2020
Adolescent	77	2017	2018	2020
Program Evaluation	116	2017	2019	2021
Stuart	64	2017	2019	2021
Human	56	2017	2019	2021
Article	39	2018	2020	2022

There is also evidence that the programs evaluated have a strong focus on the humanitarian aspect. This can be seen from the emergence of keywords such as 'adolescent' (77 incidents), 'human' (64 incidents), and 'human' (56 incidents), which shows that most educational programs in high school are designed and evaluated with their impact on the individual as a learner. The emergence of these keywords is in line with the tendency of program evaluation that not only focuses on structural or administrative aspects, but also on welfare and human development as the core of

educational goals. This can be seen in the results of the analysis using *the R Studio application* with the Bibliometrix::biblioshiny() package as shown in figure 6 below.
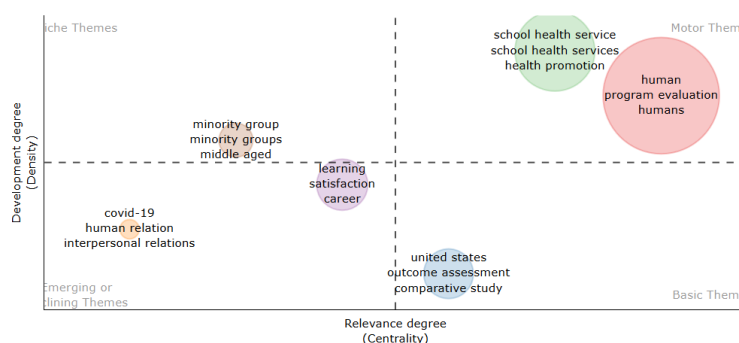

Figure 6. Thematic Map

Figure 6 shows a thematic visual map of program evaluation implemented in high school. The map consists of six clusters, with three larger clusters based on the size of CallonCentrality and CallonDensity, as well as three clusters of average size. The 'human' cluster stands out with a CallonCentrality value of around 64.39. From this map, it can be concluded that most of the programs implemented in secondary schools aim to improve human quality through aspects of health, well-being, and interventions related to curriculum/materials. The evaluators evaluate the strengths and weaknesses of the program based on its impact on people as program participants.

Discussion

This study reveals the results of interpretive-bibliometric analysis related to the program evaluation model applied in high school. The main findings of this study show the importance of evaluation in the context of education, especially in identifying the strengths and weaknesses of programs implemented in secondary schools. The combined approach of bibliometric and interpretive analysis provides deeper insight into the evaluation models used, as well as the collaborative networks between authors that contribute to the development of knowledge in this field. From the results of the analysis, it was found that the term "program evaluation" appeared as the dominant keyword in the sample of articles analyzed. This confirms that evaluation is a very important element in designing and assessing the success of educational programs. In line with previous findings by Jáuregui et al., (2022) which states that program evaluation serves to identify the advantages and disadvantages of programs implemented in schools. The existence of new or rarely used evaluation models, such as the Outcome-Mediation Cascade (OMC) and EP_PISTdu, shows that research in this area continues to evolve and adapt to the new challenges faced in education. This also reflects the dynamics of the world of education that continues to adapt to changing needs.

The use of more structured evaluation models, such as the 4Cs (Creativity, Critical Thinking, Communication, and Collaboration), offers a more holistic approach to assessing education. These models introduce a more comprehensive aspect of measuring

program success, with a focus not only on academic outcomes, but also on 21st-century skills that are indispensable in the world of work and everyday life. Therefore, the diversity of evaluation models used shows progress in efforts to better understand and measure different aspects of education (Schlickmann & Bortoluzzi, 2023). One of the big challenges in bibliometric studies is the risk of keyword duplication, which can lead to incorrect conclusions. As found in this study, many articles contain mentions of duplicate keywords, which can lead to data misinterpretation and overemphasis on certain topics. This is reminiscent of the warning given by Afriyani et al., (2022)and Sahu et al., (2020) on the importance of choosing the right keywords and avoiding duplication in bibliometric analysis to ensure that the results obtained are valid and representative. This duplication not only affects thematic mapping, but can also interfere with the interpretation of emerging research trends.

Researchers who analyze a large number of articles may not be aware of the presence of keyword duplication, which can lead to incorrect conclusions because they are based on the wrong premise. Therefore, it is important to use analytics software that can detect and address these issues. Researchers also need to conduct a systematic review of each article to ensure that keyword analysis is conducted accurately and in accordance with the context of the research. In addition, a deeper understanding of the concepts associated with program evaluation can also help reduce the risk of errors in this analysis. The study also shows that there is a significant focus on gender in program evaluation. Keywords such as "female" and "male" indicate that many evaluation programs in secondary schools are related to issues of gender equality, educational opportunities, and gender-sensitive policies. This is consistent with the findings Binderkrantz & Bisgaard, (2024) and Weinkle et al., (2020), which emphasizes the importance of gender considerations in designing and evaluating educational programs. The presence of gender in this analysis suggests that the evaluation is not only focused on academic outcomes but also takes into account the social, psychological, and economic context of the students.

This increased focus on gender in the evaluation of educational programs shows a greater awareness of the role of gender in achieving equal educational outcomes for all students, regardless of their gender. As found by Sell et al., (2023), programs that focus on gender equality contribute greatly to improving the quality of education and supporting more inclusive social development. Therefore, it is important to continue to develop evaluation models that take into account gender factors so that education policies can reduce gender gaps and create a more equitable learning environment for all students. The analysis of collaborative networks in this study reveals the importance of international cooperation in program evaluation research. These findings confirm that authors from different countries, such as the United States and Canada, often collaborate on research on program evaluation. This shows that international collaboration plays an important role in enriching the quality of research and broadening perspectives in evaluating educational programs. Collaboration between authors from different countries can enrich insights into methodologies and best practices in program evaluation.

Research by Bindu et al., (2019) shows that international collaboration encourages the exchange of ideas and improves the quality of research. In this context, collaboration between authors from different countries can strengthen the validity and relevance of research results, as well as encourage the development of more comprehensive evaluation models. In addition, international collaborations also open up opportunities for data and resource sharing, which can accelerate the research process and expand the scope of findings that can be applied in various educational contexts around the world (Brito et al., 2023; Wu et al., 2016). The thematic analysis of the sample article shows several dominant topics in the evaluation of educational programs. Among them are the evaluation of mental health programs, gender-based interventions, and programs focused on adolescents. These findings are in line with research Kelloway et al., (2023), stating that students' mental health and well-being are important factors in determining the success of educational programs. The focus on adolescents also highlights the importance of considering psychological factors in designing and evaluating educational programs.

In addition, the growing interest in programs aimed at adolescents suggests that there is a need to adapt educational programs that are more sensitive to students' psychological and social development. This underscores the importance of designing programs that focus not only on the academic aspect, but also on the emotional and social well-being of students. Therefore, educational programs need to be developed with a more holistic approach to create a learning environment that supports the overall development of students (Gençoğlu, 2019). The findings of this study provide many insights that can be used for further research in the field of evaluation of educational programs. One of the key recommendations is the development of evaluation models that are more innovative and relevant to current educational challenges. Further research can delve deeper into the effectiveness of new evaluation models found in this study, such as EP_PISTdu and TADIPHE, and how these models can be applied in a variety of educational contexts.

In addition, further research can also investigate more deeply the influence of international collaboration in improving the quality and relevance of program evaluations. Collaboration between countries can be key in developing more universal and globally applicable evaluation practices. Researchers can leverage international platforms to share findings, thereby enriching the decision-making process related to education.

CONCLUSION

Based on the results of bibliometric analysis integrated with interpretive analysis, this study succeeded in exploring various program evaluation models applied in secondary schools. Key findings reveal that evaluation models, such as the Outcome-Mediation Cascade (OMC), TADIPHE, and EP_PISTdu, are increasingly relevant in addressing complex educational challenges. In addition, the focus on gender elements in program evaluation, with an emphasis on equality and justice for women and men, is also

an important highlight in this study. The dominance of new theory-based evaluation models and methods is also proven. International collaborations have been found to play an important role in enriching the quality of program evaluation research. Authors from different countries, such as the United States and Canada, actively collaborated, creating a network that reinforced the validity and relevance of the findings. Further thematic analysis showed that programs focused on mental health, adolescents, and gender-based interventions dominated the program evaluation literature in high schools. In addition, the study emphasizes the importance of avoiding keyword duplication in bibliometric analysis to prevent incorrect conclusions. In the future, the development of more innovative and relevant evaluation models needs to be encouraged, while continuing research on international collaboration that can improve the quality of evaluation of educational programs.

However, this study has some limitations. First, the data is only taken from one database, Scopus, so there may be relevant publications from other sources that are not covered. Second, the analysis only includes English-language articles, which may exclude important studies in other languages. In addition, data processing depends on the consistency of the keywords used by the author, which is prone to duplication or term bias. For further research, it is recommended that the data coverage be extended to several other scientific databases such as the Web of Science, ERIC, or Google Scholar, as well as take into account the literature of local languages. Follow-up research can also explore more deeply the effectiveness of each identified evaluation model, both through case study approaches and field experiments. Thus, the results of future research are expected not only to expand the scope of program evaluation studies, but also to deepen theoretical and practical understanding of the application of appropriate evaluation models in various educational contexts.

REFERENCES

Afriyani, A., Wahidah, I., & Wibowo, M. T. H. (2022). Bibliomatric Analysis: Development of Research Publication on "General Elections In Indonesia". *Social Impact Journal*, *Vol. 1*(No. 1).

Al-Alawi, R., & Alexander, G. L. (2020). Systematic review of program evaluation in baccalaureate nursing programs. In *Journal of Professional Nursing* (Vol. 36, Issue 4). https://doi.org/10.1016/j.profnurs.2019.12.003

Amahoroe, R. A. (2023). Evaluation of Classical Training at the Ambon Religious Training Center using The Kirkpatrick Evaluation Model. *12 Hydrau*, *9*(2). https://doi.org/10.47655/12waiheru.v9i2.166

Amouzou, A., Bryce, J., & Walker, N. (2022). Strengthening effectiveness evaluations to improve programs for women, children and adolescents. *Global Health Action*, *15*(sup1). https://doi.org/10.1080/16549716.2021.2006423

Asrul, Ananda, R., & Rosinta. (2014). Evaluation of Piracy. In *Ciptapustaka Media*.

Binderkrantz, A. S., & Bisgaard, M. (2024). A gender affinity effect: the role of gender in teaching evaluations at a Danish university. *Higher Education*, *87*(3).

https://doi.org/10.1007/s10734-023-01025-9

Bindu, N., Sankar, C. P., & Kumar, K. S. (2019). Research collaboration and knowledge sharing in e-governance: Temporal analysis of co-author network. *Transforming Government: People, Process and Policy*, *13*(1). https://doi.org/10.1108/TG-03-2018-0022

Blizniuk, A., Ueno, M., Furukawa, S., & Kawaguchi, Y. (2014). Evaluation of a Russian version of the oral health literacy instrument (OHLI). *BMC Oral Health*, *14*(1). https://doi.org/10.1186/1472-6831-14-141

Brito, A. C. M., Silva, F. N., & Amancio, D. R. (2023). Analyzing the influence of prolific collaborations on authors productivity and visibility. *Scientometrics*, *128*(4). https://doi.org/10.1007/s11192-023-04669-7

Browne, F., Hannigan, B., & Harden, J. (2021). A realist evaluation of a safe medication administration education programme. *Nurse Education Today*, *97*. https://doi.org/10.1016/j.nedt.2020.104685

Chanthalangsy, P., Yeh, B. Il, Choi, S. J., & Park, Y. C. (2024). Program evaluation of postgraduate obstetrics and gynecology training in Lao people's democratic republic - using the CIPP model. *BMC Medical Education*, *24*(1). https://doi.org/10.1186/s12909-023-04942-6

D'Inverno, G., Smet, M., & De Witte, K. (2021). Impact evaluation in a multi-input multi-output setting: Evidence on the effect of additional resources for schools. *European Journal of Operational Research*, *290*(3). https://doi.org/10.1016/j.ejor.2020.08.042

Dolor, R. J., Ruybalid, R. L., Uyeda, L., Edson, R. G., Phibbs, C., Vertrees, J. E., Shih, M. C., Jacobson, A. K., & Matchar, D. B. (2010). An evaluation of patient self-testing competency of prothrombin time for managing anticoagulation: pre-randomization results of VA Cooperative Study #481--The Home INR Study (THINRS). *Journal of Thrombosis and Thrombolysis*, *30*(3), 263–275. https://doi.org/10.1007/s11239-010-0499-8

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*. https://doi.org/10.1016/j.jbusres.2021.04.070

Ellegaard, O., & Wallin, J. A. (2015). The bibliometric analysis of scholarly production: How great is the impact? *Scientometrics*, *105*(3). https://doi.org/10.1007/s11192-015-1645-z

Fu, Y., Mao, Y., Jiang, S., Luo, S., Chen, X., & Xiao, W. (2023). A bibliometric analysis of systematic reviews and meta-analyses in ophthalmology. *Frontiers in Medicine*, *10*. https://doi.org/10.3389/fmed.2023.1135592

Gençoğlu, C. (2019). Remedial Education Programme: As A National Catch-Up And Remedial Education Model. *Egitim Mill,* *48*(1), 853–881. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85105112375&partnerID=40&md5=73f7f6c4d5aa6a6f7af4a8ed70af8f49

Guo, Y. M., Huang, Z. L., Guo, J., Li, H., Guo, X. R., & Nkeli, M. J. (2019). Bibliometric analysis on smart cities research. *Sustainability (Switzerland), 11*(13). https://doi.org/10.3390/su11133606

Guyadeen, D., & Seasons, M. (2018). Evaluation Theory and Practice: Comparing Program Evaluation and Evaluation in Planning. In *Journal of Planning, Education and Research* (Vol. 38, Issue 1). https://doi.org/10.1177/0739456X16675930

Hidayah, N., & Romelah, R. (2022). Evaluation of Learning at Aisyiyah Kindergarten Bustanul Athfal II, Berau Regency. *Journal of Obsession: Journal of Early Childhood Education*, *6*(5). https://doi.org/10.31004/obsesi.v6i5.2770

January, N. (2024). Systematic Literature Review Using the Prism Method: The Impact of Blockchain Technology on Digital Advertising. *M-Progress Scientific Journal*, *14*(1), 1–11. https://doi.org/10.35968/m-pu.v14i1.1182

Jáuregui, A., Pacheco-Miranda, S., Argumedo-García, G., Marrón-Ponce, J. A., G.-Olvera, A., Vargas-Meza, J., Ayvar-Gama, Y. Y., Velázquez, D., David Quezada, A., & Bonvecchio-Arenas, A. (2022). Comprehensive evaluation of Salud Escolar a health school program in Mexico: Rationale, design and methods. *Preventive Medicine Reports*, *25*. https://doi.org/10.1016/j.pmedr.2021.101662

Kania, N., Rais, R., Manoppo, Y., Yusron, A., Zega, Y., Nurbaya, Nuraeni, Z., Ahmad, Munif, A., Sulfiati, Y., Anas, Suwenti, R., Maghfirah, G., Listiani, H., Dedi, & Mursidin. (2023). Educational Evaluation (A Critical Review). In *Edupedia Publisher*.

Kelloway, E. K., Dimoff, J. K., & Gilbert, S. (2023). Mental Health in the Workplace. In *Annual Review of Organizational Psychology and Organizational Behavior* (Vol. 10). https://doi.org/10.1146/annurev-orgpsych-120920-050527

Khan, M. A. (2022). ESG disclosure and Firm performance: A bibliometric and meta analysis. In *Research in International Business and Finance* (Vol. 61). https://doi.org/10.1016/j.ribaf.2022.101668

Liu, W., Xu, C., Peng, Y., & Xu, X. (2023). Evolution of Tourism Risk Communication: A Bibliometric Analysis and Meta-Analysis of the Antecedents of Communicating Risk to Tourists. *Sustainability (Switzerland), 15*(12). https://doi.org/10.3390/su15129693

Mustofa, I., Kuswara, H., Rosmawati, R., & Rosmika, E. (2023). Evaluation of Computer Training Programs Using Kirkpatrick Models to Increase the Competency Capacity of Drinking Teachers. *Nidhomul Haq : Journal of Islamic Education Management*, *8*(2). https://doi.org/10.31538/ndh.v8i2.3778

Nouraey, P., Al-Badi, A., Riasati, M. J., & Martha, R. L. (2020). Educational Program and Curriculum Evaluation Models: A Mini Systematic Review of the Recent Trends. In *Universal Journal of Educational Research* (Vol. 8, Issue 9). https://doi.org/10.13189/ujer.2020.080930

Park, S.-E., & Tang, L. (2019). How colour and visual complexity affect the evaluation of skin cancer infographics: an experiment study. *Journal of Visual Communication in Medicine*, *42*(2), 52–65. https://doi.org/10.1080/17453054.2019.1573633

Saeed, S., Chand, P., Sulaiman, A., Nisar, I., Humayun, K. N., Malik, M. G. R., & Jehan, F.

(2023). Process evaluation of paediatric fellowship training programs at a University Hospital in Pakistan. *BMC Medical Education*, *23*(1). https://doi.org/10.1186/s12909-023-04501-z

Sahu, M. K., G. K., C., & Panda, S. P. (2020). Research Trends on Coronavirus (Covid-19): A Bibliomatric Analysis between 2010–March, 2020. *Library Philosophy and Practice*, *2020*.

Schlickmann, A., & Bortoluzzi, S. C. (2023). Environmental Education Performance Evaluation in a Higher Education Institution. *Systemic Practice and Action Research*, *36*(6). https://doi.org/10.1007/s11213-023-09636-0

Sell, K., Oliver, K., & Meiksin, R. (2023). Comprehensive Sex Education Addressing Gender and Power: A Systematic Review to Investigate Implementation and Mechanisms of Impact. *Sexuality Research and Social Policy*, *20*(1). https://doi.org/10.1007/s13178-021-00674-8

Tang, J., & Hai, L. (2021). Construction and Exploration of an Intelligent Evaluation System for Educational APP through Artificial Intelligence Technology. *International Journal of Emerging Technologies in Learning*, *16*(5). https://doi.org/10.3991/ijet.v16i05.20293

Tigre, F. B., Curado, C., & Henriques, P. L. (2023). Digital Leadership: A Bibliometric Analysis. *Journal of Leadership and Organizational Studies*, *30*(1). https://doi.org/10.1177/15480518221123132

Vaiopoulou, J., Papadakis, S., Sifaki, E., Kalogiannakis, M., & Stamovlasis, D. (2023). Classification and evaluation of educational apps for early childhood: Security matters. *Education and Information Technologies*, *28*(3). https://doi.org/10.1007/s10639-022-11289-w

Vieira, E. A. O., Silveira, A. C. da, & Martins, R. X. (2019). Heuristic evaluation on usability of educational games: A systematic review. *Informatics in Education*, *18*(2). https://doi.org/10.15388/infedu.2019.20

Watt, M. (2017). An evaluation of the Australian Government's Students First agenda for strengthening the curriculum. *Curriculum Perspectives*, *37*(2), 161–170. https://doi.org/10.1007/s41297-017-0008-8

Weinkle, L. J., Stratford, J. M., & Lee, L. M. J. (2020). Voice in Digital Education: The Impact of Instructor's Perceived Age and Gender on Student Learning and Evaluation. *Anatomical Sciences Education*, *13*(1). https://doi.org/10.1002/ase.1865

Wu, Y., Venkatramanan, S., & Chiu, D. M. (2016). Research collaboration and topic trends in Computer Science based on top active authors. *PeerJ Computer Science*, *2016*(1). https://doi.org/10.7717/peerj-cs.41

Yangari, M., & Inga, E. (2021). Article educational innovation in the evaluation processes within the flipped and blended learning models. *Education Sciences*, *11*(9). https://doi.org/10.3390/educsci11090487

Yasin, A. M. Al, Shehah, A. Al, Cherian, S., & Aitchison, M. E. (2023). Program Evaluation: Nursing Informatics Structured Training Program. *Saudi Journal of Nursing and Health Care*, *6*(07). https://doi.org/10.36348/sjnhc.2023.v06i07.002