momentum
ACCELERATING EQUITY IN COMPUTING AND TECHNOLOGY

UNDERGRADUATE BPC LITERATURE DATABASE
Technical Report

Recommended citation for technical report: Newhouse, K. N. S., Karpicz, J., Gutzwa, J. A. Lehman, K. J., Stout, J. G., & Nhien, C. (2021, June). *Technical methods report for the Undergraduate BPC Literature Database: Developing a database of contemporary research on broadening participation in computing*.

For decades, there has been growing interest in diversifying the field of computing, both as it pertains to higher education and the workforce. Not only are women and Students of Color (specifically, African American/Black, Hispanic/Latinx, and Indigenous students) minoritized in computer science (CS) departments in higher education, but the percentage of women and racially minoritized communities in computing occupations has steadily decreased for most of the last quarter century (Corbett & Hill, 2015; National Science Foundation [NSF], 2015). As a result, there has been a great deal of recent focus on programs aimed at increasing the number of women and Black, Latinx, and Indigenous (BLI) students in the CS pipeline. Within higher education, there have been several programs focused on broadening the participation of women and BLI students in undergraduate CS departments, such as Extension Services through the National Center for Women and Information Technology (NCWIT, 2014) and AnitaB.org's Building, Recruiting, and Inclusion for Diversity (BRAID) initiative (AnitaB.org, 2014). Recognizing the need for CS departments to expand diversity efforts, the National Science Foundation's Directorate for Computer and Information Science and Engineering (CISE) recently implemented a requirement that grant proposals include "meaningful BPC plans," as a piece of each project's proposal (Kurose, 2017).

Despite this growth in BPC efforts nationally, there has not been a central place where stakeholders could access recent research on best practices around recruiting and retaining women and people of color in undergraduate computing. The Undergraduate BPC Literature Database aims to fill that gap. As a public resource, the matrix was designed to support practitioners, researchers, non-profits, industry, and others interested in broadening participation in computing (BPC) with identifying and assessing relevant, peer-reviewed scholarship that can inform future BPC work.

This report details the process by which the Undergraduate BPC Literature Database was developed. Note that a first version of the database was created during the 2020-2021 academic year. The initial database was created in 2020 and updated in 2022. Hence, this report references two iterations of the database and describes the methodology for the creation of the original database as well as the 2022 update.

Methods

The database was developed by a team of education researchers at the University of California, Los Angeles who are a part of *Momentum*, a mixed-methods research hub that examines efforts to diversify computing and technology fields. There were four key phases of this process, each described further below.

Phase 0: Refine

During the refine phase, the literature database team worked in consultation with the UCLA Library and other scholars in the computing education field to identify relevant parameters and best practices for conducting the systemic review. We narrowed the scope of the research to be included in the matrix in the following ways:

Parameter	**Definition**
Type of material	Refereed and/or peer-reviewed content (e.g., conference proceedings, journal articles). NO: books/monographs, ACM panels, white papers, popular press. Literature may be empirical (analyses of primary or secondary quantitative or qualitative data) or conceptual/theoretical in nature. It must into at least one of the following three categories: 1) show the impact of specific intervention(s) seeking to broaden participation in computing, 2) seek to identify and understand factors that contribute to inequitable outcomes in undergraduate computing among historically minoritized populations 3) advance theory or frameworks that inform the ways researchers think about broadening participation of historically minoritized students in computing.
Publication Years	January 2005 – August 2020
Geographic boundary	United States
Academic level	Undergraduate
Disciplinary focus	Computing, either (1) broadly defined or (2) focusing on a specific computing subfield (e.g., data science, computer engineering, computer science). Literature addressing broadening participation in STEM in general is NOT included.
Student population(s)	In alignment with NSF definitions, materials must address broadening participation in computing (recruitment, retention, completion of computing courses and majors and/or experiences in or with

	computing activities aimed at broadening participation in computing) of historically minoritized populations in computing such as: "women, minorities (African Americans/Blacks, Hispanic Americans, American Indians, Alaska Natives, Native Hawaiians, Native Pacific Islanders, and persons from economically disadvantaged backgrounds), and persons with disabilities." Articles addressing multiple identities of interest (e.g., Black women, low-income Students of Color in computing) are especially welcome. The participation of other historically minoritized and understudied subgroups in computing may be included (e.g., trans*/genderqueer students, LGBQ+ students, first-generation college students, etc.) should they appear in our searches.
Beyond the above criteria, articles will be reviewed for inclusion based on:	1) Their framing of the study around the problem of broadening participation in computing for historically minoritized populations per the definition above. 2) Descriptions of the demographic characteristics of study participants 3) Descriptions of the kinds of student engagement in computing field (intro course takers, majors, minors, undecided, mentorship, etc.) 4) Development or use of a theoretical framework to frame/address the problem of broadening participation 5) Demonstration of results with implications for research, practice, and policy about broadening participation in computing

Once the conceptual parameters of the literature database were established, we then worked with UCLA librarians to determine the proper search terms and processes we would need to execute in order to yield research that met the above parameters. Search terms used Boolean operators such that phrases in quotes must appear together; asterisks would return any word that begins with the root/stem of the word truncated by the asterisk (e.g., universit* returns items with the term "universities" and "university"). Terms in parentheses allowed the search engine to capture OR statements so the search returns all articles with one or more of those terms. We used the same set of search terms in every database we searched to maintain consistency across databases and results. The search terms we used are below:

("Computer Science education" OR "computing education")

AND

(undergraduat* or college* or universit* or "higher education")

AND

(minorit* or underrep* or divers* or ethnic* or gender* or female* or wom* or "african american*" or black* or Hispanic* or latino* or latina* or latinx* or "Disproportionate Representation" or "American Indian*" or "native american*" or "Alask*" or "Native Hawaiian*" or "Native Pacific Islander*" or disadvantage* or poverty or "low income" or disab* or trans* or lesbian* or gay or bisex* or LGB* or "first-generation")

AND

(increas* or barrier* or particip* or challeng* or broaden* or sucess* or interven*or retain* or retent* or recruit* or enroll* or access* or equit* or persist* or pedagogy or teach* or "culturally relevant" or "culturally responsive" or curriucul* or climate or major* or minor*)

AND

Peer Reviewed

AND

2005 or later

After establishing search terms, we selected key databases that would ensure we were yielding as many relevant articles with as little duplication as possible. We ultimately narrowed the list to search 12 databases that represented both multidisciplinary databases (e.g., Academic Search Complete) and disciplinary specific databases (e.g., ACM Digital Library). Certain databases, such as IEEE Explore and JSTOR, were considered but ultimately excluded because they did not have the capability to accommodate our lengthy and specific block of search terms. The databases included were:

- Education Source
- ERIC
- Academic Search Complete
- Chicano Database
- Women's Studies International
- LGBTQ Life
- Teacher Reference Center
- PsychInfo
- Sociological Abstracts

- Gender Watch
- Web of Science
- ACM Digital Library

Phase 1: Search

Once the search terms were refined and the databases selected, we executed three days of searches in August of 2020. All told, before any screening we retrieved 3729 articles. The dates, 12 databases searched, and the number of hits returned from each are shown in the table below. Bibliographic metadata and abstracts for each article were scraped and stored using Zotero, a free citation management software.

Date	**Database**	**Number of Articles Returned**
August 11, 2020	Education Source	296
	ERIC	573
	Academic Search Complete	367
	Chicano Database	0
	Women's Studies International	2
	LGBTQ Life	1
	Teacher reference Center	95
August 25, 2020	PsychInfo	640
	Sociological Abstracts	108
	Gender Watch	39
	Web of Science	157
August 27, 2020	ACM Digital Library	1451

Phase 2: Review

With the 3729 articles in Zotero, we conducted a manual review of the articles to remove any duplicates or other articles that did not met our parameters for screening. After a first pass, 626 articles were duplicates and removed, 53 articles were removed because they had no author or abstract, 11 articles were removed because they were not published in English, and 6 entries were removed because they were books or book reviews. This brought the final number of articles to be screened to 3033.

Next, to determine whether these 3033 articles met our pre-established parameters for inclusion in the literature database we used Abstrackr to screen each article. Abstrackr, a free online machine learning tool housed at Brown University's Center for Evidence Synthesis in Health, was developed for researchers in the health sciences conducting systematic reviews (Wallace et al., 2012). The software allows users to easily upload the bibliographic metadata and abstracts for each article and then provides a simple interface where multiple users can be assigned abstracts to screen. When screening, Abstrackr allows team members to assign each article one of three statuses: accept the article for inclusion; reject the article and indicate why; indicate they are unsure whether an article should be included or not. We elected to have each article be reviewed by two different team members to ensure reliability across reviewers. Thus, if both reviewers of any one article agreed that that article met the parameters we established, then the article was included; if both agreed it should not be included, it was rejected; if one or both reviewers were unsure or if they disagreed whether it should be included, it went to team leadership for final review. The article screening process took place between August 28, 2020 and September 26, 2020. Throughout this time, we met bi-weekly as a team to discuss any challenges and observations about the screening process to ensure all reviewers were on the same page about the parameters and what kinds of articles should be included.

After screening all articles, we arrived at 192 that met our parameters. Articles were most often excluded because they were about K-12 computing education, lacked a substantive focus on the populations of interest for broadening participation in computing, were studies conducted outside the United States, were about STEM more generally, or were well beyond the scope and completely unrelated to computing education at all. Once we had the list of 192, we then compiled PDF versions of each article in a folder for natural language processing analysis and saved each article as its number of record (1-192).

Phase 3: Extraction

While we were screening articles, we also worked with our data scientist consultant, Dr. Jane Stout, to determine the kinds of categories and things she might develop code to search for and extract from each article. As we already had all of the bibliographic data, we prioritized the following categories:

- Article rigor, indicated by whether the article was published in a peer reviewed journal
- The historically minoritized group(s) (i.e., BPC group) that are the focus of the paper as indicated by substantive attention in the framing, findings, and discussion of the paper (e.g., Women, LGBTQ+ students)
- The type of participant population that the study focused on/gathered data from or about (e.g., Faculty, Instructors, TAs, Undergraduate students)
- The specific methodological approach(es) used in the research design (e.g., survey, experimental), including whether the data were gathered at one or more than one institution (e.g., Multiple institutions, single institution)
- Analytic approach(es) used in the research design (e.g., logistic regression, ANOVA)
- The classification of the institution(s)involved in the study (e.g., community college, Tribal College, HBCU)

Once we established that the code would aim to identify these categories, Dr. Stout developed a process to extract this information from each PDF. Generally, PDF files were converted to text files. The content in each text file was appended to its corresponding record in an Excel sheet. The resulting .xlsx file was converted to a json file, which was read by a customized Python script that processed the title, abstract, item type, and full text for each record; the script then produced a data file with fields containing binary indicators of article categories we requested (e.g., whether the article focused on Latinx students; whether the data were analyzed via an ANOVA). One article's PDF was a photocopy and therefore could not be converted to a text file and was dropped from analysis.

Once we had an initial data file of information extracted by the Python script, we selected 40 articles (20%) in the data file to manually review. Individual team members were assigned to read an article and ensure that the code correctly identified each parameter. During this review, we identified specific categories that were less accurate and reported back to Dr. Stout to update her code; we also determined several categories (e.g., whether a sample was multi-institutional) that were too nuanced for the code to correctly extract and thus that needed to be checked by a team member. At this stage, we identified 9 articles that did not meet our parameters once we read the full paper. After two rounds of reviewing the articles in the datafile, we arrived at 182 articles for inclusion.

One final round of review was carried out by two graduate members of the literature database team in July 2021. These checks were carried out to review categories previously determined to be too nuanced for the code to correctly extract, so the team manually reviewed each of the prior 182 included articles to confirm eligibility for inclusion and correct reporting. This final phase of review identified 3 articles that did not meet parameters for inclusion. As such, the final number of articles included in the first iteration of the Undergraduate BPC Literature Database was 179.

Literature Database Update (August 2020 – March 2022)

In the spring of 2022, *Momentum* conducted an update to the Undergraduate BPC Literature Database to incorporate relevant articles that were published after the cutoff period for the initial literature database. The search for recently published articles to contribute to the database was conducted for articles published between August 2020 and March 2022. The same search parameters and processes above were carried out for this literature update given that the procedure above was followed identically with the exception of the search dates. As such, the following sections provide condensed overviews on the update process relative to numbers and dates that differ from the initial descriptions above for the initial search (January 2005 to August 2020) that established the research process.

Phase 1: Search

Between April 7, 2022 and May 6, 2022, the team conducted a search for relevant and recently published literature (between August 2020 and March 2022). This search of the 12 databases yielded 748 items prior to exclusions of any kind.

Date	Database	Number of Articles Returned
April 25, 2022	Education Source	56
April 25, 2022	ERIC	67
April 25, 2022	Academic Search Complete	44
April 25, 2022	Chicano Database	0
May 06, 2022	Women's Studies International	1
April 25, 2022	LGBTQ Life	15
April 21, 2022	Teacher reference Center	10
April 21, 2022	PsychInfo	7
April 21, 2022	Sociological Abstracts	5
May 06, 2022	Gender Watch	5

May 03, 2022	Web of Science	146
April 07, 2022	ACM Digital Library	392

Phase 2: Review

The 748 returned items were assessed within Zotero for duplicates, resulting in 195 duplicates. After removing duplicates, the sample of articles was reduced to 553 items. Then the article titles and abstracts for each of those 553 items were manually screened in Abstrackr to conclude that 63 articles met our parameters. Given that these articles constitute additions to an existing data base of articles with identifiers 1-192, they were given the article identifiers of 193-255.

Phase 3: Extraction

To execute extraction, we again worked with Dr. Jane Stout to help with excluding articles that did not actually meet our parameters despite being previously included based on titles and abstracts alone. In addition to Dr. Stout's coded analysis as well as manual article checks from one of our *Momentum* team members, there were an additional 10 articles excluded for not meeting the identified parameters; thus, we added a total of 53 new articles in this most recent phase. This addition of 53 to the initial 179 articles brings the total number of articles in the Undergraduate BPC Literature Database to 232.

Appendix A

Figure 1: Graphic representation of article review process for initial review (January 2005 and August 2020)

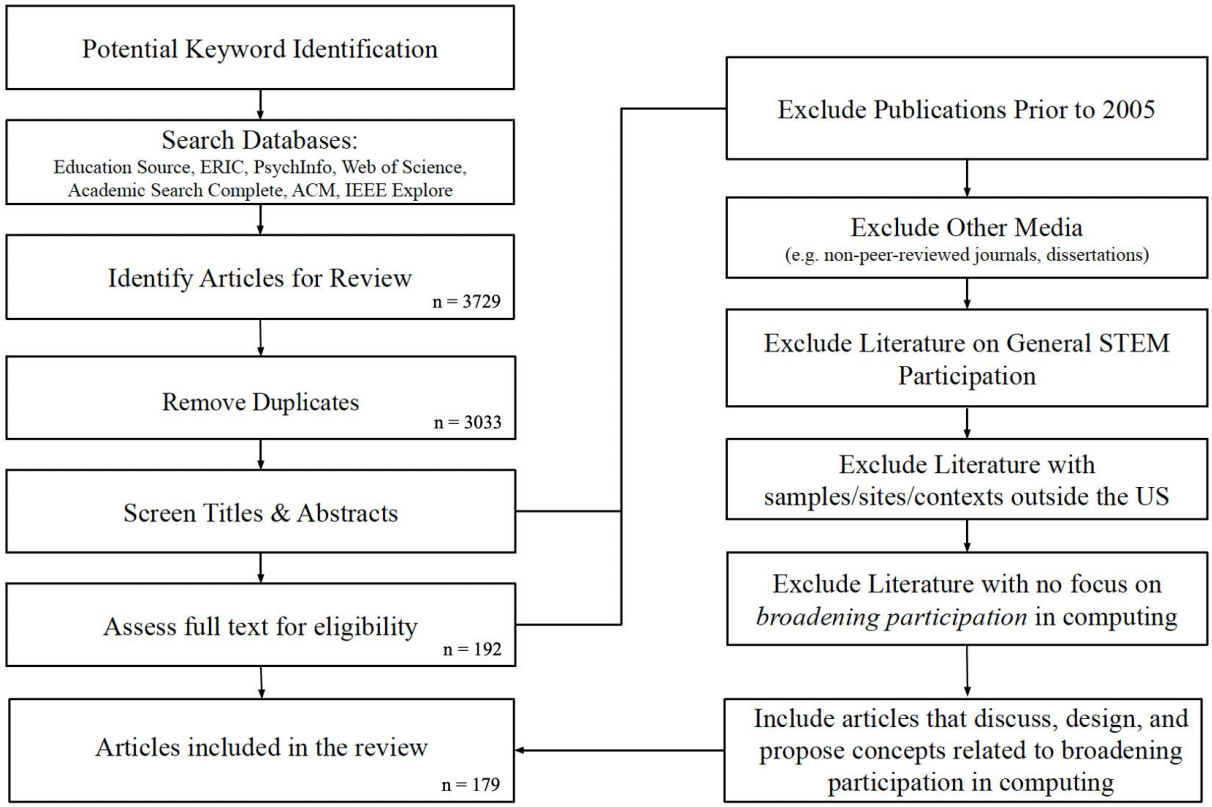
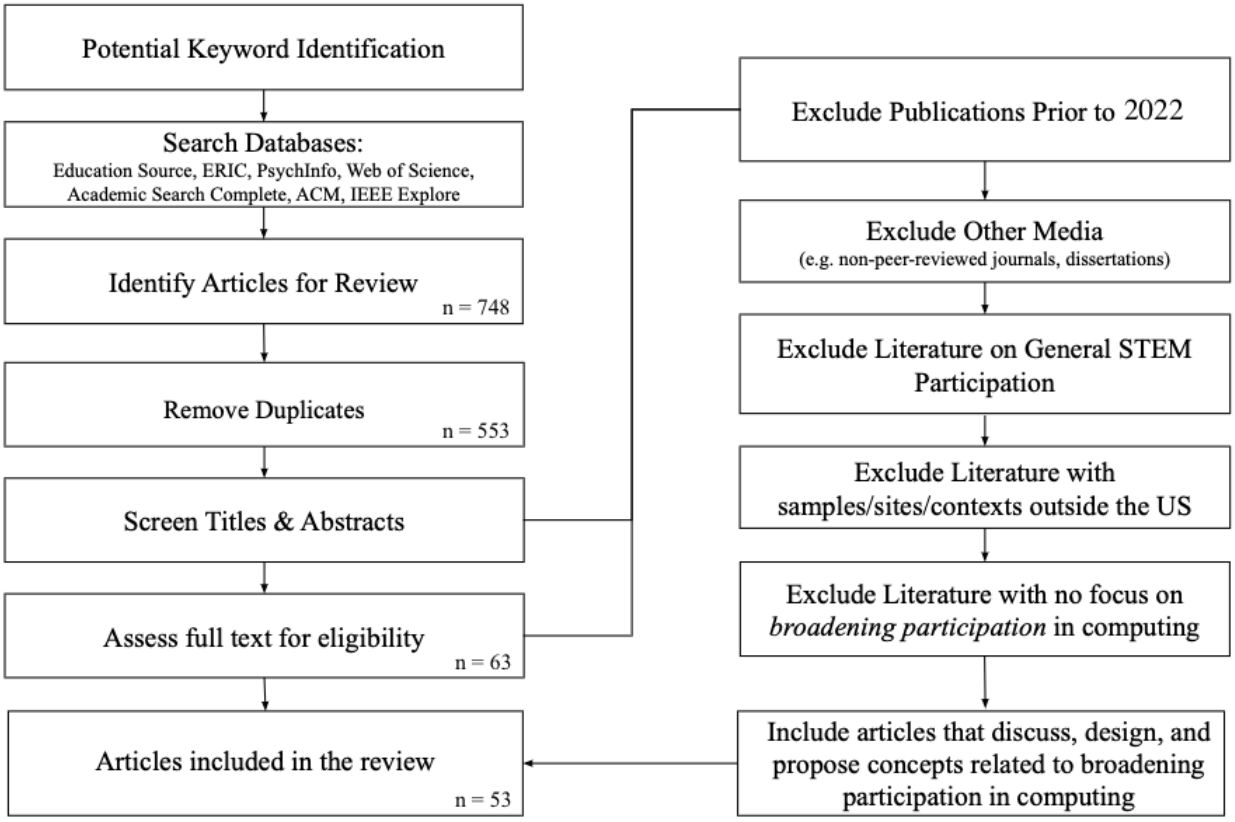


Figure 2: Graphic representation of article review process for literature update (August 2020 and March 2022)



References

AnitaB.org, (2014). BRAID: A Diversity Program. *AnitaB.org.* Retrieved from http://anitaborg.org/braid-building-recruiting-and-inclusion-for-diversity/.

Corbett, C., & Hill, C. (2015). *Solving the Equation: The Variables for Women's Success in Engineering and Computing*. American Association of University Women. 1111 Sixteenth Street NW, Washington, DC 20036.

Kurose, J. (2017). Dear Colleague Letter: Pursuing Meaningful Actions in Support of Broadening Participation in Computing (BPC). *National Science Foundation (July 3, 2017).* Retrieved from https://www.nsf.gov/pubs/2017/nsf17110/nsf17110.pdf.

National Center for Women and Information Technology, (2014). Why Join: NCWIT Programs. *NCWIT.* Retrieved from https://www.ncwit.org/why-join-ncwit-programs.

National Science Foundation (2015). National Survey of College graduates. *National Center for Science and Engineering Statistics.*Wallace, B. C., Small, K., Brodley, C. E., Lau, J, and Trikalinos, T. A. (2012). Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. In *Proceedngs of the ACM International Health Informatics Symposium (IHI)*, p.819--824.

Williams, T. (2020, June 19). 'Underrepresented Minority' Considered Harmful, Racist Language. Retrieved from https://cacm.acm.org/blogs/blog-cacm/245710-underrepresented-minority-considered-harmful-racist-language/fulltext.