



Measured at the Door: Durability, Measurement Practice and the Limits of Evidence on Women's Personality Development Programmes

Swati Saini¹, Dr. Hari Chand²

¹Research Scholar, ²Associate Professor

¹⁻²Department of Education, Sunrise University

Abstract

Women's personality development programmes, life skills education and soft skills training have attracted repeated evaluation, and the verdict is reasonably consistent: they work. This paper puts an earlier question first. What is such a literature able to establish, given the way it measures? Its material is a systematic review of 42 studies, chosen under a protocol fixed in advance and reported to PRISMA 2020, and the evidence base here becomes an object of study rather than a quarry for effect estimates. Three findings emerge. The first concerns how narrowly and how unevenly measurement is spread. There are 41 separate outcome instruments for 42 studies; the narrowest constructs are the ones the two well-validated scales address; and reliability evidence hardly ever accompanies the author-built indices on which the field's central claims rest. Second, the timing is wrong. Nothing was measured after the programme closed in 32 of the 42 studies, which puts the central practical question about these interventions beyond the reach of the published record. Where a second measurement was taken, in ten studies, every dimension shows a smaller effect. That is consistent with attenuation, yet it is equally confounded with design quality, and here the two cannot be prised apart. Third, a bibliometric map of the same corpus catches the field's vocabulary shifting across two decades, away from collective and economic terms and toward individual and psychological ones, so that what critical scholarship has described qualitatively acquires a quantitative trace.

Keywords: evidence quality; measurement practice; durability of effect; women's empowerment; life skills education; bibliometric analysis; reporting standards

1. Introduction

Programmes aiming to build women's confidence, communication and assertiveness attract frequent evaluation, and the verdicts tend to be kind. Neighbouring literatures show the same thing twice over: psychosocial outcomes usually improve, and the methods used to show it are usually weak (Nasheeda et al., 2019). What follows is a call for better studies, after which the matter is dropped.

This paper goes another way. The evidence base becomes its object rather than its quarry, and the question put to it is what it can establish. Why the distinction matters is that the answers diverge. Do these programmes work? That asks about interventions. Can we tell? That asks about measurement, and an answer is available now, out of the published record, without anyone waiting for better studies.

Three features come under examination. First, what is measured, and on what instruments. Second, the timing of measurement, which proves to constrain everything the field is able to



claim. Third, two decades of movement in the field's conceptual vocabulary, which a bibliometric map of the same corpus brings into view. Convention treats the three as unrelated complaints. They are connected, and the connection is this: a field that attends to the participant and not to her circumstances, and attends at the moment she walks out, will generate precisely the drift of vocabulary that the map records.

At the end comes a minimum reporting standard, kept short on purpose. Describing an ideal study is not the aim. Naming the two or three changes that would render the existing stock of studies interpretable is.

2. Method

A systematic review underlies the analysis. Its protocol was prepared before any searching began, and it is reported against the PRISMA 2020 statement (Page et al., 2021). Searching covered eight electronic databases, together with institutional repositories, six core journals hand-searched, and citation searching in both directions. Using several databases instead of one was deliberate. Adequate recall in the social sciences normally takes four or more sources in combination, and which combination is optimal shifts with the topic (Bramer et al., 2017). Here that mattered: about a third of the unique records was the most any single database yielded.

Two reviewers worked independently at the screening stage. Of 2,010 records retrieved, 1,483 survived de-duplication to be screened and 196 were read in full; eligibility and quality appraisal then left 42 studies, carried in 47 publications, for the synthesis.

Material for the present paper comes from three components of that review, each described below at its point of use. For quantitative and mixed designs, quality was appraised against the fourteen-item checklist of Kmet et al. (2004); qualitative studies went to its ten-item companion, with each score becoming a proportion of the maximum applicable and inclusion requiring 0.55. Separate extraction fields recorded, study by study, the instruments, the timing of measurement, who administered it, and whether reliability evidence was present. Co-occurrence analysis then took the author keywords and indexed terms, run through the association strength normalisation and clustering that VOSviewer implements (van Eck & Waltman, 2010), with the procedural guidance of Donthu et al. (2021) followed throughout.

A boundary belongs here at the start. Describing a body of literature is what a bibliometric mapping does. Findings it does not synthesise. About the effects of the interventions it establishes nothing. Munn et al. (2018) note how easily that line goes blurred, mapping results being reported as if they bore evidential weight. Mapping happens in Section 3 and appraisal in Sections 4 to 6; neither yields any claim about programme effectiveness.

3. The Shape of the Evidence Base

3.1 Growth, geography and design

The included studies fall across 15 countries and the years 2006 to 2025. India alone accounts for 17 of the 42 and South Asia for 21, while sub-Saharan Africa accounts for 7 and high-income Western settings for 5. Students predominate among the participants, cities among the sites. Need is not what either pattern tracks. Both track instead the ready availability of

assembled participants and administratively accessible outcomes, which tells us about the economics of evaluation rather than about where need lies.

Judged against what the appraisal instrument asks, the designs are weak. Dominance belongs to quasi-experimental and pre-experimental work. Randomised trials number only 2 of the 42. Evaluation research into educational and social interventions routinely looks this way, randomisation often being impracticable there. The cost to inference is real all the same. What a one-group pre-post design shows is that participants changed, not that the programme changed them.

Publication across the twenty-year window clusters rather than spreads. The first decade holds roughly a third of the included studies and the second roughly two thirds. Growth of that order partly follows the general swelling of social science publication and is no evidence of a turn toward this topic. More can be learned from what the growth consists of. Comparison groups and validated instruments turn up somewhat more often in the later studies, which also come more often from outside the handful of countries dominating the earlier work, so growth and improvement have both occurred, unevenly, and how far is quantified in Section 6.

3.2 Disciplinary fragmentation

Six disciplinary fields appear, the majority of the work coming from psychology and education. Co-occurrence analysis exposes how that fragmentation is structured. Three occurrences was the minimum threshold, forty terms cleared it, and four clusters emerged, which Table 1 sets out.

Table 1. Composition of the keyword clusters

Cluster	Principal terms	Reading
1. Empowerment and economic participation	women empowerment; gender equality; India; self-help groups; microfinance; entrepreneurship; rural women; agency; autonomy	The development-studies core, oriented to collective and economic routes to empowerment
2. Psychological constructs	self-efficacy; self-esteem; psychological empowerment; assertiveness; emotional intelligence; self-confidence; motivation; resilience; intervention study; quasi-experimental	The measurement core. Design terms sit inside this cluster, which locates experimental evaluation in psychological outcomes rather than across the field
3. Programme content and delivery	personality development; life skills education; soft skills; communication skills; training programme; skill development; capacity building; adolescent girls; employability	Four vocabularies for substantially one class of intervention, dense enough internally to justify searching them as one construct

Cluster	Principal terms	Reading
4. Leadership and workplace	leadership development; women leadership; decision making; career advancement; organisational commitment; mentoring	The management literature, largely disconnected from Cluster 1 despite addressing overlapping outcomes

Note. Mapping produced in VOSviewer using association strength normalisation (van Eck & Waltman, 2010).

Two things about the map bear on the argument. Although four separate vocabularies generate the intervention terms, those terms gather in one dense cluster, confirming that the labels record disciplinary provenance and not distinct interventions. The second thing weighs more. At opposite sides of the map sit Clusters 1 and 2, linked directly only rarely. Internal change and economic outcomes are measured by what amount to two separate literatures, written in different disciplines, with few studies bridging them, which is the structural counterpart to the concentration of measurement documented in Section 4.

3.3 The vocabulary shift

Lay the average publication year of each term's records over the same map and the field's language is seen to move. Microfinance, self-help groups and rural women sit among the terms tied to collective and financial routes to empowerment, and their average years are early. Soft skills, employability, resilience, wellbeing and mentoring sit among the terms tied to individual capability and its measurement, and their average years are late. Collective to individual, economic to psychological: that is the direction of travel.

Empowerment, Batliwala (2007) argued on conceptual grounds, had been steadily emptied of the collective and political content it was coined to carry, until what remained was an individual attribute installable on commission. Her qualitative account finds a quantitative counterpart in the overlay, and convergence between two wholly unlike kinds of evidence is worth a pause, since it establishes the drift as a measurable property of the corpus and not as an impression held by critics. A neighbouring conclusion comes from Cornwall (2016), whose review of what works in women's empowerment programming finds success where interventions alter relationships and structures and not dispositions alone. Earlier still, Mosedale (2005) proposed assessing empowerment relationally rather than as an attribute of individuals, on the ground that what alters is a woman's position with respect to others and not the woman by herself.

4. Measurement Practice

4.1 What is measured, and with what

Across the 42 studies there are 41 distinct outcome instruments. More than two uses is rare. Table 2 lists the ones reaching three, naming the construct each addresses and recording whether reliability evidence came with it.

Table 2. Outcome instruments used in three or more studies

Instrument	Construct measured	Studies	Reliability evidence reported
Generalised Self-Efficacy Scale (Schwarzer & Jerusalem, 1995)	Self-efficacy	9	8 of 9
Rosenberg Self-Esteem Scale (Rosenberg, 1965)	Self-esteem	8	7 of 8
Author-constructed empowerment index	Composite empowerment	7	2 of 7
Author-constructed confidence questionnaire	Self-confidence	6	1 of 6
Psychological Capital Questionnaire (Luthans et al., 2007)	Psychological capital	4	4 of 4
Emotional intelligence inventory (various)	Emotional intelligence	4	3 of 4
Semi-structured interview schedule	Qualitative outcomes	9	Not applicable
Employment or enrolment record	Objective attainment	5	Not applicable
All instruments used once or twice	Various	32	6 of 32

Note. Counts exceed 42 because most studies used more than one instrument.

That final column holds the first of this paper's findings, and its direction is the wrong one. Much of the quantitative measurement rests on two well-validated scales, which almost always come with reliability evidence attached, and facing them are empowerment indices and confidence questionnaires written by the authors themselves, drawn on nearly as often and validated almost never. The consequence turns on which constructs sit where. Narrow, well defined and well instrumented: that describes self-efficacy and self-esteem. Composite empowerment and general confidence, which is where the field's central claims actually rest, are left to the instruments of weakest psychometric standing. The outcomes measured best are the narrowest ones.

A second consequence runs quieter. Gains reported on unvalidated instruments come out systematically larger than those reported on validated ones. About programmes that tells us nothing. It is a warning about instruments. The weaker measures will pull upward any synthesis that pools the two together.



Proliferation carries a further consequence for synthesis, usually filed as a technical nuisance when it should be read as a finding. With 41 instruments spread over 42 studies, nearly every study is measuring something slightly different. Compute a standardised difference on the Rosenberg scale and another on a confidence questionnaire the authors wrote, and the two are numerically comparable while being substantively unlike. Pooled, they produce a figure whose referent will not hold still. So the heterogeneity that blocks meta-analysis in this field belongs to the measurement, which varies, and not to the interventions, which are reasonably similar.

4.2 Who administers the measure

Administration went into a field of its own, because interpretation turns on it quite apart from the instrument. For rather more than a third of the included studies, whoever delivered the training also handed out the outcome measure. Self-report gathered that way opens the door to acquiescence and social desirability, and later analysis cannot shut it.

Direction survives the qualification: positive effects show up where assessors were independent as readily as where they were not. Magnitude very likely does not. Treat any mean reported out of a corpus composed this way as an upper bound rather than an estimate, and note that reviews pooling magnitudes without flagging administration report something other than what they seem to (Pigott & Polanin, 2020).

4.3 What is never measured

In an extraction sheet the empty column frequently tells the most. Not one quantitative study among all 42 measured what surrounded the participant: her household's response to her changed behaviour, the openness of institutions to her, whether support continued once the programme ended, whether the labour market she was being prepared for was within reach. All of this surfaces in the qualitative studies, and surfaces there with striking consistency, yet nobody turned it into an instrument.

The gap is not marginal. Programme content barely differs between the two sets of qualitative accounts, and what those accounts single out as separating change that lasted from change that faded is exactly these conditions. Measuring the participant exhaustively while leaving her circumstances alone amounts to a decision about where the causal action is expected to sit, and that decision was never argued for; it was simply inherited. The demography literature already holds instruments for household decision-making and spousal support, and adapting them would take little.

5. The Durability Gap

5.1 When measurement happens

No measurement of any kind followed the close of the programme in 32 of the 42 included studies. Ten looked again, between three and twelve months later. Against the claims the field makes, Figure 1 sets the measurement window.

Figure 1. The measurement window and the span of the claim

	Before the programme	At the close of the programme	Six months and beyond
--	-----------------------------	--------------------------------------	------------------------------

	Before the programme	At the close of the programme	Six months and beyond
Where this literature measures	Baseline in most studies	Post-test in all 42 studies	Follow-up in 10 of 42 studies
What the claims concern	—	—	Employment, enterprise, leadership, sustained autonomy
Consequence			
The window in which the field measures and the window in which its claims fall do not overlap. Whether what these programmes produce survives contact with ordinary life is therefore a question the published record cannot answer.			

Note. Constructed by the authors from the extraction sheet of the underlying review.

Elision being routine here, the implication is worth putting plainly. Long-term transformation in women's lives is what policy documents invoke to justify programmes of this kind, while measurement in the evidence underpinning those justifications happens almost entirely at the moment the programme ends. A gap therefore separates the temporal reach of the claim from the temporal reach of the evidence, and careful synthesis, however careful, does not close it.

5.2 What the follow-up studies show

Smaller effects appear on every dimension among the ten studies with follow-up: psychological +0.47 against +0.68 for the post-test-only group, social +0.31 against +0.51, economic +0.08 against +0.24. The obvious reading is attenuation. Gains measured as a programme closes decay, and so less is found by a study that returns. The qualitative accounts suit that reading too, describing a spell of elevated confidence that held where circumstances supported it and drained away where they did not.

That reading cannot rest on the comparison alone, and it is here that this paper departs from the usual handling of such a finding. Those ten studies do not represent the literature. They are better resourced and better designed, an independent evaluator is likelier to have been involved, and better design produces smaller effects for reasons wholly unconnected to elapsed time. Three parts make up the honest statement. Smaller effects under follow-up are consistent with attenuation. Attenuation is theoretically plausible and the qualitative accounts support it. And this evidence base cannot separate attenuation from design quality.

The finding is that the question cannot be settled. It also demonstrates this paper's general argument more clearly than anything else, the confound existing only because follow-up is uncommon. Make measurement at six months routine and the follow-up subset would look like the rest of the literature, leaving the comparison interpretable.



5.3 What would settle it

An answer is available, and the design that would deliver it is unremarkable. Recruit a prospective cohort, set beside it a non-equivalent comparison group of eligible non-participants, measure at baseline, at the close of the programme, at six months and at twelve, and carry at least one objectively verifiable outcome alongside the psychological scales. Persistence of the reported gains would then be established. Put items on household response and institutional openness into every wave and what persistence rests on would be established as well. Nothing there asks much technically. Instruments for all of it are already in circulation, and set against a conventional evaluation the extra outlay is a single further wave of data collection and a brief supplementary questionnaire.

No such study appears to have been conducted, and in a literature this large and this old that is the single most informing fact reported here. The constraint binding this evidence base, it suggests, is not what methods can do but how evaluations come to be commissioned, a matter Section 7.2 takes up.

6. Methodological Quality

6.1 Appraisal results

Summary appraisal scores average 0.68 over the included set. The good or strong bands, at 0.75 and above, were reached by eleven studies; adequate describes the other 31, and below 0.55 there is nothing, the 13 studies that fell there having gone at the appraisal stage. What the mean conveys is less than what it hides. The item-level pattern informs more, and Table 3 groups the fourteen quantitative items according to how the included studies fared on them.

Table 3. Pattern of item-level performance on the quantitative appraisal checklist

Performance	Checklist items	Interpretation
Consistently strong	Objectives described; design evident and appropriate; subject selection described; participant characteristics described; results reported in sufficient detail	Reporting of what was done is generally adequate
Not applicable by design	Randomisation described; blinding of investigators; blinding of subjects	A training programme cannot be blinded; these items are scored as not applicable rather than as failures in most studies
Mixed	Outcome measures well defined and robust; sample size appropriate; analytic methods described and justified; conclusions supported by results	Varies with publication type; weakest in agency reports
Consistently weak	Control of confounding; estimate of variance reported	Where enrolment is voluntary the



Performance	Checklist items	Interpretation
		disposition that produced enrolment is an unaddressed rival explanation; absent variance, the spread behind a reported mean is unknown

Note. Items as specified by Kmet et al. (2004). The full item-by-study matrix is reported in the appendix of the underlying review.

One item shows genuinely weak performance where the weakness is not structural, and that is control of confounding. Enrolment here is usually a matter of choice. The woman who enrolls is unlike the woman who does not, and whichever disposition drove her to enrol stands as a live alternative explanation for any change that follows. Confronting it is rare. Where it happens, the studies cluster among the eleven in the good and strong bands.

6.2 Bias flags

Separately from the scores, four sources of bias peculiar to this literature were recorded study by study, since the checklists miss them and a study otherwise well conducted can still carry one. In 32 of the 42 there is nothing beyond the post-test. Self-selection into a voluntary programme goes unadjusted in a little over half the set. The trainer administers the outcome measure in rather more than a third. A smaller group, concentrated among the agency reports, was evaluated by the funding or delivering body with no independent assessor. Most studies carry more than one of these flags; few carry none.

Put beside Section 4, these flags pin down the composition of the evidence fairly precisely. A good deal of what this literature records amounts to participants' statements about themselves, often on an unvalidated instrument, made to whoever had just trained them, at the instant training finished, with no adjustment for the disposition that brought them there in the first place. Weight attaches to that, since participants' own accounts are not nothing. But it falls considerably short of the conclusions the field draws from it.

7. Discussion

7.1 What this evidence base can and cannot establish

As the corpus stands, three claims hold. Participants leave these programmes reporting higher self-efficacy and self-esteem, consistently across designs, settings and provider types, and the finding holds when the set is restricted to studies with independent assessors. Measured gains come out larger on unvalidated than on validated instruments. And where external outcomes appear at all, they appear in the longer programmes that sat inside some continuing structure. Four claims made routinely do not hold. One: the effects persist. Two: they convert into employment, enterprise or leadership position. Three: one programme model beats another, which almost no study can speak to, variants having been compared against nothing instead of against each other. Four: the programmes are cost-effective, which no included study can



support, since none reports cost per participant and nothing about value for money follows from this literature at all.

Table 4. Claims about these programmes, set against what the evidence base supports

Claim	Status	Basis
Participants report higher self-efficacy and self-esteem at the close of the programme	Supported	39 of 42 studies; survives restriction to studies with independent assessors
Measured gains are larger on unvalidated than on validated instruments	Supported	Instrument comparison, Section 4.1
External outcomes appear where programmes ran longer and were embedded in a continuing structure	Supported, with moderate confidence	Clustering of the studies reporting such outcomes
The effects persist beyond the close of the programme	Not established	32 of 42 studies take no later measurement; the follow-up subset is confounded with design quality
The programmes produce employment, enterprise or leadership position	Not established	Measured by fewer than half the studies, and rarely beyond the post-test
One programme model outperforms another	Not established	Almost no study compares variants against each other rather than against nothing
The programmes represent good value for money	Not assessable	No included study reports cost per participant

Note. “Not established” denotes absence of evidence and should not be read as evidence of absence.

Absence of evidence and evidence of absence are not the same thing, the difference does real work in this argument, and it ought to be preserved. Nothing among the unsupportable claims has been refuted. Each was simply never put to the test, by a practice of measurement that halts too soon and faces only one way. Treating an absence as an absence, rather than converting it into a negative finding, is a discipline the qualitative metasummary approach makes explicit (Sandelowski et al., 2007); the confidence attaching to each synthesised statement should likewise be rated and reported, not merely implied (Lewin et al., 2018).

7.2 Why the field measures what it measures

Separate shortcomings is how the three features above are usually listed. Better to see one pattern with one cause behind it, and the cause is not carelessness.

Most evaluations here are run by the organisations delivering the programmes, under funding agreements that call for an evaluation while specifying none. Every choice this paper documents is, inside that constraint, the rational one. Nothing is cheaper than a validated scale, but an unvalidated scale can be written to fit the programme, and measuring at the close costs little and catches everyone whereas measuring at six months costs more and loses people. Over the participant the evaluator has control; over her household and her labour market, none. And a favourable result follows near enough automatically from the first three choices, which is what the funder's reporting cycle rewards.

On this reading, what Section 3.3 documents as vocabulary drift is not intellectual development at all. The field's turn toward the individual and the psychological followed, in part, from what its evaluation economics left measurable. Measurement practice in this area, Kabeer (2005) observed, had collapsed the distinction between agency and achievement and was treating indicators of one as evidence for the others. Where that collapse is sustained, and how it came about, is what the present analysis suggests.

7.3 A minimum reporting standard

Brevity below is intentional. An ideal design is not what the list describes. It names the fewest requirements under which the existing stock of studies would become interpretable, and a routine programme evaluation can meet every one of them.

Table 4. Proposed minimum reporting standard for evaluations of developmental programmes for women

No.	Requirement	Rationale	Marginal cost
R1	At least one outcome measured no earlier than six months after the programme closes	Closes the durability gap and removes the design confound in the follow-up subset (Section 5)	One additional round of data collection
R2	At least one objectively verifiable outcome, independent of participant self-report	Breaks the dependence on self-report documented in Section 4	Access to records the provider usually already holds
R3	Outcome measures administered by someone other than the trainer, with the administrator identified in the report	Removes a bias that no later analysis repairs	Negligible; a scheduling matter
R4	Household and institutional conditions measured alongside	Instruments the variable the qualitative evidence	One short additional questionnaire



No.	Requirement	Rationale	Marginal cost
	psychological outcomes, using existing instruments	identifies as decisive (Section 4.3)	
R5	Programme content, contact hours, session count, delivery mode and cost per participant reported	Permits comparison between studies and any statement about value for money	Reporting only; the data already exist

Note. R1 and R2 are the two conditions that would change what this literature contains within a single funding cycle.

Reporting requirements rather than research requirements describe three of the five: the data are already there and simply go unpublished. Funders, not researchers, are better placed to impose them, the requirement belonging in the funding agreement and not in the journal. Among the changes this analysis identifies, the highest ratio of consequence to cost attaches to specifying the evaluation instead of merely requiring that one happen.

7.4 Limitations

Boundaries belonging to the underlying review bind this analysis too, and English and Hindi were the limits of the searching, which under-represents regional evaluation evidence and probably feeds the geographical concentration described in Section 3.1. Publication bias is likely as well. Write-up is less likely for an evaluation that finds nothing, acceptance less likely again, and least likely of all is posting by an organisation whose funding turns on the result. Grey literature searching softens this without removing it, and a synthesis of this kind has no statistical test for it available.

Two further limitations are this paper's own. The corpus the bibliometric mapping describes was defined by the review's eligibility criteria, which makes the vocabulary shift it records a property of the literature on women's developmental programmes rather than of empowerment research at large. The other concerns Section 7.2: the explanation offered there for the field's measurement habits interprets the pattern, and while consistent with it the explanation is not tested by it. Testing would call for data on evaluation commissioning, which this review never collected.

One last caveat holds for analyses of this kind generally. Nothing above reaches past the published record. Whether that record sits close to what these programmes actually do, or far from it, is unknown, and no sentence here should be taken as describing the programmes themselves.

8. Conclusion

Careful attention goes to the participant in the literature on women's personality development programmes, paid at the moment she leaves the room, with instruments whose strength sits where the constructs are narrowest. Her circumstances draw no attention at all, and three quarters of the time no second reading is taken. Those three facts hang together, and between them they fix the limits of what the field can say.



Where the bound falls is sharper than the field admits. Women come out of these programmes feeling more capable, and that much is established with some confidence. Whether anything then follows is not established, and the reason lies not in evidence that cuts the other way but in evidence that scarcely reaches the question. Take readings only at the door and none of what waits beyond it, and the resulting literature settles neither that these programmes empower women nor that they fail to.

No methodological advance is required to close the gap, nor any large increase in funding. What is required amounts to a further round of data collection, one instrument taken over from an adjacent field, and a funder willing to put both into the agreement. Where long-run effects have actually been demonstrated, Heckman and Kautz (2012) note, the programmes were sustained ones, multi-component, and followed up over years. Nothing in the literature reviewed here matches that description, and what can most usefully be said of it is that a known and inexpensive remedy exists.

References

1. Batliwala, S. (2007). Taking the power out of empowerment: An experiential account. *Development in Practice*, 17(4–5), 557–565.
2. Bramer, W. M., Rethlefsen, M. L., Kleijnen, J., & Franco, O. H. (2017). Optimal database combinations for literature searches in systematic reviews: A prospective exploratory study. *Systematic Reviews*, 6, 245.
3. Cornwall, A. (2016). Women's empowerment: What works? *Journal of International Development*, 28(3), 342–359.
4. 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, 285–296.
5. Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour Economics*, 19(4), 451–464.
6. Kabeer, N. (2005). Gender equality and women's empowerment: A critical analysis of the third Millennium Development Goal. *Gender & Development*, 13(1), 13–24.
7. Kmet, L. M., Lee, R. C., & Cook, L. S. (2004). *Standard quality assessment criteria for evaluating primary research papers from a variety of fields* (HTA Initiative No. 13). Alberta Heritage Foundation for Medical Research.
8. Lewin, S., Booth, A., Glenton, C., Munthe-Kaas, H., Rashidian, A., Wainwright, M., Bohren, M. A., Tunçalp, Ö., Colvin, C. J., Garside, R., Carlsen, B., Langlois, E. V., & Noyes, J. (2018). Applying GRADE-CERQual to qualitative evidence synthesis findings: Introduction to the series. *Implementation Science*, 13(Suppl. 1), 2.
9. Luthans, F., Avolio, B. J., Avey, J. B., & Norman, S. M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology*, 60(3), 541–572.
10. Mosedale, S. (2005). Assessing women's empowerment: Towards a conceptual framework. *Journal of International Development*, 17(2), 243–257.



11. Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, 18, 143.
12. Nasheeda, A., Abdullah, H. B., Krauss, S. E., & Ahmed, N. B. (2019). A narrative systematic review of life skills education: Effectiveness, research gaps and priorities. *International Journal of Adolescence and Youth*, 24(3), 362–379.
13. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71.
14. Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, 90(1), 24–46.
15. Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.
16. Sandelowski, M., Barroso, J., & Voils, C. I. (2007). Using qualitative metasummary to synthesize qualitative and quantitative descriptive findings. *Research in Nursing & Health*, 30(1), 99–111.
17. Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio* (pp. 35–37). NFER-NELSON.
18. van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538.