# What are the key measures or approaches used to assess individual-level outcomes targeted by capacity development interventions (e.g., capabilities, performance, career growth)?

## Executive Summary

This review examines how individual-level outcomes of capacity development - such as capabilities, on-the-job performance, and career progression - are defined and measured. Its aim is to clarify what should be assessed and how, so that stakeholders can judge whether interventions work, compare results across settings, and design stronger programs.

Three families of guidance shape current practice. First, outcome frameworks (notably four-level training models) keep attention on whether learning translates beyond satisfaction and knowledge into changed behavior and results. Second, multi-domain models (e.g., knowledge–attitudes–skills with cognitive/affective/psychomotor domains) highlight the need to move beyond factual recall to include values and application. Third, motivational and work-design perspectives (such as job demands–resources) highlight sustained change, illustrated through longitudinal tracking of self-efficacy, job crafting, and the application of strengthened competencies.

Across studies, seven outcome domains recur with corresponding measures: (1) knowledge gains (pre/post tests and item banks); (2) skills and competencies (simulations, workplace assessments, and scored work products); (3) attitudes, self-efficacy, and confidence (validated scales rather than ad hoc items); (4) behavioral performance (direct observation, multi-rater/360 feedback, and verified task indicators); (5) career advancement and productivity (promotions, role changes, publications, grants); (6) soft skills and leadership traits (psychometric inventories and structured scenarios); and (7) economic or value-added metrics (cost–benefit, value-added per employee, or broader social return approaches).

Several cross-cutting insights emerge. The strongest evidence pairs validated instruments with observed behavior and tangible outputs; mixed-methods designs and longitudinal follow-ups are pivotal for credibility. At the same time, comparison groups - while ideal - are often impractical for implementers; feasible alternatives include clearly defined baselines, repeated measures, external verification of key outcomes, and prudent triangulation to reduce bias. Behavioral and career outcomes are under-measured or weakly verified not because they are unimportant, but because they are inherently difficult and resource-intensive to track; proportionate strategies (e.g., brief follow-ups at targeted intervals, sampled verification of critical behaviors, and use of existing administrative records) are more realistic than blanket long-term monitoring of all participants. Likewise, economic indicators should be used selectively and only where a plausible pathway to monetizable or efficiency gains exists; in many evaluation-capacity contexts, non-monetary evidence of performance and behavior change will be more decision-relevant.

These findings matter because they move evaluation beyond surface-level knowledge checks to a focus on real-world impact - whether participants applied skills in practice, advanced in their careers, and generated measurable value. In practical terms, this means prioritizing valid, observable indicators that are feasible to collect in context, and using cautious causal language commensurate with the strength of the design. This shift strengthens accountability and ensures that investments in capacity development are assessed by their lasting outcomes rather than their immediate outputs.

## Introduction

This literature review addresses the question: *What are the key measures and approaches used to assess individual-level outcomes targeted by capacity development interventions, such as capabilities, performance, and career growth?* Addressing this question is critical because credible and consistent measurement is essential for understanding whether capacity development efforts achieve their intended effects, for comparing results across contexts, and for informing the design of future initiatives.

The review proceeds in four stages. First, it examines conceptual frameworks that guide measurement choices, including models that structure outcome domains and causal pathways. Second, it analyzes the principal categories of outcomes - ranging from knowledge and skills to attitudes, behaviors, career trajectories, soft skills, and economic metrics - together with the instruments most frequently employed to assess them. Third, it reviews the methodological approaches and rigor of outcome measurement, highlighting strengths, limitations, and emerging practices. Finally, the review synthesizes findings and identifies implications for evaluation of capacity development.

The scope of the review is deliberately bounded to the individual level, while recognizing that individuals are embedded in organizational and systemic contexts. It draws primarily on recent peer-reviewed studies across diverse sectors and regions, with an emphasis on approaches that are innovative, rigorous, or widely cited.

By mapping the landscape of measures and methods, this review seeks to clarify how individual-level capacity outcomes are operationalized and assessed, and to contribute to the development of more robust, comparable, and actionable evaluation practices.

## Conceptual Frameworks for Individual-level Outcome Measurement

Decisions about *what* to measure and *how* to measure it are often guided by established evaluation frameworks and models. Three conceptual frameworks appear frequently in the literature and strongly influence the selection of outcome indicators: Kirkpatrick’s four-level model, the Knowledge–Attitude–Skills (KAS) triad combined with Bloom’s learning domains, and the Job Demands–Resources theory. Each framework offers a distinct perspective on the causal pathway from an intervention to individual outcomes, thereby shaping the types of evidence that evaluators collect.

### Kirkpatrick’s Four Levels

Perhaps the most ubiquitous model in training evaluation is Kirkpatrick’s four-level hierarchy (Kirkpatrick, 1994). This model delineates outcomes in a sequence from immediate reaction (Level 1) and learning gains (Level 2), through behavior change on the job (Level 3), to results (Level 4) such as improved organizational performance. Kirkpatrick’s framework remains widely used in both research and practice to categorize training outcomes (Shewchuk et al., 2023). For instance, a recent review of knowledge-brokering (K\*) training programs found that most evaluations still rely on Kirkpatrick’s categories to report outcomes (Shewchuk et al., 2023, p. 5). However, many studies focus on the lower levels (learner satisfaction and knowledge tests) and provide limited evidence of sustained behavior change or organizational results (Shewchuk et al., 2023, pp. 7–8). Kirkpatrick’s model usefully emphasizes that effective training should ultimately translate into on-the-job performance and broader results, but it also highlights a well-recognized challenge: attributing organizational results to an individual training is difficult due to many confounding factors.

### KAS and Bloom’s Taxonomy

A second framework combines the classic triad of Knowledge–Attitudes–Skills (KAS) with Bloom’s taxonomy of learning domains (cognitive, affective, psychomotor) (Bloom, 1956). This approach encourages a multi-dimensional view of individual change. In practice, it means that an intervention’s outcomes are classified by the type of change (knowledge gained, attitude shifted, or skill acquired) and the domain of learning (intellectual/cognitive, emotional/affective, or physical/practical skills). For example, Kuppuswami and Ferreira (2022) describe an evaluation of a gender-equality capacity-building program where each survey item completed by participants was mapped to both a specific capacity topic and to one of Bloom’s domains. This mapping made it possible to analyze whether reported changes fell primarily in the cognitive realm (e.g. factual knowledge of gender policies), the affective realm (e.g. commitment to gender equity values), or the behavioral realm (e.g. ability to act against discrimination). The KAS+Bloom framework thus provides a structured lens to ensure that evaluations capture more than one dimension of learning.

It is particularly useful in interventions aiming for transformative outcomes, such as changes in values or attitudes in addition to skills. However, one limitation noted in such evaluations is the reliance on self-reported data: even if a questionnaire covers multiple domains of learning, participants’ subjective reporting may not always reflect actual changes in their behavior or beliefs. In the gender training example, increases in self-reported knowledge and willingness to challenge bias were observed (Kuppuswami & Ferreira, 2022), but without external validation, it remains uncertain how those self-perceptions translate into action. This underscores a broader issue across frameworks: multi-domain outcome models broaden what is measured, but ensuring the validity of those measures (especially for internal states like attitudes) remains a challenge.

### Job Demands–Resources (JD-R) and Personal Resources

Whereas Kirkpatrick and Bloom focus on what changes (levels or domains of learning), the JD-R model offers a measurement-oriented lens for selecting individual-level indicators in work contexts, emphasizing how job demands and available resources relate to motivation, well-being, and performance. For outcome measurement, JD-R points evaluators to track specific constructs - such as job crafting behaviors, access to and use of job resources (e.g., feedback, autonomy, development opportunities), and personal resources (e.g., self-efficacy) - with validated scales and repeated measures to assess maintenance over time. In capacity development, JD-R has been used to operationalize outcomes like frequency of job-crafting actions, perceived resource availability and utilization, and changes in self-efficacy rather than to argue theoretical causality.

An illustrative study by van Wingerden, Bakker, and Derks (2017) demonstrates JD-R-aligned measurement choices, including validated self-efficacy scales and behavior-based indicators of job crafting collected at multiple time points (immediately after the program and again one year later). From a metrics standpoint, the study exemplifies three practices relevant here: (1) using psychometrically sound instruments for personal and job resources; (2) capturing behavior frequency (e.g., seeking feedback, experimenting with new strategies) as an outcome in its own right; and (3) employing longitudinal follow-ups to test whether observed changes are maintained (van Wingerden et al., 2017). These design choices are germane to this review because they indicate what to measure (resources, crafting, self-efficacy), how to measure it (validated scales plus behavior items), and when to measure it (immediate and delayed follow-ups).

In practical terms for evaluators, JD-R’s contribution is a structured menu of outcome constructs and corresponding measurement approaches: select a small set of resource and job-crafting indicators that align with the training objectives; pair them with a validated self-efficacy scale; collect baseline and at least one follow-up; and, where feasible, add a modest external check (e.g., brief supervisor verification of one or two salient behaviors) to reduce sole reliance on self-report. This keeps the focus on feasible, practitioner-oriented metrics while maintaining evidentiary discipline. Nonetheless, because many JD-R applications rely on self-reported frequencies and perceptions, results should be interpreted cautiously and - where possible - triangulated with observation or simple documentation checks (e.g., evidence of new practices or routines being adopted).

## Categories and Measures of Individual-Level Outcomes

This section examines the principal domains in which individual-level outcomes of capacity development are assessed, along with the indicators and instruments most commonly employed. Organized into seven subsections - *Knowledge and Learning Gains*; *Skills and Competency Development*; *Attitudes, Self-Efficacy, and Confidence*; *Behavioural Performance*; *Career Advancement and Research Productivity*; Soft Skills and Leadership Traits; and *Economic and Value-Added Metrics* - it provides a structured overview of the range of outcomes targeted in evaluation. Collectively, these domains capture the breadth of change that capacity development initiatives seek to foster, from cognitive knowledge acquisition to the demonstration of new skills, shifts in confidence and motivation, observable behavior in professional settings, longer-term career advancement, and even economic or value-based contributions. By outlining both the outcomes themselves and the tools used to measure them, the section highlights the diversity of evidence that can inform judgments about the effectiveness of capacity development interventions.

### Knowledge and Learning Gains

Cognitive learning – the acquisition of knowledge – is the most frequently assessed outcome in capacity development interventions. Knowledge gains are relatively straightforward to measure and are often seen as a prerequisite to other capacity outcomes (skills or behavior change). The standard method is to administer knowledge tests before and after the intervention and then compute the improvement. Such tests can take various forms, including multiple-choice quizzes, short-answer questions, case-based problems, or even true/false items, depending on the subject matter. For example, in training programs for public-sector employees in Botswana, participants completed paper-and-pencil exams covering key content both at the start (pre-test) and end (post-test) of the course (Tshukudu, 2009). Questions ranged from factual recall to applied scenario responses, and results were sometimes summarized using “gain scores” to quantify the percentage improvement for each participant (Tshukudu, 2009).

This approach controls for differing pre-training knowledge levels and focuses on the incremental learning attributable to the training. In other contexts, such as academic capacity strengthening initiatives, more specialized knowledge assessments are used. Davis and D’Lima (2020), in a systematic review of training in dissemination and implementation (D&I) science, noted that many workshops included a pre-course quiz on key concepts (e.g. knowledge of stakeholder engagement strategies or familiarity with implementation frameworks) and then repeated the same quiz at course completion. Consistently higher scores on the post-test are taken as evidence of knowledge acquisition (Davis & D’Lima, 2020). These are relatively routine applications of knowledge testing, but they underscore the central role of pre/post knowledge exams as a metric for training effectiveness across disciplines.

Beyond traditional tests, some programs use more interactive methods to assess knowledge. For instance, in a community health training on water and sanitation practices (Community-Led Total Sanitation, or CLTS), Crocker et al. (2016) report that instead of a written test, trainees participated in a structured debrief interview after the training. They were asked to recount the steps of the CLTS process and explain key concepts in their own words, shortly after completing the workshop. This conversational assessment revealed which concepts were vividly retained (participants uniformly remembered the dramatic “triggering” exercises used to ignite community action) and which were less salient (many struggled to recall the later follow-up and monitoring stages in detail) (Crocker et al., 2016). Such narrative or oral assessments can probe the depth of understanding and highlight areas of confusion that a multiple-choice test might miss. However, the interpretation of qualitative knowledge assessments is more subjective. In the CLTS example, the evaluators treated the ability to recount all steps correctly as a sign of strong knowledge gain, but also gleaned insights about relative emphasis or engagement with different modules (Crocker et al., 2016). Generally, knowledge assessments are considered the most direct evidence of learning; they are easy to implement at scale and lend themselves to quantitative analysis (mean scores, % improvement). Their limitations are well-known too: a high score on a test does not guarantee the ability to apply that knowledge in practice, and test results can be influenced by factors like test-taking skills or question phrasing. To mitigate some issues, programs sometimes design criterion-referenced tests closely tied to learning objectives, or use item analyses to ensure reliability. In sum, measuring knowledge gain is an indispensable but initial step in evaluating capacity development. It tells us whether the cognitive content was absorbed – a necessary condition for downstream changes in behavior or performance, but not sufficient on its own for capacity to be realized.

### Skills and Competency Development

When training moves from the classroom to the workplace, the focus shifts to skill development – what participants can *do* with their knowledge. Skills are often more difficult to measure than knowledge because they must be demonstrated through action, not just described. The literature shows a spectrum of approaches to evaluating skills, from simple self-assessments to complex performance simulations. At a basic level, some interventions use self-reported proficiency ratings or supervisor observations to gauge skill acquisition. For example, Tshukudu (2009) describes a post-training “skills profile” matrix used in Botswana’s public service: participants, together with their supervisors, rated the trainee’s ability to perform specific job tasks before and after training. These tasks (e.g. using a particular software, or handling a customer inquiry) were listed in a matrix, and proficiency was marked on a scale from novice to expert (Tshukudu, 2009, pp. 231–232). Such matrices provide a structured way to capture perceived skill growth on the job and can be tailored to any job function. However, they rely on subjective judgments and can be prone to bias (e.g. supervisors over-rating their staff, or participants overestimating their own abilities).

At the other end of the methodological spectrum are objective performance assessments. In some high-stakes or high-skill fields, training programs utilize assessment centers or practical examinations to measure skill gains. A notable example comes from global leadership development programs in multinational companies. According to Cumberland et al. (2016), it is common for such programs to include simulated management exercises: participants might be placed in a role-play scenario leading a multicultural team or negotiating a business deal, while trained assessors observe their behaviors. The assessors use predefined rubrics to score competencies like communication, problem-solving, cultural intelligence, and team leadership (Cumberland et al., 2016). Because these simulations mirror real challenges and require participants to apply a range of soft and hard skills, they are considered a gold standard for evaluating complex competencies. Similarly, in healthcare training (for example, upskilling primary health care workers), simulation-based assessments are widely used. Trainees may be tested via Objective Structured Clinical Examinations (OSCEs), where they rotate through stations performing clinical tasks on mannequins or standardized patients under observation (Finn et al., 2021). Evaluators then use checklists to mark whether each step or criterion of the skill was demonstrated correctly (Finn et al., 2021). This provides fine-grained data on technical proficiency (e.g. did the nurse follow all infection control steps while administering an injection?) and can highlight specific gaps that still need improvement.

Between self-reports and full simulations, there are also work products and practical exercises used as evidence of skill. In research capacity-building programs, a common approach is to have participants produce something tangible – such as a draft grant proposal, a data analysis, or a policy brief – and then assess the quality of those outputs. Pulford et al. (2020) note that in some research training, trainees’ grant proposals or scientific manuscripts are scored against a standard rubric as a way to measure research skill development. The assumption is that a stronger proposal (for example, clearly written aims, sound methodology) reflects higher research capacity. This is a direct outcome of training that can be more meaningful to stakeholders than a test score. The drawback is the time and expertise required to grade such outputs reliably.

In summary, measuring skills requires balancing practicality and rigor. Simulations and direct assessments yield high-quality evidence of skill performance but are resource-intensive; surveys and self-assessments are easy to deploy but less credible. Many programs choose mixed approaches (e.g. a self-assessment combined with a supervisor’s assessment, or a test combined with a sample work product review) to get a fuller picture. Regardless of method, skill evaluation is crucial because it bridges the gap between knowing and doing - ultimately, capacity development’s success hinges on people not only understanding new concepts but being able to perform new tasks or old tasks better.

### Attitudes, Self-Efficacy, and Confidence

Not all outcomes of capacity building are about knowledge and skills; often, interventions aim to change how participants feel or perceive themselves in relation to their work. This domain includes outcomes like attitude changes, increases in self-efficacy or confidence, and shifts in motivation or commitment. Such changes are important because they can influence whether and how individuals apply their new knowledge or skills. However, by their nature, attitudes and self-perceptions are typically internal states, making them harder to measure objectively. The literature primarily relies on self-report questionnaires to assess these outcomes, and while common, this approach requires cautious interpretation.

Attitude change is frequently targeted in interventions that have a social or behavioral component. For instance, a capacity-building program on gender equity might seek to not only teach facts about gender disparities but also to encourage more positive attitudes toward women’s leadership. In evaluating a series of gender-equality workshops, Kuppuswami and Ferreira (2022) included survey items that asked participants the extent to which they personally endorse statements like “Women should have equal opportunities to lead in my organization.” By comparing responses before and after the training, they found significant positive shifts, indicating more supportive attitudes after the program. Similarly, in public-sector training in Botswana, evaluators used semantic differential scales to capture changes in trainees’ perceptions of the training and its relevance (Tshukudu, 2009). Participants were asked to rate aspects of the training on bipolar adjective scales (e.g. “meaningless – meaningful”, “boring – engaging”). Consistent movement towards the positive end after the course was interpreted as a sign that participants found the training more relevant and motivating than they initially expected (Tshukudu, 2009, pp. 233–236). These attitude measures can be considered proxies for engagement and buy-in - if a person’s attitude towards the subject matter becomes more favorable, one might predict they are more likely to implement what they learned. However, these are still subjective measures; a participant might report a positive attitude shift without it actually translating into any change in their actions or decisions at work.

Self-efficacy, or one’s belief in their ability to perform certain tasks, is another common outcome measured in capacity development. Interventions often strive to boost individuals’ confidence that they *can* execute new skills or take on new challenges. The JD-R intervention discussed earlier explicitly measured teachers’ self-efficacy in handling job demands (van Wingerden et al., 2017). They used a validated self-efficacy scale with items like “I am confident that I can handle many challenges at my job,” rated on a Likert scale. The results showed a statistically significant increase in self-efficacy scores among teachers who underwent the job crafting training, even one year later, whereas no such increase was observed in a control group (van Wingerden et al., 2017). This suggests the intervention had a durable impact on participants’ own beliefs about their capabilities, which is encouraging for long-term capacity retention. Self-efficacy scales have the advantage of being theory-based and often rigorously tested for reliability and validity (many draw from established instruments in psychology). By using a validated scale, researchers can be more confident that changes in the score reflect real changes in the underlying construct (in this case, confidence in managing one’s work), rather than random variation or measurement error.

Many other training evaluations include simpler confidence ratings or self-assessed competence items. For example, in training programs for health research or implementation science, participants might be asked to rate how confident they feel in performing specific tasks (like “conducting a literature search” or “developing a monitoring and evaluation plan”) on a scale from 1 (not confident) to 5 (very confident) before and after training (Davis & D’Lima, 2020). Davis and D’Lima’s review found that almost all the D&I training programs they examined reported increases in self-rated confidence across various skill areas (2020). However, a noted shortcoming was that many of these confidence questionnaires were developed ad hoc by the program designers and had unknown psychometric properties (Davis & D’Lima, 2020). The lack of standardization or validation means one cannot be entirely sure if a rise in “confidence” score truly indicates a meaningful gain or if it might be influenced by response bias (for instance, participants feeling they should report being more confident after an investment in training).

In some cases, attitude and self-efficacy measures can serve as early indicators of deeper change. They often shift before observable behavior does. For example, a person may first internalize the value of evidence-based decision-making (attitude) and become confident in their ability to use data (self-efficacy), and only later will we see them actually implement a new data analysis process at work (behavior). Because of this temporal dynamic, many evaluators collect attitude and confidence data immediately post-training, as a sign that the training has at least influenced the participants’ mindsets in the intended direction. These outcomes can be critical for participant engagement as well: a training that fails to move attitudes or confidence at all might indicate that participants remain unconvinced or unsure about using what they learned, suggesting limited potential for impact.

Nonetheless, the limitations of self-reported attitudes and confidence are significant. Participants may give “socially desirable” responses (e.g. claiming more progressive attitudes or higher confidence because they think that’s expected). Also, increased confidence is not always positive - there is the phenomenon of overconfidence where individuals might feel very confident but still lack actual skill (the Dunning-Kruger effect). Therefore, best practice is to triangulate these measures with other data. For instance, one might pair a confidence survey with a skills test to see if the most confident participants are indeed the most skilled (often they are, but not always). Another strategy is collecting multi-source feedback: some leadership programs use 360-degree feedback where not only does the individual rate their own capabilities and attitudes, but their peers and subordinates also provide ratings of the individual’s behaviors (Cumberland et al., 2016). Discrepancies between self-perception and others’ observations can be illuminating (e.g. a manager might feel confident in their communication skills, while their team still finds them lacking - pointing to a need for further development).

In summary, attitudinal and self-perception outcomes are valuable pieces of the evaluation puzzle. They capture the often-intangible shifts in mindset that accompany capacity development. The evidence base shows these are commonly measured via Likert-scale surveys and show consistent improvements post-intervention in areas like motivation, commitment, and self-belief (DeCorby-Watson et al., 2018; van Wingerden et al., 2017). However, because they are subjective, it is important to interpret them alongside more objective indicators. A positive change in attitude or self-efficacy is encouraging, but the ultimate goal is that such changes support actual improvements in performance and outcomes, which is the focus of the next section.

### Behavioural Performance

The true test of capacity development is whether individuals apply their enhanced knowledge and skills in their work, leading to observable changes in behavior or performance. Measuring behavioral outcomes is therefore a critical bridge between learning and organizational impact. Unlike test scores or survey responses, behavioral measures aim to capture what people actually do differently as a result of an intervention. This is inherently challenging: behaviors must be observed or reliably reported, and attributing changes in performance to a specific training can be complex. Despite these challenges, various approaches to assessing behavior change appear in the literature, including direct observation, follow-up interviews, performance metrics, and structured assessments.

One rigorous approach is direct observation in the workplace after a training. For example, in the Kenyan CLTS capacity-building program, evaluators conducted site visits and in-depth interviews several months after the training to see how participants were managing sanitation projects (Crocker et al., 2016). By examining meeting records, project reports, and interviewing colleagues, the evaluators documented concrete new practices that reflected skills emphasized during training, such as initiating interdepartmental committees and developing harmonized monitoring indicators (Crocker et al., 2016). These qualitative data provide not only confirmation of behavior change but also contextual insight into how and why it occurred. The trade-off is resource intensity, which often limits feasibility for routine program evaluation.

Surveys that ask about the frequency of specific behaviors provide a more scalable but less rigorous option. For instance, van Wingerden et al. (2017) used repeated surveys to track how often participants engaged in job-crafting behaviors such as seeking feedback or reorganizing tasks. While such frequency-based measures are grounded in concrete actions, they remain self-reported and thus subject to recall and response bias. More robust practice would combine self-reports with at least one form of independent verification, such as simple document checks or supervisor sign-off on selected behaviors.

Some training programs incorporate structured behavioral assessments tied closely to the intended learning objectives. In leadership or management development, this might involve simulated workplace scenarios where participants’ behaviors are scored against predefined rubrics. In technical or service-oriented roles, output metrics - such as the number of cases successfully resolved or error rates reduced - can serve as performance proxies if they are closely linked to the targeted skills. These types of outcome measures are stronger when they are specific, directly connected to training content, and accompanied by clear documentation of how performance indicators map onto the competencies being developed.

Multi-source (360-degree) feedback instruments have also been used, particularly in leadership contexts, but their applicability for capacity development evaluation is mixed. While they can reveal how others perceive changes in an individual’s behaviors, their usefulness depends heavily on whether raters actually observe the targeted behaviors and whether the assessed competencies align closely with the intervention’s objectives. Without this alignment and sufficient rater reliability, 360-degree reviews may provide only generic workplace feedback rather than evidence of training-related behavior change. For practitioner audiences, these tools should be considered cautiously and only when conditions allow for targeted, reliable application.

Finally, the literature acknowledges that rigorous behavioral measures - such as tracer studies, long-term follow-ups, and multi-source verification - are rarely feasible in practice (Morkel & Ramasobama, 2017). The key practical takeaway is that measures must strike a balance: they should be credible enough to signal genuine behavioral change, but also realistic within program resources. Scaled-down approaches, such as targeted follow-ups with a sample of participants, modest document checks, or lightweight observational rubrics, can provide practitioners with usable evidence without imposing prohibitive demands.

### Career Advancement and Research Productivity

Many capacity development efforts aspire to influence participants’ longer-term career trajectories or professional productivity. These outcomes - such as promotions, job role changes, research funding, and publications - signal whether capacity gains extend beyond immediate knowledge or skill improvements. However, they are also the most difficult to measure and attribute directly to a training intervention, since career progression depends on many external factors.

The literature identifies a few common approaches. In research capacity-strengthening programs, bibliometric indicators such as publications, citation counts, and grant success are often used (Pulford et al., 2020). These indicators are attractive because they are objective and can be tracked in public databases. Yet they are rarely applied with standardized definitions or time frames, and Pulford et al. (2020) found that only about 1% of such indicators met quality criteria (specific, time-bound, and sensitive to the intervention). This highlights a major limitation: career and productivity outcomes can provide useful contextual information, but without careful specification they risk becoming ambiguous or misleading measures of training effectiveness.

Fellowship and scholarship programs sometimes track alumni placement, retention, or leadership roles, particularly in efforts to assess “brain circulation” versus “brain drain” (Ripoll Lorenzo, 2012). These indicators may illustrate whether trained individuals return to contribute capacity in their home institutions or sectors. While valuable, such measures again face attribution challenges - participation in training may coincide with career mobility but not necessarily cause it.

For practitioners, the key lesson is that career and productivity indicators should be used sparingly and with caution. They are best treated as complementary, high-level signals rather than core outcome measures. When included, they should be clearly defined (e.g., number of peer-reviewed publications within two years of program completion, or proportion of alumni promoted within three years) and, where possible, benchmarked against comparison groups or baseline trends to improve interpretability.

In practice, most training implementers will not have the resources or longitudinal access required to track career trajectories in a robust way. For this reason, career and productivity outcomes should not be viewed as routine evaluation requirements, but rather as optional indicators suited for selective use in large-scale, well-resourced programs with a clear mandate to demonstrate long-term impact.

### Soft Skills and Leadership Traits

Beyond technical competencies and formal career metrics, some capacity development initiatives explicitly seek to strengthen “soft skills” such as communication, teamwork, adaptability, and leadership traits. These attributes can be critical for applying technical knowledge effectively and for sustaining change in organizational or community settings. However, their relevance as evaluation outcomes depends on whether they are explicitly among the intervention’s learning objectives. Where they are peripheral or incidental, measuring them may not add value.

When relevant, several measurement approaches appear in the literature. Psychometrically validated instruments like the Global Competencies Inventory (GCI) or the Intercultural Development Inventory (IDI) provide structured ways to assess interpersonal and cross-cultural skills (Cumberland et al., 2016). These tools offer quantitative scores and normative comparisons, but their use requires specialized expertise and is mostly confined to leadership development contexts. More commonly, programs rely on simpler self-assessments embedded in post-training surveys (e.g., “I feel more confident communicating with my team”), which are easy to administer but vulnerable to response bias. Qualitative methods such as Most Significant Change narratives (Ripoll Lorenzo, 2012) can also surface evidence of shifts in confidence, adaptability, or collaboration that are difficult to capture quantitatively.

For practitioners, the pragmatic lesson is to only measure soft skills and leadership traits when they are directly tied to the program’s goals and feasible to assess credibly. In such cases, combining at least two sources of evidence - such as self-perception data triangulated with a short situational exercise, peer observation, or supervisor input - can increase confidence in findings. Where programs do not target soft skills explicitly, evaluation resources are better focused on more central outcomes like knowledge, technical skills, and observed behaviors.

### Economic and Value-Added Metrics

In evaluation discourse, economic and value-added metrics are often proposed as ways to demonstrate the return on investment (ROI) of training or capacity development. These approaches seek to translate learning outcomes into monetary terms or efficiency gains. However, their relevance to evaluation capacity development (ECD) work is limited and highly context-dependent. Improved evaluation skills and competencies do not automatically or directly yield measurable cost savings or financial returns, and the data required to establish such links are rarely available in practice.

The literature does contain examples where economic metrics have been considered. For instance, Tshukudu (2009) outlined a model for comparing the estimated value of increased productivity with training costs in Botswana’s public service. Similarly, some research capacity-strengthening initiatives have tracked grant income as a proxy for economic benefit (Pulford et al., 2020). Yet these applications depend on settings where outputs are already closely tied to financial or productivity measures. In most ECD contexts - such as training evaluators in NGOs or public agencies - the connection between improved competencies and quantifiable financial gains is too diffuse to measure reliably.

Broader frameworks such as Social Return on Investment (SROI) have been discussed in relation to capacity building (Ripoll Lorenzo, 2012). However, there is little evidence that SROI has been successfully applied to individual-level capacity development interventions, and even advocates note its methodological and practical limitations. For this reason, it should not be presented as a viable option here.

For practitioners, the key takeaway is that economic metrics should not be considered standard or expected in the evaluation of capacity development at the individual level. Where a credible and direct link to measurable efficiency or financial outcomes exists - such as reduced error rates in service delivery or demonstrable grant acquisition - economic indicators may be useful as supplementary evidence. In most cases, however, evaluators should focus on more direct and observable outcomes like knowledge, skills, behaviors, and performance, which are both feasible and decision-relevant.

## Methodological Approaches and Rigor in Outcome Measurement

The effectiveness of any outcome measure depends not only on what is being measured but also on how the data are collected and the overall rigor of the evaluation design. Different methodologies - quantitative surveys, qualitative interviews, observational studies, and mixed methods - offer complementary strengths and weaknesses. A recurring theme in the literature is that many evaluations of capacity development rely on convenient but lower-strength methods, leaving room for improvement in ensuring data quality, objectivity, and depth.

Quantitative surveys and tests dominate as data collection tools, especially for knowledge, attitudes, and self-efficacy outcomes. Surveys can reach large numbers at relatively low cost, and standardized tests provide concrete scores for statistical analysis. Validated scales (such as established self-efficacy or leadership instruments) enhance the credibility of findings. However, many evaluations still rely on ad hoc questionnaires with uncertain reliability, often applied to small samples or using single items to capture complex constructs (Davis & D’Lima, 2020).

Another limitation is the absence of baseline data or comparison groups. DeCorby-Watson et al. (2018) emphasized that outcome measurement is stronger when supported by a baseline and, where feasible, a comparison. While randomized or quasi-experimental designs are ideal, they are rarely practical for training implementers. More realistic options include establishing clear pre/post measures, using repeated follow-ups, or incorporating modest external verification of outcomes to improve confidence in results.

Qualitative methods - such as interviews, focus groups, and case studies - add depth by capturing outcomes that are difficult to quantify and by explaining how and why changes occur. The Kenyan CLTS case (Crocker et al., 2016) illustrates how in-depth interviews and field observations revealed specific behavior changes that surveys alone might have missed. Similarly, Gerrie et al. (2022) showed how qualitative evidence helped refine a theory of change and identify outcomes valued by participants. Using coding rubrics to analyze qualitative data can reduce subjectivity and allow some comparability across respondents.

Mixed methods approaches combine the breadth of surveys with the depth of qualitative insights. For clarity, this review uses “mixed methods” in its conventional sense: intentionally integrating qualitative and quantitative evidence to enable triangulation. For example, survey data may suggest that most participants applied new skills, while interviews reveal the contexts in which they were applied and barriers encountered. This combination strengthens credibility and provides richer evidence for decision-makers.

Follow-up duration is another critical quality consideration. Many evaluations stop at immediate post-tests, yet capacity gains may fade or only manifest after participants have opportunities to apply them. Longer-term follow-ups - at six months, one year, or beyond - are more demanding but can reveal whether outcomes persist (van Wingerden et al., 2017). Even modestly extending follow-up can substantially improve the usefulness of findings.

Finally, triangulation with external or administrative data is often underused. HR records, performance dashboards, or digital learning analytics can corroborate self-reported changes. For instance, retention data can validate claims of improved staff stability, or e-learning usage data can contextualize test scores. Integrating such sources helps reduce over-reliance on participants’ own accounts.

The practical takeaway is that practitioners do not need experimental designs to strengthen outcome measurement. Instead, they can improve rigor by: (1) collecting baseline and at least one follow-up measure; (2) using validated instruments where available; (3) triangulating self-reports with at least one independent data source; and (4) aligning data collection with program objectives to ensure outcomes measured are decision-relevant.

## Conclusion

This review examined how individual-level outcomes of capacity development are measured across domains such as knowledge, skills, attitudes, behaviors, career trajectories, and, in select contexts, economic indicators. A central insight is that while many tools and frameworks exist, their value for practitioners depends less on conceptual breadth and more on feasibility, credibility, and alignment with program objectives.

For practitioners, three priorities emerge. First, outcome measures should be proportionate and purposeful: select a small number of indicators that directly reflect intended learning objectives rather than attempting to cover every possible domain. Second, methodological rigor can be strengthened through practical steps - using validated instruments where available, establishing baselines, adding at least one follow-up, and triangulating self-reports with external or observational data. Third, evaluation resources should be concentrated on outcomes that are both observable and decision-relevant: knowledge and skills are necessary but insufficient on their own, while behavioral application and credible evidence of practice change provide the most meaningful signals of lasting capacity gains.

By grounding measurement choices in these principles, evaluators can produce evidence that is not only more reliable but also more useful for decision-makers. Rather than aiming for comprehensive or idealized measurement, the goal is to generate credible insights that help stakeholders understand whether capacity development efforts are producing the intended individual-level changes and how those changes can be supported and sustained.

## References

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook I: Cognitive domain. David McKay Company.

Crocker, J., Shields, K. F., Venkataramanan, V., Saywell, D., & Bartram, J. (2016). Building capacity for water, sanitation, and hygiene programming: Training evaluation theory applied to CLTS management training in Kenya. *Social Science & Medicine, 166*, 66–76. <https://doi.org/10.1016/j.socscimed.2016.08.008>

Cumberland, D. M., Herd, A., Alagaraja, M., & Kerrick, S. A. (2016). Assessment and development of global leadership competencies in the workplace: A review of literature. *Advances in Developing Human Resources, 18*(3), 301–317. <https://doi.org/10.1177/1523422316645883>

Davis, R., & D’Lima, D. (2020). Building capacity in dissemination and implementation science: A systematic review of the academic literature on teaching and training initiatives. *Implementation Science, 15*, Article 97. <https://doi.org/10.1186/s13012-020-01051-6>

DeCorby-Watson, K., Mensah, G., Bergeron, K., Abdi, S., Rempel, B., & Manson, H. (2018). Effectiveness of capacity building interventions relevant to public health practice: A systematic review. BMC Public Health, 18, 684. <https://doi.org/10.1186/s12889-018-5591-6>

Finn, M., Gilmore, B., Sheaf, G., & Vallières, F. (2021). What do we mean by individual capacity strengthening for primary health care in low- and middle-income countries? A systematic scoping review to improve conceptual clarity. *Human Resources for Health, 19*, 5. <https://doi.org/10.1186/s12960-020-00547-y>

Gerrie, R. M., Concannon, L., Copsey, J. A., Wright, T., & Young, R. P. (2022). Using a theory of change to evaluate the impact of a conservation training programme: A practitioner’s perspective. *Oryx, 56*(5), 720–727. <https://doi.org/10.1017/S0030605321001551>

Kirkpatrick, D. L. (1994). *Evaluating Training Programs: The Four Levels*. San Francisco, CA: Berrett-Koehler.

Kuppuswami, D., & Ferreira, F. (2022). Gender equality and women’s empowerment capacity building of organisations and individuals. *Journal of Learning for Development, 9*(3), 394–419.

Morkel, C., & Ramasobama, M. (2017). Measuring the effect of evaluation capacity building initiatives in Africa: A review. *African Evaluation Journal, 5*(1), a187. <https://doi.org/10.4102/aej.v5i1.187>

Pulford, J., Price, N., Amegee Quach, J., & Bates, I. (2020). Measuring the outcome and impact of research capacity strengthening initiatives: A review of indicators used or described in the published and grey literature. *F1000Research, 9*, 517. <https://doi.org/10.12688/f1000research.24144.1>

Ripoll Lorenzo, S. (2012). *Evaluating individual approaches to capacity development: A literature review* (Working Paper No. 6). Agricultural Learning and Impacts Network.

Shewchuk, S., Wallace, J., & Seibold, M. (2023). Evaluations of training programs to improve capacity in K\*: A systematic scoping review of methods applied and outcomes assessed. *Humanities and Social Sciences Communications, 10*, Article 887. <https://doi.org/10.1057/s41599-023-02403-5>

Tshukudu, T. T. (2009). *A model for evaluating training and development initiatives in the Botswana public service* (Doctoral thesis, Nelson Mandela Metropolitan University).

van Wingerden, J., Bakker, A. B., & Derks, D. (2017). The longitudinal impact of a job crafting intervention. *European Journal of Work and Organizational Psychology, 26*(1), 107–119. <https://doi.org/10.1080/1359432X.2016.1224233>