Government Language Training Program

Program Evaluation Manual

项目评估手册

A Comprehensive Guide to Program Evaluation, Continuous Improvement, and Performance Measurement for Government Language Training

Document Version	2.0
Effective Date	January 2026
Classification	Internal Use — Program Administration
Prepared By	Office of Language Training Evaluation
Approved By	Director, Government Language Training Program

Table of Contents

1. Introduction

1.1 Purpose of This Manual

This Program Evaluation Manual establishes the policies, procedures, and standards for evaluating the Government Language Training Program (GLTP). It provides a systematic framework for assessing program effectiveness, measuring learner outcomes, and driving continuous improvement across all training delivery modalities. The manual serves as the authoritative reference for program administrators, instructors, evaluators, and stakeholders involved in the evaluation process.

The primary objectives of this manual are to: (a) define a coherent evaluation philosophy grounded in continuous improvement; (b) specify measurable performance indicators aligned with organizational goals; (c) establish standardized data collection, analysis, and reporting procedures; (d) ensure that evaluation findings translate into actionable improvements; and (e) promote transparency and accountability in program operations.

Evaluation within the GLTP is not a compliance exercise but a strategic function. Every evaluation activity is designed to generate evidence that informs decision-making at the operational, tactical, and strategic levels. This manual integrates both formative and summative evaluation approaches, ensuring that the program benefits from real-time feedback as well as periodic comprehensive assessments.

1.2 Evaluation Philosophy: The Continuous Improvement Model

The GLTP adopts a continuous improvement model as its foundational evaluation philosophy. This model recognizes that language training programs operate in dynamic environments where learner needs, organizational priorities, and resource constraints evolve over time. Rather than treating evaluation as a terminal event, the continuous improvement model embeds evaluation into every phase of the training lifecycle.

At its core, the continuous improvement model operates on the principle that systematic assessment, combined with reflective practice and evidence-based decision-making, produces progressively better outcomes. The model draws from established quality improvement frameworks, particularly the Plan-Do-Study-Act (PDSA) cycle developed by W. Edwards Deming, and adapts them to the specific context of government language training.

Core Principles of the Continuous Improvement Model

- **Evidence-Based Decision Making:** All programmatic decisions are informed by data, not assumptions. Evaluation generates the evidence base upon which improvements are predicated.
- **Stakeholder Engagement:** Evaluation is a collaborative process that involves learners, instructors, supervisors, program administrators, and organizational leadership as active participants.

- **Iterative Refinement:** Improvements are implemented in incremental cycles, allowing for rapid testing, feedback, and adjustment. This reduces the risk of large-scale failures and promotes agility.
- **Transparency and Accountability:** Evaluation findings are shared openly with stakeholders, and corrective actions are tracked to completion. Accountability mechanisms ensure that identified issues are addressed in a timely manner.
- **Longitudinal Perspective:** The program evaluates outcomes over extended time horizons, recognizing that language proficiency gains and their application in professional settings may not be immediately observable.
- **Contextual Sensitivity:** Evaluation methods and criteria are adapted to the specific language training context, including target language, proficiency level, learner population, and organizational mission.

1.3 Stakeholder Roles in Evaluation

Effective program evaluation requires the active participation of multiple stakeholder groups, each contributing unique perspectives, expertise, and data. The following table delineates the primary roles and responsibilities of each stakeholder group within the evaluation framework.

Stakeholder Group	Evaluation Roles & Responsibilities	Key Touchpoints
Program Leadership (Director, Deputy Director)	Set evaluation priorities and resource allocation; approve evaluation plans; review annual evaluation reports; make strategic decisions based on findings	Annual evaluation plan approval; Quarterly KPI review; Annual report sign-off
Evaluation Team (Evaluators, Analysts)	Design evaluation instruments; collect and analyze data; prepare reports; recommend improvements; maintain evaluation data systems	Instrument design; Data collection and analysis; Report preparation
Instructors	Administer formative assessments; provide learner performance data; participate in evaluation feedback sessions; implement instructional improvements	Ongoing formative assessment; End-of-course data submission; Feedback sessions
Learners	Complete satisfaction surveys; participate in focus groups; consent to longitudinal tracking; provide self-assessment data	Survey completion; Focus group participation; Self-assessment
Supervisors	Provide on-the-job performance assessments; verify behavioral change; validate training relevance; support longitudinal data collection	Post-training assessment; 6-month follow-up; Annual validation
Organizational Stakeholders (HR, Budget Office)	Align evaluation with organizational goals; fund evaluation activities; use findings for workforce planning; review cost-effectiveness analyses	Strategic alignment review; Budget approval; Workforce planning input

The evaluation framework relies on a distributed model of data collection, where stakeholders contribute information at different stages of the training and post-training process. The Evaluation Team serves as the central coordinating body, ensuring that data flows efficiently and that findings are synthesized into coherent, actionable recommendations. Stakeholder engagement is sustained throughregular communication, transparent reporting, and opportunities for input into the evaluation designprocess.

2. Evaluation Framework

2.1 Overview: Kirkpatrick's Four Levels Adapted for Language Training

The GLTP evaluation framework is built upon Kirkpatrick's four-level model of training evaluation, widely recognized as the industry standard for assessing training effectiveness. Originally developed by Donald Kirkpatrick in the 1950s and refined over subsequent decades, the model provides a structured approach to evaluating training at progressively deeper levels of impact. The GLTP has adapted this model to address the unique characteristics of language training, including the complexity of proficiency development, the significance of cultural competence, and the long time horizons required for measurable behavioral and organizational impact.

The four levels, adapted for the GLTP context, are as follows:

Evaluation Level	Focus Area	Description
Level 1: Reaction	Learner Satisfaction	Measures learner satisfaction with training delivery, content, materials, and instructor effectiveness. Captures immediate perceptions and identifies areas for course-level improvement.
Level 2: Learning	Knowledge & Skill Acquisition	Assesses the knowledge, skills, and language proficiency gained by learners as a direct result of training. Uses standardized proficiency assessments and criterion-referenced tests.
Level 3: Behavior	On-the-Job Application	Evaluates the extent to which learners apply acquired language skills in their professional roles. Relies on supervisor observations, performance evaluations, and self-reports.
Level 4: Results	Organizational Impact	Measures the broader organizational outcomes attributable to language training, including mission effectiveness, interagency communication, and return on investment.

Table 2.1 — Kirkpatrick's Four Levels Adapted for the GLTP

The four levels are not independent but interconnected. Positive reactions (Level 1) create favorable conditions for learning (Level 2), which in turn enables behavioral application (Level 3), ultimately contributing to organizational results (Level 4). However, the relationship is not strictly linear; strong organizational support can enhance behavioral application even when initial reactions are modest, and learning gains may not always translate to behavioral change without appropriate transfer conditions.

2.2 Level 1 — Reaction (Learner Satisfaction)

2.2.1 Definition

Level 1 evaluation measures the degree to which learners find the training favorable, engaging, and relevant to their professional needs. In the context of language training, reaction encompasses satisfaction with course content, instructional methods, materials and resources, technology platforms, instructor competence, schedule and pacing, and the overall learning environment. While reaction data alone cannot validate training effectiveness, they serve as a critical early indicator of potential issues and a prerequisite for meaningful engagement at subsequent evaluation levels.

2.2.2 Measurement Methods

- **End-of-Course Surveys:** Standardized Likert-scale questionnaires administered at the conclusion of each course module. Surveys capture satisfaction across multiple dimensions including content relevance, instructional quality, material usefulness, and overall experience.
- **Mid-Course Pulse Checks:** Brief surveys administered at the midpoint of longer courses to identify emerging issues before they affect completion or satisfaction.
- **Exit Interviews:** Semi-structured interviews with a sample of learners at program completion to gather qualitative insights on the training experience.
- **Digital Platform Analytics:** Automated tracking of learner engagement metrics including login frequency, time-on-task, resource access patterns, and completion rates for online components.

2.2.3 Data Sources

Primary data sources for Level 1 evaluation include: (1) learner-completed satisfaction survey instruments; (2) instructor observation logs documenting classroom dynamics; (3) learning management system (LMS) analytics; (4) informal feedback channels such as suggestion boxes and course forums; and (5) comparative benchmarking data from similar government training programs.

2.2.4 Frequency

Level 1 data are collected at the following intervals: end-of-course surveys for every course offering; mid-course pulse checks for courses exceeding 80 instructional hours; exit interviews for a minimum of 15% of program completers per cohort; and digital analytics continuously throughout the enrollment period.

2.2.5 Success Criteria

Criterion	Target	Data Source
Overall learner satisfaction	Mean score >= 4.0 on 5-point scale	End-of-course survey
Content relevance	>= 85% rating "Agree" or "Strongly Agree"	End-of-course survey
Instructor effectiveness	Mean score >= 4.2 on 5-point scale	End-of-course survey
Materials quality	>= 80% rating "Agree" or "Strongly Agree"	End-of-course survey

Criterion	Target	Data Source
Course completion rate	>= 90% of enrolled learners	Enrollment records
Would recommend to peers	>= 88% responding "Yes"	End-of-course survey

Table 2.2 — Level 1 Success Criteria

2.3 Level 2 — Learning (Knowledge and Skill Acquisition)

2.3.1 Definition

Level 2 evaluation assesses the extent to which learners acquire the intended knowledge, skills, attitudes, confidence, and commitment as a result of participating in the language training program. For the GLTP, this level focuses on measurable gains in language proficiency as defined by the Interagency Language Roundtable (ILR) scale, as well as the acquisition of domain-specific vocabulary, cultural competence, and professional communication skills relevant to government service.

2.3.2 Measurement Methods

- **Pre/Post Proficiency Testing:** Standardized ILR-scale assessments administered at program entry and completion. Tests cover speaking, listening, reading, and writing modalities.
- **Criterion-Referenced Assessments:** Module-specific tests that measure mastery of targeted competencies, including vocabulary, grammar, and functional language use.
- **Portfolio Assessment:** Collection of learner work products (written reports, recorded presentations, simulated diplomatic exchanges) evaluated against rubrics aligned with ILR descriptors.
- **Self-Assessment Instruments:** Can-Do statements aligned with the ACTFL/ILR proficiency guidelines that allow learners to rate their own perceived competence.
- **Oral Proficiency Interviews (OPI):** Certified OPI assessments conducted by trained testers to provide independent validation of speaking proficiency gains.

2.3.3 Data Sources

Data sources for Level 2 evaluation include: (1) institutional proficiency test scores (pre and post); (2) Defense Language Proficiency Test (DLPT) or equivalent results; (3) Oral Proficiency Interview ratings; (4) instructor-assigned course grades and competency ratings; (5) learner portfolio artifacts; and (6) self-assessment surveys. All proficiency data are linked to learner identifiers to enable longitudinal tracking.

2.3.4 Frequency

Pre-assessments are conducted at program entry; post-assessments at program completion. Criterion-referenced assessments occur at the end of each instructional module (typically every 4-6 weeks). Portfolio reviews are conducted at mid-program and end-of-program. OPI assessments are administered at entry, mid-program (for programs exceeding 6 months), and completion.

2.3.5 Success Criteria

Criterion	Target	Data Source
Proficiency gain (ILR levels)	>= 0.5 ILR level gain per 480 contact hours	Pre/Post proficiency tests
Module assessment pass rate	>= 80% of learners scoring >= 70%	Criterion-referenced tests
Speaking proficiency gain	>= 0.5 ILR level gain (OPI)	OPI ratings
Reading proficiency gain	>= 0.5 ILR level gain	DLPT or equivalent
Listening proficiency gain	>= 0.5 ILR level gain	DLPT or equivalent
Writing proficiency gain	>= 0.5 ILR level gain	Writing assessment
Portfolio quality rating	>= 80% rated "Competent" or above	Portfolio rubric evaluation

Table 2.3 — Level 2 Success Criteria

2.4 Level 3 — Behavior (On-the-Job Application)

2.4.1 Definition

Level 3 evaluation examines whether learners apply the knowledge and skills acquired during training to their professional roles and responsibilities. In the context of government language training, this means assessing whether trained personnel use their target language skills in workplace tasks such as diplomatic communication, document translation, interagency coordination, and public engagement. Level 3 evaluation is critical because learning gains that do not transfer to the workplace represent a failure of the training investment, regardless of how impressive the proficiency scores may be.

2.4.2 Measurement Methods

- **Supervisor Observation Reports:** Structured observation instruments completed by supervisors at 3, 6, and 12 months post-training. Reports assess frequency and quality of target language use in professional settings.
- **Performance Evaluation Integration:** Language skill utilization is incorporated into annual performance evaluations, providing a standardized mechanism for tracking behavioral application.
- **Self-Report Surveys:** Learner-completed surveys at 3, 6, and 12 months post-training that assess self-perceived language use, confidence, and barriers to application.
- **Work Product Analysis:** Review of professional outputs (reports, correspondence, meeting records) that require target language use, evaluated against quality rubrics.
- **Critical Incident Reports:** Documentation of specific workplace situations where target language skills were applied successfully or where skill gaps were identified.

2.4.3 Data Sources

Level 3 data sources include: (1) supervisor-completed behavioral observation checklists; (2) annual performance evaluation records; (3) learner self-report surveys; (4) work product samples; (5) critical incident reports; (6) peer assessments from colleagues who interact with the learner in target language contexts; and (7) language skill utilization tracking systems maintained by the learner's organizational unit.

2.4.4 Frequency

Post-training behavioral assessments are conducted at 3 months (initial transfer assessment), 6 months (sustained application assessment), and 12 months (long-term integration assessment). Self-report surveys follow the same schedule. Work product analysis is conducted semi-annually. Critical incident reporting is ongoing.

2.4.5 Success Criteria

Criterion	Target	Data Source
Target language use frequency	>= 3 times per week in professional tasks	Supervisor observation
Quality of language application	Rated "Effective" or above on 5-point scale	Supervisor observation
Self-reported confidence	>= 80% reporting "Confident" or "Very Confident"	Self-report survey
Work product quality	>= 75% rated "Acceptable" or above	Work product analysis
Skill maintenance at 12 months	>= 90% maintaining >= 80% of post-training proficiency	Follow-up assessment
Supervisor satisfaction with application	>= 85% rating "Satisfied" or "Very Satisfied"	Supervisor survey

Table 2.4 — Level 3 Success Criteria

2.5 Level 4 — Results (Organizational Impact)

2.5.1 Definition

Level 4 evaluation measures the extent to which the language training program contributes to organizational goals and mission effectiveness. This is the most challenging level of evaluation, as organizational outcomes are influenced by numerous factors beyond training. Nevertheless, the GLTP is committed to assessing results-level impact through a combination of quantitative metrics, qualitative indicators, and, where feasible, quasi-experimental designs that can attribute observed changes to training participation.

2.5.2 Measurement Methods

- **Mission Effectiveness Metrics:** Tracking of key mission indicators that are directly influenced by language capability, such as successful completion of language-dependent diplomatic engagements, quality and timeliness of translated intelligence products, and effectiveness of foreign liaison activities.
- **Cost-Effectiveness Analysis:** Comparison of training costs against measurable outcomes, including cost per proficiency level gained, cost per certified graduate, and return on investment calculations.
- **Organizational Readiness Assessment:** Evaluation of the organization's language readiness posture, including the percentage of positions filled by qualified language-capable personnel, compared against established requirements.
- **Stakeholder Outcome Surveys:** Surveys of organizational leaders and mission owners to assess perceived impact of language training on organizational performance.
- **Comparative Benchmarking:** Comparison of GLTP outcomes with other government language training programs and, where available, with private-sector benchmarks.

2.5.3 Data Sources

Level 4 data sources include: (1) organizational performance metrics maintained by mission elements; (2) financial and budget data from the program's accounting system; (3) personnel readiness databases; (4) stakeholder outcome surveys; (5) external benchmarking data; and (6) congressional and inspector general reports on language readiness.

2.5.4 Frequency

Level 4 evaluation is conducted on an annual cycle, with preliminary data reviews conducted semi-annually. Mission effectiveness metrics are tracked quarterly. Cost-effectiveness analyses are performed annually as part of the budget review process. Stakeholder outcome surveys are administered annually. Comprehensive organizational impact assessments are conducted every three years.

2.5.5 Success Criteria

Criterion	Target	Data Source
Language readiness fill rate	>= 85% of designated positions filled	Personnel readiness database
Cost per ILR level gain	<= $15,000 per 0.5 ILR level gain	Financial analysis
Mission success rate (language-dependent)	>= 90% rated "Successful"	Mission effectiveness metrics
Stakeholder perceived impact	>= 80% rating "Significant" or "High"	Stakeholder survey
Return on investment	Positive ROI within 3 years of training	Cost-effectiveness analysis
Interagency communication effectiveness	>= 85% rated "Effective" or above	Interagency survey

Table 2.5 — Level 4 Success Criteria

2.6 Integration Across Levels

The four evaluation levels are designed to function as an integrated system, with data from each level informing and validating findings at adjacent levels. For example, low satisfaction scores at Level 1 may help explain below-target proficiency gains at Level 2, while strong Level 2 outcomes that do not translate to Level 3 behavioral change may indicate deficiencies in transfer support mechanisms. The Evaluation Team is responsible for synthesizing data across all four levels to produce a holistic assessment of program effectiveness.

The integration framework operates on three dimensions: (1) **Vertical alignment** — ensuring that Level 1 insights feed into Level 2 analysis, Level 2 findings inform Level 3 interpretation, and Level 3 data support Level 4 conclusions; (2) **Temporal coherence** — aligning data collection schedules so that findings from earlier levels are available to contextualize later-level assessments; and (3) **Causal inference** — using the cumulative evidence from all four levels to build a logical chain from training inputs to organizational outcomes.

3. Evaluation Methods and Instruments

3.1 Learner Satisfaction Surveys

Learner satisfaction surveys are the primary instrument for Level 1 evaluation. The GLTP employs a suite of standardized survey instruments designed to capture satisfaction across multiple dimensions of the training experience. All surveys use a combination of Likert-scale items, multiple-choice questions, and open-ended response fields to enable both quantitative and qualitative analysis.

3.1.1 Survey Instruments

The GLTP uses three primary survey instruments:

1. **End-of-Course Satisfaction Survey (ECSS):** A 35-item instrument administered at the conclusion of every course. Items cover content relevance, instructional quality, materials and resources, technology support, learning environment, and overall satisfaction. The ECSS uses a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree) for 28 items, plus 4 multiple-choice items and 3 open-ended items.

2. **Mid-Course Pulse Survey (MCPS):** A 12-item brief survey administered at the midpoint of courses exceeding 80 instructional hours. The MCPS focuses on pacing, content difficulty, instructor responsiveness, and emerging concerns. All items use a 5-point Likert scale plus one open-ended item.

3. **Program Exit Survey (PES):** A 20-item survey administered to learners completing the full training program. The PES captures overall program satisfaction, perceived value, recommendation likelihood, and suggestions for improvement. It includes 15 Likert-scale items, 3 multiple-choice items, and 2 open-ended items.

3.1.2 Administration Procedures

Surveys are administered electronically through the program's Learning Management System (LMS). The following procedures ensure consistent and high-quality data collection:

- Surveys are activated automatically upon course completion or at designated midpoints.
- Learners receive an initial invitation and up to two reminder notices at 3-day intervals.
- The survey window remains open for 14 calendar days from the initial invitation.
- Responses are anonymous; no individual identifiers are linked to survey responses in reporting.
- A minimum response rate of 70% is required for results to be considered representative.
- Instructors do not have access to individual responses during the course delivery period.
- The Evaluation Team monitors response rates in real time and implements follow-up strategies as needed.

3.1.3 Sample ECSS Items

Item #	Item Text	Response Type
1	The course content was relevant to my professional responsibilities.	Likert 1-5
2	The instructor presented the material in a clear and organized manner.	Likert 1-5
3	The course materials (textbook, handouts, digital resources) supported my learning.	Likert 1-5
4	The pace of instruction was appropriate for the content complexity.	Likert 1-5
5	I had sufficient opportunities to practice speaking in the target language.	Likert 1-5
6	The technology platforms used were reliable and user-friendly.	Likert 1-5
7	I feel more confident in my ability to use the target language professionally.	Likert 1-5
8	What aspects of the course were most valuable to your learning?	Open-ended
9	What improvements would you suggest for this course?	Open-ended

Table 3.1 — Sample End-of-Course Satisfaction Survey Items

3.2 Proficiency Gain Analysis

Proficiency gain analysis is the cornerstone of Level 2 evaluation. The GLTP measures proficiency gains using a pre/post assessment design, where learners are tested at program entry and completion using standardized instruments aligned with the ILR proficiency scale. The analysis encompasses both overall proficiency gains and skill-specific gains across the four modalities: speaking, listening, reading, and writing.

3.2.1 Assessment Instruments

- **Defense Language Proficiency Test (DLPT):** The standardized test used across government agencies for reading and listening proficiency. Administered at program entry and completion.
- **Oral Proficiency Interview (OPI):** A standardized interview-based assessment of speaking proficiency conducted by certified OPI testers. Administered at entry, mid-program, and completion.
- **Writing Proficiency Test (WPT):** A standardized assessment of writing proficiency using prompted writing samples evaluated against ILR descriptors.
- **GLTP Internal Assessments:** Criterion-referenced tests developed by GLTP instructors aligned with course objectives and ILR descriptors. These provide finer-grained measurement between ILR levels.

3.2.2 Gain Score Calculation

Proficiency gains are calculated as the difference between post-training and pre-training assessment scores. For ILR-based assessments, gains are expressed in ILR levels (e.g., a gain from ILR 1+ to ILR

2+ represents a 1.0 ILR level gain). For criterion-referenced internal assessments, gains are expressedas percentage point improvements on a 100-point scale. Both raw gain scores and adjusted gain scores(controlling for initial proficiency level and aptitude) are reported. Effect sizes (Cohen's d) arecalculated to provide a standardized measure of the magnitude of proficiency gains.

3.2.3 Analysis Procedures

1. **Data Cleaning:** Verify completeness and accuracy of all test scores. Flag and investigate outliers (scores > 3 SD from the mean).
2. **Descriptive Statistics:** Calculate means, standard deviations, and ranges for pre- and post-test scores by modality, language, and cohort.
3. **Gain Score Analysis:** Compute individual and group gain scores. Test for statistical significance using paired-samples t-tests.
4. **Effect Size Calculation:** Compute Cohen's d for all gain comparisons. Interpret using conventional thresholds ($d = 0.2$ small, 0.5 medium, 0.8 large).
5. **Subgroup Analysis:** Examine gains by language category (Category I-IV), initial proficiency level, and delivery modality (classroom, online, blended).
6. **Normative Comparison:** Compare observed gains against institutional norms and DLPT historical data.

3.3 On-the-Job Performance Assessment

On-the-job performance assessment is the primary method for Level 3 evaluation. It measures the extent to which learners transfer their language skills to workplace applications. The GLTP uses a multi-method approach that combines supervisor evaluations, self-assessments, and objective performance data.

3.3.1 Supervisor Feedback Instruments

The GLTP employs two supervisor feedback instruments:

1. **Post-Training Behavioral Observation Checklist (BOC):** A 25-item checklist completed by supervisors at 3, 6, and 12 months post-training. Items assess frequency and quality of target language use in specific workplace tasks. Responses use a 5-point frequency scale (1 = Never, 5 = Always) and a 5-point quality scale (1 = Unacceptable, 5 = Exceptional).
2. **Annual Language Skill Utilization Report (ALSR):** A narrative report completed by supervisors as part of the annual performance evaluation process. The ALSR documents specific instances of target language use, evaluates the quality of application, and identifies areas for continued development.

3.3.2 Administration Procedures

- The Evaluation Team distributes BOC instruments to supervisors via secure email at the designated post-training intervals (3, 6, and 12 months).
- Supervisors are provided with a brief orientation guide explaining the instrument and rating criteria.
- Learners are responsible for notifying their supervisors when they complete training and are due for a behavioral assessment.
- The Evaluation Team follows up with non-responding supervisors at 2-week intervals for up to 6 weeks.
- A minimum response rate of 60% is required for BOC results to be included in program-level reporting.
- The ALSR is integrated into the organization's existing annual performance evaluation cycle and is coordinated with the HR office.

3.4 Cost-Effectiveness Analysis

Cost-effectiveness analysis (CEA) provides a systematic framework for evaluating the economic efficiency of the language training program. The GLTP conducts CEA annually as part of the budget review process and as a key component of Level 4 evaluation. The analysis compares the total cost of training delivery against measurable outcomes to determine the cost per unit of proficiency gain and the overall return on investment.

3.4.1 Cost Categories

Cost Category	Components	Data Source
Direct Instructional Costs	Instructor salaries and benefits; adjunct/contract instructor fees	Accounting system
Curriculum Development	Materials development, revision, and copyright licensing	Project accounting
Technology Costs	LMS licensing, software subscriptions, hardware maintenance	IT budget
Facility Costs	Classroom space, equipment, utilities (allocated share)	Facilities management
Learner Opportunity Costs	Estimated value of learner time in training (not in productive duty)	Salary data x hours
Assessment Costs	Testing materials, OPI administration, scoring services	Assessment budget
Administrative Overhead	Program management, evaluation staff, support services	Administrative budget

Table 3.2 — Cost Categories for Cost-Effectiveness Analysis

3.4.2 Cost-Effectiveness Metrics

- **Cost per Proficiency Gain:** Total program cost divided by the aggregate ILR level gains across all completers. Expressed as dollars per 0.5 ILR level gain.
- **Cost per Certified Graduate:** Total program cost divided by the number of learners who achieve the target certification or proficiency standard.
- **Return on Investment (ROI):** Estimated monetary value of training outcomes (e.g., reduced need for external translation services, improved mission effectiveness) minus total training costs, divided by total training costs, expressed as a percentage.
- **Cost-Effectiveness Ratio:** Ratio of training costs to a composite outcome index that combines proficiency gains, certification rates, and on-the-job performance scores.

3.5 Longitudinal Tracking

Longitudinal tracking is essential for understanding the long-term impact of language training and the sustainability of proficiency gains. The GLTP maintains a longitudinal database that tracks learner outcomes for a minimum of five years following program completion. This database integrates data from all four evaluation levels and enables time-series analyses of proficiency maintenance, career progression, and organizational contribution.

3.5.1 Tracking Components

- **Proficiency Maintenance Tracking:** Annual or biennial proficiency re-assessment of program graduates to monitor skill retention and attrition.
- **Career Progression Monitoring:** Tracking of position assignments, promotions, and language-designated billet assignments for program graduates.
- **Continuing Education Records:** Documentation of post-program language sustainment activities, including immersion programs, online courses, and self-study.
- **Organizational Contribution Metrics:** Tracking of mission assignments, language-dependent task completions, and professional achievements attributable to language capability.

3.5.2 Data Management

All longitudinal data are stored in the GLTP Evaluation Database, a secure, access-controlled system maintained by the Evaluation Team. Data are linked through a unique learner identifier that preserves anonymity while enabling individual-level tracking. Data retention policies comply with applicable records management regulations, with a minimum retention period of 10 years for evaluation records. Annual data quality audits are conducted to ensure completeness and accuracy of the longitudinal dataset.

4. Data Collection and Analysis

4.1 Data Collection Schedule

The data collection schedule is designed to ensure that evaluation data are collected systematically and at appropriate intervals across the training lifecycle. The following master schedule specifies the timing, responsible party, and instruments for each data collection event.

Time Point	Instruments / Data Collected	Responsible Party	Population
Program Entry	Pre-assessment battery (DLPT, OPI, WPT, Self-Assessment)	Evaluation Team	Every cohort
Mid-Course (>80 hrs)	Mid-Course Pulse Survey (MCPS) Criterion-referenced test	Evaluation Team Instructor	Every course >80 hrs
Program Completion	Post-assessment battery ECSS, PES Portfolio review	Evaluation Team Evaluation Team Instructor	Every cohort
3 Months Post	BOC (supervisor) Self-report survey	Supervisor Learner	Every graduate
6 Months Post	BOC (supervisor) Self-report survey Work product analysis	Supervisor Learner Evaluation Team	Every graduate
12 Months Post	BOC (supervisor) Self-report survey Proficiency re-test	Supervisor Learner Evaluation Team	Every graduate
Annually	ALSR (supervisor) Stakeholder outcome survey Cost-effectiveness analysis KPI compilation	Supervisor Eval. Team Eval. Team Eval. Team	All active graduates
Triennially	Comprehensive organizational impact assessment External benchmarking	Evaluation Team Evaluation Team	All programs

Table 4.1 — Master Data Collection Schedule

4.2 Quantitative Analysis Methods

Quantitative analysis provides the empirical foundation for program evaluation. The GLTP employs a range of statistical methods to analyze proficiency data, satisfaction scores, performance metrics, and cost data. All quantitative analyses are conducted using validated statistical software, and results are reviewed by at least two qualified analysts before inclusion in official reports.

4.2.1 Gain Score Analysis

Gain scores are the primary quantitative metric for Level 2 evaluation. A gain score is computed as the difference between post-test and pre-test scores for each learner. The GLTP calculates gain scores at the individual, cohort, and program levels. Statistical significance of group-level gains is assessed using paired-samples t-tests (for normally distributed data) or Wilcoxon signed-rank tests (for non-normaldistributions). Confidence intervals (95%) are reported for all mean gain scores. The GLTP alsocalculates percentage of learners achieving the minimum expected gain (0.5 ILR levels per 480 contacthours) as a practical indicator of program effectiveness.

4.2.2 Effect Size Calculation

Effect sizes complement significance testing by providing a measure of the practical magnitude of observed differences. The GLTP calculates Cohen's d for all pre/post comparisons using the formula: d = (M_post - M_pre) / SD_pooled, where SD_pooled is the pooled standard deviation. Effect sizes are interpreted as follows: d = 0.2 indicates a small effect, d = 0.5 a medium effect, and d = 0.8 a large effect. In addition, the GLTP reports the percentage of learners exceeding the minimal detectable change (MDC) for each assessment, providing a clinically meaningful threshold for proficiency gains.

4.2.3 Pass Rate Analysis

Pass rates are calculated for all criterion-referenced assessments and certification examinations. The GLTP reports raw pass rates (percentage of learners meeting the passing threshold) and adjusted pass rates (controlling for initial proficiency level and language category). Pass rate trends are analyzed over time using logistic regression to identify statistically significant changes. The program also computes the relative pass rate (observed pass rate divided by the expected pass rate based on historical norms) as a benchmarking metric.

4.2.4 Subgroup Comparisons

The GLTP conducts subgroup analyses to identify differential outcomes across learner populations. Standard subgroup variables include: language category (I, II, III, IV), initial proficiency level, delivery modality (classroom, online, blended), instructor experience level, and learner demographic characteristics. Comparisons are conducted using analysis of variance (ANOVA) for continuous outcomes and chi-square tests for categorical outcomes. Effect sizes and confidence intervals are reported for all subgroup comparisons. The Evaluation Team applies a Bonferroni correction for multiple comparisons to control the family-wise error rate.

4.3 Qualitative Analysis Methods

Qualitative analysis provides depth and context that complement quantitative findings. The GLTP employs rigorous qualitative methods to explore learner experiences, identify underlying mechanisms of change, and generate rich descriptions of program impact.

4.3.1 Focus Groups

Focus groups are conducted with learner cohorts at program completion and at 6 months post-training. Each focus group includes 6-10 participants and is facilitated by a trained evaluator using a semi-structured protocol. Focus group discussions explore themes such as: perceived value of trainingcomponents, barriers to on-the-job application, suggestions for curriculum improvement, andexperiences with the transfer of learning to the workplace. Sessions are audio-recorded (with consent)and transcribed for analysis.

4.3.2 Individual Interviews

Individual interviews are conducted with a purposive sample of key informants, including high-performing graduates, learners who struggled with the program, instructors, and supervisors. Interviews use a semi-structured format that allows for in-depth exploration of individual experiences and perspectives. The GLTP conducts a minimum of 20 individual interviews per evaluation cycle. Interview data are analyzed using thematic analysis, with independent coding by at least two analysts and reconciliation of coding discrepancies through consensus.

4.3.3 Case Studies

Case studies provide detailed, contextualized accounts of how language training translates into professional capability. The GLTP develops 3-5 case studies per evaluation cycle, selecting cases that represent diverse outcomes (highly successful, moderately successful, and challenging). Each case study documents the learner's training trajectory, proficiency development, workplace application, and organizational contribution. Case studies draw on multiple data sources, including test scores, survey responses, supervisor reports, work products, and interview data, to construct a comprehensive narrative.

4.4 Data Quality Standards

The integrity of evaluation findings depends on the quality of the underlying data. The GLTP maintains rigorous data quality standards that apply to all phases of data collection, processing, and analysis.

Quality Dimension	Standard	Verification Method
Completeness	>= 95% of required data fields populated for each record	Automated validation checks; manual audit
Accuracy	>= 98% of data values verified against source documents	Random audit of 10% of records per quarter
Timeliness	Data entered within 5 business days of collection	LMS and database timestamps
Consistency	< 2% discrepancy rate across linked data sources	Cross-reference validation
Validity	All instruments meet minimum psychometric standards	Annual reliability and validity analysis
Security	All PII encrypted at rest and in transit	Annual security audit; continuous monitoring

Table 4.2 — Data Quality Standards

4.5 Reporting Templates

Standardized reporting templates ensure consistency and comparability across evaluation reports. The GLTP uses the following templates, each designed for a specific purpose and audience:

- **Course Evaluation Summary (CES):** A 2-page template summarizing Level 1 and Level 2 data for an individual course offering. Includes mean satisfaction scores, proficiency gain statistics, and instructor-specific metrics. Distributed to instructors and course coordinators.
- **Cohort Evaluation Report (CER):** A 5-8 page template providing a comprehensive assessment of a training cohort's outcomes across all four levels. Includes statistical analyses, qualitative findings, and improvement recommendations. Distributed to program leadership and stakeholders.
- **Annual Program Evaluation Report (APER):** A 20-30 page template synthesizing evaluation findings across all cohorts and courses for the fiscal year. Includes trend analyses, KPI performance, cost-effectiveness results, and strategic recommendations. Distributed to program leadership, stakeholders, and oversight bodies.
- **Executive Summary Brief (ESB):** A 1-2 page high-level summary of key findings and recommendations for senior leadership. Uses visual dashboards and bullet-point highlights.
- **Data Quality Report (DQR):** A quarterly report documenting data quality metrics, identified issues, and corrective actions. Distributed to the Evaluation Team and program leadership.

5. Continuous Improvement Process

5.1 Plan-Do-Study-Act (PDSA) Cycle for Program Improvement

The GLTP employs the Plan-Do-Study-Act (PDSA) cycle as its primary framework for continuous improvement. This iterative, four-stage model provides a structured approach to identifying improvement opportunities, testing changes on a small scale, evaluating results, and implementing successful changes broadly. The PDSA cycle is applied at both the course level (for instructional improvements) and the program level (for systemic improvements).

5.1.1 Plan

In the Plan phase, the improvement team identifies a specific aspect of the program that requires improvement, establishes measurable objectives, and designs an intervention. Planning activities include: (a) analyzing evaluation data to identify areas of underperformance; (b) conducting root cause analysis to determine contributing factors; (c) reviewing research evidence and best practices for potential solutions; (d) defining specific, measurable improvement targets; (e) designing the intervention with clear implementation steps; and (f) establishing the data collection plan for evaluating the intervention.

5.1.2 Do

In the Do phase, the planned intervention is implemented on a small scale (typically with one or two course sections or a single cohort). Key activities include: (a) communicating the change to affected stakeholders; (b) providing necessary training or resources for implementation; (c) implementing the intervention as designed; (d) documenting any deviations from the plan; (e) collecting data during implementation; and (f) recording observations and informal feedback from participants.

5.1.3 Study

In the Study phase, the improvement team analyzes the data collected during the Do phase to determine whether the intervention achieved its intended objectives. Activities include: (a) comparing outcomes against the established targets; (b) analyzing both quantitative and qualitative data; (c) identifying unexpected outcomes (positive or negative); (d) assessing the feasibility and sustainability of the intervention; (e) gathering stakeholder feedback on the change; and (f) documenting lessons learned.

5.1.4 Act

In the Act phase, the improvement team makes decisions about the intervention based on the Study phase findings. Three possible courses of action are: (a) **Adopt** — if the intervention achieved its targets and is sustainable, implement it broadly across the program; (b) **Adapt** — if the intervention showed promise but fell short of targets, modify the approach and conduct another PDSA cycle; (c) **Abandon** — if the intervention did not achieve its targets and no feasible modifications are apparent, discontinue the approach and return to the Plan phase with an alternative strategy. All Act phase decisions are documented and archived for future reference.

Phase	Key Activities	Data Inputs	Outputs
Plan	Identify improvement opportunity Set measurable objectives Design intervention Define data collection plan	Evaluation data analysis Root cause analysis Stakeholder input	Improvement plan document Data collection plan
Do	Implement intervention (small scale) Document deviations Collect data Record observations	Implementation logs Participant feedback Observation notes	Implementation report Preliminary data
Study	Analyze outcomes vs. targets Assess unexpected results Evaluate feasibility Gather stakeholder feedback	Intervention data Stakeholder interviews Cost analysis	Analysis report Lessons learned
Act	Decide: Adopt, Adapt, or Abandon Document decision rationale Plan next cycle if adapting Implement broadly if adopting	Study phase findings Stakeholder consensus Resource assessment	Decision document Implementation plan Next PDSA cycle plan

Table 5.1 — PDSA Cycle Framework

5.2 Root Cause Analysis Procedures

When evaluation data indicate underperformance, the GLTP conducts a systematic root cause analysis (RCA) to identify the underlying factors contributing to the problem. RCA is a prerequisite for effective corrective action planning, as addressing symptoms rather than root causes typically leads to ineffective or unsustainable improvements.

5.2.1 RCA Methods

- **5-Why Analysis:** A structured questioning technique that iteratively asks "Why?" to drill down from a surface-level problem to its root cause. The GLTP uses this method for relatively straightforward issues with likely single root causes.
- **Fishbone (Ishikawa) Diagram:** A visual tool that categorizes potential causes into major categories (People, Process, Materials, Technology, Environment, Measurement). The GLTP uses this method for complex problems with multiple potential contributing factors.
- **Pareto Analysis:** A statistical technique that identifies the vital few causes that account for the majority of the problem. The GLTP uses Pareto analysis when multiple root causes are identified and prioritization is needed.
- **Failure Mode and Effects Analysis (FMEA):** A systematic method for identifying potential failure modes in training processes, assessing their severity and likelihood, and prioritizing

corrective actions. The GLTP uses FMEA for proactive risk identification.

5.2.2 RCA Process

1. Problem Definition: Clearly state the performance gap using specific, measurable terms.

2. Data Gathering: Collect relevant evaluation data, process documentation, and stakeholder perspectives.

3. Cause Identification: Apply selected RCA method(s) to identify potential root causes.

4. Cause Verification: Test identified root causes against available data to confirm or refute.

5. Root Cause Confirmation: Confirm the root cause(s) through stakeholder review and consensus.

6. Documentation: Record the RCA process, findings, and confirmed root causes in the Improvement Log.

5.3 Corrective Action Planning

Corrective action planning translates root cause findings into specific, implementable improvement actions. Each corrective action plan includes the following elements:

Plan Element	Description
Problem Statement	Clear description of the identified performance gap
Root Cause(s)	Confirmed root cause(s) from the RCA process
Corrective Action	Specific action(s) to address the root cause
Responsible Party	Individual or team accountable for implementation
Timeline	Start date, milestones, and completion date
Resources Required	Personnel, budget, materials, and technology needed
Success Metrics	Measurable indicators that the action has achieved its objective
Monitoring Plan	How and when progress will be tracked
Escalation Criteria	Conditions under which the issue will be escalated to higher authority

Table 5.2 — Corrective Action Plan Template

5.4 Implementation Monitoring

Corrective actions are monitored through a structured implementation tracking system. The Evaluation Team conducts monthly reviews of all active corrective actions, assessing progress against milestones, identifying barriers, and recommending adjustments. Implementation status is classified using a four-tier system:

- **On Track (Green):** Implementation is proceeding according to plan, with milestones met on schedule.
- **At Risk (Yellow):** Minor delays or obstacles have been identified, but corrective measures are in place to recover the schedule.
- **Off Track (Red):** Significant delays or barriers are jeopardizing successful implementation. Escalation and intervention are required.
- **Completed (Blue):** The corrective action has been fully implemented and verified.

5.5 Re-Evaluation Protocols

Following the implementation of corrective actions, the GLTP conducts re-evaluation to verify that the actions achieved their intended effect. Re-evaluation protocols specify: (a) the evaluation design (typically a pre/post comparison or a comparison between affected and unaffected cohorts); (b) the outcome measures and data sources; (c) the timeline for re-evaluation (typically 3-6 months after full implementation); (d) the criteria for determining success; and (e) the actions to be taken if the re-evaluation shows insufficient improvement. Re-evaluation results are documented in the Improvement Log and communicated to all relevant stakeholders.

6. Program-Level Key Performance Indicators

6.1 KPI Overview

Key Performance Indicators (KPIs) provide a concise, quantitative summary of program performance. The GLTP has established seven core KPIs that are tracked continuously and reported to program leadership on a quarterly basis. Each KPI has a defined target, measurement method, data source, and reporting frequency. KPI performance is reviewed by program leadership quarterly, and trends are analyzed annually as part of the Annual Program Evaluation Report.

6.2 KPI Definitions, Targets, and Measurement Methods

KPI	Definition	Target	Data Source	Frequency
KPI-1: Proficiency Gain Rate	Percentage of learners achieving >= 0.5 ILR level gain per 480 contact hours	>= 75%	Pre/Post proficiency test scores	Per cohort; aggregated quarterly
KPI-2: Certification Pass Rate	Percentage of learners passing the target certification exam on first attempt	>= 70%	Certification exam results	Per exam administration; aggregated quarterly
KPI-3: Time-to-Proficiency	Median number of instructional hours required to achieve target ILR level	<= 720 hours for Category I-II <= 1320 hours for Category III-IV	Enrollment and proficiency records	Per cohort; aggregated annually
KPI-4: Learner Satisfaction	Mean score on the End-of-Course Satisfaction Survey (5-point scale)	>= 4.0	ECSS results	Per course; aggregated quarterly
KPI-5: Instructor Effectiveness	Mean instructor rating on learner satisfaction surveys (5-point scale)	>= 4.2	ECSS instructor-specific items	Per course; aggregated quarterly
KPI-6: Cost per Proficiency Gain	Total program cost divided by aggregate ILR level gains across completers	<= $15,000 per 0.5 ILR level gain	Financial records; proficiency data	Annually
KPI-7: Retention Rate	Percentage of program graduates who remain in language-designated positions 12 months post-training	>= 80%	Personnel records; supervisor reports	Annually

Table 6.1 — Program-Level Key Performance Indicators

6.3 KPI-1: Proficiency Gain Rate — Detailed Specification

The Proficiency Gain Rate is the most fundamental indicator of program effectiveness, measuring the proportion of learners who achieve meaningful proficiency improvements. A gain of 0.5 ILR levels represents the minimum practically significant improvement, corresponding to a discernible step up in functional language capability.

Calculation Method

Proficiency Gain Rate = (Number of learners achieving >= 0.5 ILR level gain) / (Total number of learners with matched pre/post scores) x 100. Gains are calculated separately for each modality (speaking, listening, reading, writing) and then aggregated. Learners who withdraw before completing the post-assessment are excluded from the calculation but are tracked separately for attrition analysis.

Disaggregation

The Proficiency Gain Rate is reported at the aggregate level and disaggregated by: language category (I, II, III, IV); initial proficiency level (0, 0+, 1, 1+); delivery modality (classroom, online, blended); and instructor experience (< 2 years, 2-5 years, > 5 years). Disaggregation enables identification of differential outcomes and targeted improvement efforts.

6.4 KPI-2: Certification Pass Rate — Detailed Specification

The Certification Pass Rate measures the effectiveness of the program in preparing learners to meet external certification standards. For government language programs, this typically refers to the Defense Language Proficiency Test (DLPT) or equivalent agency-specific certification examinations.

Calculation Method

Certification Pass Rate = (Number of learners passing certification exam on first attempt) / (Total number of learners attempting certification) x 100. First-attempt pass rate is used as the primary metric because it reflects the program's ability to prepare learners to standard. Overall pass rate (including retests) is reported as a supplementary metric.

6.5 KPI-3: Time-to-Proficiency — Detailed Specification

Time-to-Proficiency measures the efficiency of the training program in producing qualified language professionals. This KPI is expressed as the median number of instructional hours required for learners to achieve the target ILR proficiency level, broken down by language category.

Calculation Method

Time-to-Proficiency = Median instructional hours from program entry to the date of the first assessment showing achievement of the target ILR level. The median is used rather than the mean to reduce the influence of outliers. Separate calculations are performed for each language category to account for differences in language difficulty.

6.6 KPI-4: Learner Satisfaction — Detailed Specification

Learner satisfaction is measured through the End-of-Course Satisfaction Survey (ECSS). The overall satisfaction score is the mean of all Likert-scale items on the ECSS. This KPI provides a continuous pulse on learner perceptions of program quality and serves as an early warning indicator for potential issues.

Calculation Method

Learner Satisfaction Score = Mean of all Likert-scale items on the ECSS, averaged across all respondents within the reporting period. Course-level scores are weighted by enrollment size when aggregating to the program level.

6.7 KPI-5: Instructor Effectiveness — Detailed Specification

Instructor effectiveness is measured through the instructor-specific items on the ECSS. This KPI provides feedback on instructional quality and identifies instructors who may benefit from professional development or mentoring.

Calculation Method

Instructor Effectiveness Rating = Mean of instructor-specific ECSS items (typically items 2, 5, and 7), averaged across all respondents for each instructor. Ratings are aggregated across all courses taught by the instructor within the reporting period. A minimum of 10 responses per instructor is required for the rating to be considered reliable.

6.8 KPI-6: Cost per Proficiency Gain — Detailed Specification

Cost per Proficiency Gain provides a measure of the economic efficiency of the training program. This KPI is calculated annually and is a key input to budget planning and resource allocation decisions.

Calculation Method

Cost per 0.5 ILR Level Gain = Total Annual Program Cost / (Total ILR Level Gains Across All Completers / 0.5). Total program cost includes all direct instructional costs, curriculum development, technology, facilities, assessment, and administrative overhead as defined in Section 3.4. Learner opportunity costs are reported separately and are not included in the primary KPI calculation.

6.9 KPI-7: Retention Rate — Detailed Specification

The Retention Rate measures the extent to which program graduates remain in positions that utilize their language skills. High retention rates indicate that the training investment is being sustained through continued organizational application of the acquired skills.

Calculation Method

Retention Rate = (Number of graduates in language-designated positions at 12 months post-training) / (Total number of graduates) x 100. A language-designated position is defined as one that requires or significantly benefits from target language proficiency as documented in the position description.

6.10 KPI Performance Dashboard

KPI performance is monitored through a dashboard that displays current values, targets, trends, and status indicators for each KPI. The dashboard is updated quarterly and is accessible to program leadership through the GLTP's secure web portal. Status indicators follow a three-tier color scheme:

Status	Criteria	Action
Green (On Target)	KPI value meets or exceeds the target	Continue monitoring; no action required
Yellow (Approaching Target)	KPI value is within 10% of the target but below it	Investigate contributing factors; develop contingency plan
Red (Below Target)	KPI value is more than 10% below the target	Initiate root cause analysis; develop corrective action plan

Table 6.2 — KPI Status Indicator Framework

7. Reporting and Communication

7.1 Reporting Frequency and Formats by Audience

Evaluation findings are communicated to different stakeholder audiences through tailored reports and communication channels. The frequency, format, and content of reports are calibrated to the information needs and decision-making responsibilities of each audience.

Audience	Reports	Format	Primary Use
Program Leadership (Director, Deputy)	Quarterly KPI Report Annual Program Evaluation Report Executive Summary Brief	Dashboard + narrative 20-30 page report 1-2 page brief	Strategic decision-making Resource allocation Policy development
Instructors	Course Evaluation Summary Instructor Effectiveness Report Best Practices Bulletin	2-page summary Individual report Periodic bulletin	Instructional improvement Professional development Peer learning
Organizational Stakeholders (HR, Budget Office)	Annual Program Evaluation Report Cost-Effectiveness Report Workforce Readiness Report	20-30 page report Specialized report Data summary	Workforce planning Budget justification Readiness assessment
Learners	Course Evaluation Summary Program Outcomes Overview	Brief summary Infographic	Informed participation Program credibility
Oversight Bodies (IG, Congressional)	Annual Program Evaluation Report Special Reviews (as requested)	Full report Custom reports	Accountability Compliance verification

Table 7.1 — Reporting Matrix by Audience

7.2 Annual Program Evaluation Report Structure

The Annual Program Evaluation Report (APER) is the flagship evaluation document, providing a comprehensive assessment of program performance over the fiscal year. The APER follows a standardized structure to ensure completeness and comparability across years.

Section	Title	Content	Pages
1	Executive Summary	High-level overview of key findings, KPI performance, and strategic recommendations	2-3
2	Program Overview	Enrollment data, course offerings, staffing, and budget summary for the year	2-3
3	Level 1: Learner Satisfaction	Satisfaction survey results, trends, and comparative analysis	3-4

Section	Title	Content	Pages
4	Level 2: Learning Outcomes	Proficiency gain analysis, certification results, and subgroup comparisons	4-5
5	Level 3: Behavioral Application	On-the-job performance data, supervisor feedback, and transfer analysis	3-4
6	Level 4: Organizational Impact	Mission effectiveness, cost-effectiveness analysis, and ROI	3-4
7	KPI Performance Summary	Dashboard of all KPIs with current values, targets, trends, and status	2-3
8	Continuous Improvement Activities	Summary of PDSA cycles, corrective actions, and improvement outcomes	2-3
9	Recommendations	Prioritized recommendations for program improvement with action plans	2-3
10	Appendices	Detailed data tables, methodology notes, and supplementary analyses	5-10

Table 7.2 — Annual Program Evaluation Report Structure

7.3 Dashboard Design for Real-Time Monitoring

The GLTP maintains a web-based evaluation dashboard that provides program leadership with real-time access to key evaluation metrics. The dashboard is designed to support rapid identification of emerging issues and data-driven decision-making. The dashboard comprises the following components:

7.3.1 KPI Scorecard

The KPI Scorecard displays current values for all seven program-level KPIs, along with target values, trend arrows (improving, stable, declining), and color-coded status indicators (Green/Yellow/Red). The scorecard is updated in real time as new data are entered into the evaluation database. Clicking on any KPI opens a detailed view showing historical trends, disaggregated data, and explanatory notes.

7.3.2 Enrollment and Completion Tracker

This component displays real-time enrollment figures, course fill rates, withdrawal rates, and completion rates by course, language, and delivery modality. It includes a Gantt-style timeline showing the status of all active courses and upcoming cohorts. Alerts are generated when withdrawal rates exceed threshold values or when courses are at risk of low enrollment.

7.3.3 Proficiency Gain Visualizer

The Proficiency Gain Visualizer provides interactive charts showing pre/post proficiency distributions, gain score histograms, and trend lines for each modality and language. Users can filter by language category, initial proficiency level, delivery modality, and time period. The visualizer includes statistical annotations showing mean gains, effect sizes, and significance levels.

7.3.4 Satisfaction Trend Monitor

This component displays rolling averages of learner satisfaction scores across key dimensions (content, instructor, materials, technology, overall) with trend lines and control limits. When satisfaction scores drop below the lower control limit, an automated alert is sent to program leadership, triggering a review.

7.3.5 Improvement Tracker

The Improvement Tracker displays the status of all active corrective actions and PDSA cycles. It includes a visual timeline of improvement activities, progress bars for each corrective action, and a summary of improvement outcomes. The tracker supports filtering by status (On Track, At Risk, Off Track, Completed) and by responsible party.

7.3.6 Dashboard Access and Security

Dashboard access is role-based, with the following access levels: (a) **Full Access** — Program Director and Deputy Director can view all components and all data; (b) **Standard Access** — Evaluation Team members and program managers can view all components for their assigned programs; (c) **Limited Access** — Instructors can view course-level data and their own effectiveness ratings; (d) **Summary Access** — Organizational stakeholders can view summary dashboards and published reports. All dashboard access is logged and audited quarterly.

8. Appendices

Appendix A: Sample Learner Satisfaction Survey Instrument

The following is the complete End-of-Course Satisfaction Survey (ECSS) instrument. This survey is administered electronically at the conclusion of every course offering.

End-of-Course Satisfaction Survey (ECSS)

Instructions: Please rate each statement below using the following scale: 1 = Strongly Disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, 5 = Strongly Agree. Your responses are anonymous and will be used to improve the quality of our training programs.

Section: Course Content

1. The course objectives were clearly stated at the beginning. [1] [2] [3] [4] [5]
2. The course content was relevant to my professional responsibilities. [1] [2] [3] [4] [5]
3. The course content was appropriate for my proficiency level. [1] [2] [3] [4] [5]
4. The amount of material covered was manageable within the allotted time. [1] [2] [3] [4] [5]
5. The course provided sufficient opportunities for practical application. [1] [2] [3] [4] [5]

Section: Instructional Quality

1. The instructor demonstrated thorough knowledge of the subject matter. [1] [2] [3] [4] [5]
2. The instructor presented the material in a clear and organized manner. [1] [2] [3] [4] [5]
3. The instructor was responsive to questions and learner needs. [1] [2] [3] [4] [5]
4. The instructor used a variety of teaching methods to engage learners. [1] [2] [3] [4] [5]
5. The instructor provided timely and constructive feedback on my performance. [1] [2] [3] [4] [5]

Section: Materials and Resources

1. The textbook and supplementary materials supported my learning effectively. [1] [2] [3] [4] [5]
2. The audio/video resources were of high quality and relevant. [1] [2] [3] [4] [5]
3. The digital platform (LMS) was user-friendly and reliable. [1] [2] [3] [4] [5]
4. Sufficient practice exercises and self-study materials were provided. [1] [2] [3] [4] [5]

Section: Learning Environment

1. The physical (or virtual) learning environment was conducive to learning. [1] [2] [3] [4] [5]
2. The class size was appropriate for effective language instruction. [1] [2] [3] [4] [5]
3. The schedule and pacing of the course were well-organized. [1] [2] [3] [4] [5]
4. I felt comfortable participating and making mistakes in this course. [1] [2] [3] [4] [5]

Section: Overall Satisfaction

1. Overall, I am satisfied with this course. [1] [2] [3] [4] [5]
2. I feel more confident in my ability to use the target language professionally. [1] [2] [3] [4] [5]
3. I would recommend this course to a colleague. [1] [2] [3] [4] [5]
4. This course met my expectations for language training. [1] [2] [3] [4] [5]

Open-Ended Questions

34. What aspects of this course were most valuable to your learning?

35. What improvements would you suggest for this course?

36. Please share any additional comments about your learning experience.

Appendix B: Supervisor Behavioral Observation Checklist

The following instrument is completed by supervisors at 3, 6, and 12 months post-training to assess on-the-job application of language skills.

Behavioral Observation Checklist (BOC)

Instructions: For each item below, rate the learner's frequency and quality of target language use using the scales provided. Base your ratings on direct observation or verified reports.

Frequency Scale: 1 = Never, 2 = Rarely, 3 = Sometimes, 4 = Often, 5 = Always

Quality Scale: 1 = Unacceptable, 2 = Below Standard, 3 = Meets Standard, 4 = Exceeds Standard, 5 = Exceptional

Item #	Behavior	Rating
1	Uses target language in routine professional communications	Freq: ___ Quality: ___
2	Comprehends spoken target language in meetings and briefings	Freq: ___ Quality: ___
3	Reads and interprets target language documents accurately	Freq: ___ Quality: ___
4	Produces written correspondence in the target language	Freq: ___ Quality: ___
5	Applies cultural awareness in professional interactions	Freq: ___ Quality: ___
6	Serves as language resource for team members	Freq: ___ Quality: ___
7	Handles language-dependent tasks independently	Freq: ___ Quality: ___
8	Participates effectively in bilingual meetings	Freq: ___ Quality: ___
9	Translates or interprets as required by position	Freq: ___ Quality: ___
10	Maintains language proficiency through self-study or practice	Freq: ___ Quality: ___

Table B.1 — Behavioral Observation Checklist Items

Overall Assessment:
Based on your observations, how would you rate this individual's overall on-the-job application of target language skills?
[ ] Exceptional   [ ] Exceeds Standard   [ ] Meets Standard   [ ] Below Standard   [ ] Unacceptable

Comments: Please provide specific examples of how this individual has applied target language skills in the workplace, or note any barriers to application.

Supervisor Name: ________________ **Date:** ________________ **Relationship to Learner:** ________________

Appendix C: Annual Program Evaluation Report Template

Report Section	Content Requirements
Cover Page	Program name, report title, fiscal year, date, classification, prepared by, approved by
Table of Contents	Auto-generated from section headings
1. Executive Summary	Key findings (3-5 bullet points); KPI status summary table; top 3 recommendations
2. Program Overview	Enrollment statistics (headcount, FTE, demographics); course offerings by language; staffing summary (FTE instructors, support staff); budget execution summary
3. Level 1: Learner Satisfaction	Overall satisfaction scores with trend chart; satisfaction by dimension (content, instructor, materials, environment); course-level satisfaction heatmap; comparison to prior year; qualitative themes from open-ended responses
4. Level 2: Learning Outcomes	Proficiency gain rates by modality and language; gain score distributions; effect sizes by language category; certification pass rates; subgroup analysis; comparison to benchmarks
5. Level 3: Behavioral Application	Supervisor BOC results (mean scores by dimension); self-report survey results; work product quality ratings; transfer barriers and facilitators; case study highlights
6. Level 4: Organizational Impact	Mission effectiveness metrics; cost-effectiveness analysis; ROI calculation; readiness fill rates; stakeholder outcome survey results
7. KPI Performance	KPI scorecard with current values, targets, and status; trend charts for each KPI; disaggregated data tables; explanatory notes for KPIs below target
8. Continuous Improvement	PDSA cycles completed; corrective actions status summary; improvement outcomes; lessons learned; planned improvements for next year
9. Recommendations	Prioritized list of recommendations (ranked by impact and feasibility); action plans for each recommendation; resource requirements; timeline
10. Appendices	Detailed data tables; statistical analysis outputs; survey instruments; methodology notes; data quality report

Table C.1 — Annual Program Evaluation Report Template

Appendix D: KPI Tracking Spreadsheet Template

The following template defines the structure of the KPI tracking spreadsheet maintained by the Evaluation Team. This spreadsheet is the authoritative source for KPI data used in quarterly reporting and dashboard updates.

Sheet 1: KPI Summary

Column	Field	Example	Description
A	KPI ID	KPI-1	Unique identifier for each KPI
B	KPI Name	Proficiency Gain Rate	Full name of the KPI
C	Reporting Period	Q1 FY2026	Quarter and fiscal year
D	Current Value	78.5%	Most recent calculated value
E	Target Value	75.0%	Established target for the period
F	Prior Period Value	76.2%	Value from the previous reporting period
G	Status	Green	Green/Yellow/Red based on threshold
H	Trend	Improving	Improving/Stable/Declining
I	Data Source	Proficiency test database	System or document from which data were drawn
J	Date Updated	2026-01-15	Date the value was last updated
K	Notes		Explanatory comments or caveats

Table D.1 — KPI Summary Sheet Structure

Sheet 2: Historical KPI Data

The Historical KPI Data sheet maintains a time-series record of KPI values for trend analysis. Each row represents one reporting period (quarter) for one KPI. Columns include: KPI ID, KPI Name, Fiscal Year, Quarter, Value, Target, Status, and Notes. This sheet supports automated trend chart generation and statistical process control analysis.

Sheet 3: KPI Disaggregation

The KPI Disaggregation sheet provides KPI values broken down by key dimensions: language category (I, II, III, IV), delivery modality (classroom, online, blended), initial proficiency level, and instructor experience level. Columns include: KPI ID, Fiscal Year, Quarter, Disaggregation Dimension, Disaggregation Category, Value, Sample Size, and Notes. This sheet supports equity analysis andtargeted improvement planning.

Sheet 4: Action Tracker

The Action Tracker sheet records all corrective actions and improvement initiatives linked to KPI performance. Columns include: Action ID, Linked KPI, Problem Description, Root Cause, Corrective Action, Responsible Party, Start Date, Target Completion Date, Actual Completion Date, Status (On Track/At Risk/Off Track/Completed), Outcome Metrics, and Notes. This sheet integrates with the Improvement Tracker component of the evaluation dashboard.

End of Document

Government Language Training Program — Program Evaluation Manual ( )

Version 2.0 — January 2026