Development and Pilot Testing of a Tool to Measure the Mid- to Long-Term Impact of Invitational Training Programs for Healthcare Policy Experts from Low- and Middle-Income Countries

Jung, Hyejin

Department of Medical Law and Ethics

Graduate School

Yonsei University

Development and Pilot Testing of a Tool to Measure the
Mid- to Long-Term Impact of Invitational Training
Programs for Healthcare Policy Experts from Low- and
Middle-Income Countries

Advisor Kim Soyoon

A Dissertation Submitted
to the Department of Medical Law and Ethics
and the Committee on Graduate School
of Yonsei University in Partial Fulfillment of the
requirements for the Degree of
Doctor of Philosophy in Medical Law and Ethics

Jung, Hyejin
July 2025

Development and Pilot Testing of a Tool to Measure the Mid- to Long-Term Impact of Invitational Training Programs for Healthcare Policy Experts from Low- and Middle-Income Countries

**This Certifies that the Dissertation
of Jung, Hyejin is Approved**

Committee Chair   **Kim, Soyoon**

Committee Member   **Kim, Taehyun**

Committee Member   **Kim, Heejin**

Committee Member   **Yu, Byungwook**

Committee Member   **Yoon, Bo young**

**Department of Medical Law and Ethics
Graduate School
Yonsei University
July 2025**

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Development and Pilot Testing of a Tool to Measure Mid- to Long-Term Impact of Invitational Training Programs for Healthcare Policy Experts from Low- and Middle-Income Countries

Background:

International training programs for health professionals in low- and middle-income countries (LMICs) often lack robust methods to assess their long-term impacts (Rotem, 2010; Vallejo, 2016; Kim et al., 2020). The Dr LEE Jong-wook Fellowship Program (LJWFP), operated by the Korea Foundation for International Healthcare (KOFIH), aims to build leadership and governance capacity in LMIC healthcare sectors. However, evaluations of the sustained effects of the program are limited.

Objective:

This study aimed to develop and validate a comprehensive evaluation tool capable of measuring the mid- to long-term impact of LJWFP, focusing on career progression, policy influence, academic contributions, professional networking, and contextual factors influencing implementation.

Methods:

The methodology for this study was structured into three sequential phases to develop and validate a tool for assessing the mid- to long-term impacts of international healthcare policy training programs (Kirkpatrick, 1959; Stufflebeam, 1971; Kaufman, 2005; Oh & Yoon, 2024). First, an initial pool of evaluation items was generated through an extensive literature review of established training evaluation models (such as Kirkpatrick, Kaufman,

CIPP, and OECD DAC), a qualitative analysis of interviews, and open-ended responses from program alumni. These items were refined based on expert input. Five specialists in global health and medical education participated in focus group interviews to review and suggest improvements, after which a two-round Delphi process with ten experts was conducted. In the Delphi rounds, experts rated the relevance, clarity, and comprehensiveness of each item using a 5-point Likert scale, with the content validity ratio (CVR) and consensus thresholds guiding item selection. This resulted in a final set of nine questionnaire items, covering career advancement, action plan implementation, knowledge sharing, academic output, organizational contributions, international cooperation, joint projects, and professional networking. Open-ended voluntary questions were added to each question to obtain a more in-depth understanding of the qualitative results.

For data collection, the finalized tool was pilot tested with 52 alumni of the LJWFP, who completed their training in 2023. The survey, administered electronically, included nine closed-ended Likert-scale items and open-ended questions to capture qualitative insights into implementation experiences and challenges. Quantitative data were analyzed using descriptive statistics and exploratory factor analysis (EFA) to confirm the tool's structure (with KMO and Bartlett's tests for sampling adequacy) and reliability testing using Cronbach's alpha. Additional analyses included t-tests and ANOVA to compare outcomes by region, gender, and program type as well as correlation and regression analyses to examine the relationships between the indicators. The qualitative responses were subjected to thematic analysis to identify implementation challenges, contextual influences, and sustainability needs.

Results:
The evaluation tool demonstrated high content validity and moderate reliability across five

factors: career progression, academic achievement, policy contribution, network utilization, and joint project proposals (Lawshe, 1975; Yusoff, 2019; Costello & Osborne, 2005). The participants showed strong outcomes in terms of professional development and knowledge-sharing. There were no notable variations among gender or regional differences, whereas graduate program participants showed significantly higher academic achievement. Early indicators such as satisfaction and application rates were correlated with long-term outcomes. Personal achievements, such as professional and academic achievements, showed a strong correlation with long-term impacts such as policy contributions. This study underscores the value of integrating open-ended qualitative components within the questionnaire to capture implementation challenges, sustainability factors, and the broader context of impact. These insights are expected to be used to enhance follow-up evaluations, customize support, and inform strategic improvements in global health capacity-building initiatives.

Conclusion:

This validated tool offers a robust framework for assessing the long-term impact of international healthcare training programs by combining quantitative measurements with qualitative contextualization. This study underscores the value of integrating open-ended qualitative components to capture implementation challenges, sustainability factors, and the broader context of impact. These insights are critical for enhancing follow-up evaluations, customizing support, and informing strategic improvements in global health capacity building initiatives.

Keywords:

LJWFP, Impact Evaluation, Global Health, Capacity Building, Delphi Method, Healthcare Leadership, LMICs, KOFIH

1. Introduction

1.1. Research background

Global health capacity building has emerged as a fundamental component of international development assistance, particularly for addressing the critical shortage of qualified healthcare professionals in low- and middle-income countries (LMICs). The World Health Organization (WHO) estimates that approximately 57 countries face severe healthcare workforce shortages, with the greatest deficits concentrated in sub-Saharan Africa and parts of Asia. (Frenk et al., 2010; WHO, 2010) This shortage not only affects immediate healthcare service delivery but also compromises the development of sustainable health systems capable of addressing evolving public health challenges.

The right to health represents an essential human right enshrined in multiple international frameworks, including the Universal Declaration of Human Rights, the International Covenant on Economic, Social, and Cultural Rights, and the Convention on the Rights of Persons with Disabilities. Public health, as a multidisciplinary field focusing on disease prevention, life prolongation, and community well-being enhancement, requires sophisticated leadership and governance structures for effective functioning. Therefore, education and training of health policymakers and practitioners responsible for managing these systems has become a strategic priority for international development organizations.

The Dr LEE Jong-wook Fellowship Program (LJWFP), operated by the Korea Foundation for International Healthcare (KOFIH), aims to build leadership and governance

capacity in the LMIC healthcare sectors (Oh & Yoon, 2024; Nakanjako et al., 2021; Kang, 2017). However, evaluations of the sustained effects of the program are limited (Oh & Yoon, 2024; Vallejo, 2016). The program's comprehensive approach encompasses eight specialized tracks: Health Professional Education, Health Policy and Management, Clinical Experts, Biomedical Engineering, Leadership, Health Financing and Economics, Infectious Disease Specialists, and Graduate degree programs. Over its 18-year operational history, KOFIH has trained 1,672 healthcare professionals across 30 countries, demonstrating substantial reach and systematic impact on global health workforce development.

However, despite the program's extensive scope and strategic importance, a systematic evaluation of its long-term impact remains limited. Current evaluation practices in international development training programs often focus on short-term, individual-level outcomes, overlooking sustained systemic impacts and organizational transformations. This limitation significantly constrains the ability of program administrators and policymakers to optimize training investments and maximize contributions to sustainable health systems worldwide.

1.2. Purpose of Study

The primary objective of this study was to develop and validate a comprehensive evaluation tool capable of measuring the mid- to long-term impacts of LJWFP, with particular emphasis on career progression, policy influence, academic contributions, and professional networking outcomes. This research sought to address the critical gap between training

program implementation and systematic impact assessment, which has characterized much of the international development training literature.

Specifically, this study aimed to create a robust measurement framework that could capture the multidimensional nature of training program impacts, while remaining practical for implementation across diverse cultural and institutional contexts. The evaluation tool was designed to assess both individual-level outcomes, such as career advancement and skill application, and broader systemic contributions, including policy development, organizational change, and international collaboration.

The research objectives encompassed several key components: first, establishing a theoretically grounded evaluation framework based on established training evaluation models; second, achieving expert consensus on the most relevant and valid impact indicators through systematic consultation processes; third, developing and validating a practical measurement instrument through pilot testing; and fourth, generating evidence-based insights into the program's effectiveness and areas for improvement.

This study aims to contribute to the broader field of international development evaluation by providing a replicable methodology for assessing training program impacts that extend beyond the traditional short-term metrics. The findings are expected to offer strategic insights for enhancing global health capacity-building initiatives while supporting evidence-based decision making for program administrators, policymakers, and funding organizations.

2. Methods

This study followed a systematic approach to develop, pilot, and validate an evaluation tool to assess the effectiveness of the fellowship program. This methodology employed a structured three-phase framework to create and authenticate an assessment instrument for measuring the efficacy of fellowship programs. These interconnected phases encompassed the following: 1) development of the measurement tool, 2) pilot testing and data collection, and 3) analysis of the evaluation results. Interpretive and statistical methodologies were employed to establish scientific integrity throughout the study.

Figure 1 illustrates the methodology employed to develop and validate the evaluation tool. This systematic approach begins with theoretical foundation building through a literature review and progresses through expert consultation, iterative refinement, empirical testing, and final validation.

The development process follows a sequential framework, in which each stage builds upon the findings and outcomes of the previous phase. This interconnected approach ensures that the final evaluation tool incorporates both scholarly research and professional expertise, while maintaining scientific rigor throughout the development process.
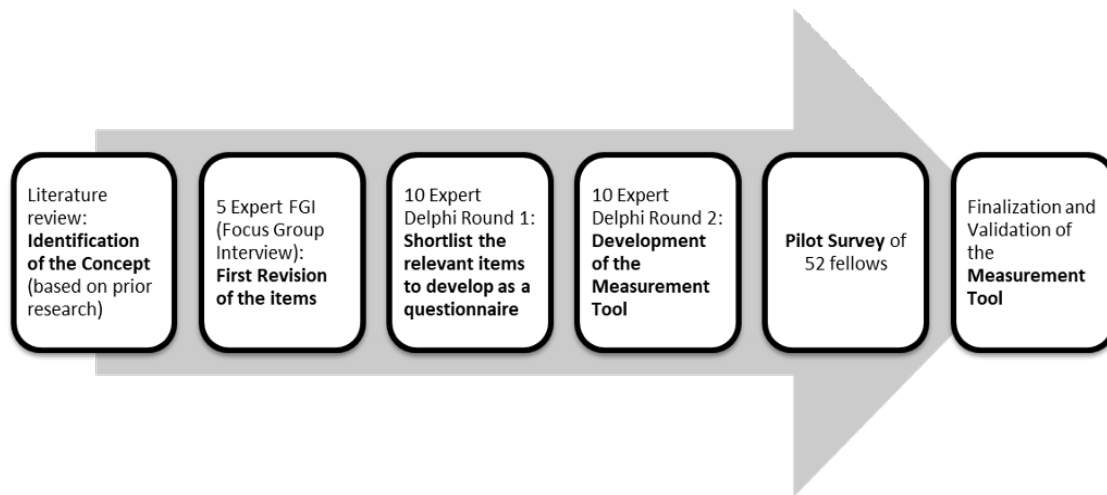
Figure 1. Measurement Tool Development Process

2.1. Development of the Measurement Tool

The assessment instrument was derived from an analysis of existing evaluation frameworks and surveys aligned with the core objectives of the LJWFP. The developmental sequence commenced by synthesizing literature findings with qualitative insights gathered through graduate interviews to establish a conceptual foundation for the measurement approach. Items were systematically identified from the current studies on impact evaluation and the specific purpose of the LJWFP, ensuring theoretical grounding and practical relevance.

To ensure the tool's relevance, qualitative data from previous interviews and open-ended survey responses from successful graduates were analyzed through thematic analysis to extract key competencies, learning experiences, and perceived program impact. Based on these

insights, an initial pool of evaluation items was generated to form a foundation for expert consultation and refinement.

2.1.1. Focus Group Interview

The development of the measurement tool began with a systematic expert consultation through focus group interviews designed to establish content validity and ensure comprehensive coverage of relevant impact dimensions. Five experts in global health and medical education were recruited to participate in structured focus group interviews, representing diverse expertise to ensure a comprehensive review of the initial item sets.

The expert panel consisted of three global health specialists and two medical education experts, all with substantial experience, ranging from 15 to over 20 years. All participants held professional positions in their academic institutions.

Table 1 presents the composition of the expert panel engaged in the Focus Group Interview (FGI), which served as the foundational phase for content validation of the LJWFP impact evaluation tool. The panel consisted of five experts, three specialized in Global Health (GH) and two in Medical Education (ME). All of them had some involvement with LJWFP, either as a Project Manager (PM) or as an advisor of selected fellows in their specialty area.

Table 1. List of FGI Panel Experts

ID	Affiliation	Name	Field of Expertise	Years of Experience	Position	LJWFP Involvement
1	A University	KOO	GH	20 +	Professor	PM
2	B University	YOO	GH	15 ~ 20	Professor	Advisor
3	C University	YOO	ME	15 ~ 20	Professor	Advisor
4	D University	KOO	GH	20 +	Professor	PM
5	C University	KOO	ME	20 +	Professor	PM

Note: GH = experts related to Global Health; ME = experts related to medical education, especially international training.

Two separate focus group sessions were conducted on an online meeting platform to accommodate the varying availability of experts. Each session lasted approximately 30 minutes and was recorded for subsequent transcription and analysis. The semi-structured format allowed facilitators to guide discussions using predefined item categories, while encouraging natural conversation flow and emergent themes.

During the focus group sessions, the experts provided open-ended feedback on predetermined items and evaluated their clarity, relevance, and appropriateness within the context of global health and medical education. They also offered suggestions for additional items that would enhance the comprehensiveness of the set, and recommendations for refining item wording and organization to improve validity and utility. The facilitator employed probing

questions to elicit the rationale behind expert recommendations and encouraged dialogue among participants when divergent perspectives emerged.

Following the focus group interviews, thematic analysis was conducted to identify patterns in expert feedback. This process involved the systematic coding of transcripts, categorization of feedback by item and suggestion type, and synthesis of recommendations for each item. The items were subsequently modified based on expert input, with particular attention paid to enhancing clarity, relevance, and comprehensiveness. The final item set was based on the collective expertise of professionals, strengthening its content validity before subsequent rounds of the Delphi method.

2.1.2. Delphi Rounds

A panel of ten global health and medical education experts engaged in two rounds of the Delphi method to achieve a consensus on the most essential evaluation criteria. Through this process, the measurement tool was refined to include a final set of items. The Delphi survey was distributed to ten expert panel members on February 21, 2025; all members responded with a 100% response rate by February 25, 2025. To ensure convenience in collecting responses and to avoid bandwagon or halo effects, data were collected individually through email.

Table 2 identifies the ten experts who participated in the Delphi validation rounds. The composition includes faculty members, clinicians, and officers from academic institutions and

health-related agencies that span the fields of Global Health and Medical Education. The table details their institutional affiliations, areas of specialization, years of experience, professional titles, and involvement with the LJWFP. This multi-institutional and interdisciplinary composition is critical for ensuring balanced perspectives of item relevance and clarity. The inclusion of participants with both academic- and field-level engagement facilitated consensus development on complex constructs, such as policy impact and international collaboration.

Table 2. List of Delphi Panel Experts

ID	Affiliation	Name	Field of Expertise	Years of Experience	Main Position	LJWFP Involvement
1	A University	COO	GH	5~10	Researcher	None
2	B University	NOO	ME	20 +	Professor	PM
3	C Hospital	KOO	GH	20 +	Clinician	PM
4	D University	KOO	GH	15~20	Professor	Advisor
5	D Center	YOO	GH	15 ~ 20	Clinician	Trainer
6	E University	YOO	GH	15~20	Professor	PM
7	A University	KOO	ME	20 +	Professor	PM
8	K Agency	LOO	GH	10 ~ 15	Officer	Trainer
9	Y University	POO	ME	10 ~ 15	Professor	Advisor
10	A University	KOO	GH	5 ~ 10	Professor	None

Note: GH = experts in Global Health; ME = experts in medical education, especially international training.

To analyze the results, the content validity ratio (CVR) was used to validate the items, and agreement and convergence were analyzed to gather expert opinions. The minimum CVR value depends on the number of panelists, and the content validity is determined when the CVR exceeds this minimum threshold. For this study with ten experts, a threshold of 0.62 was applied, following established statistical criteria was applied. The validity of the Delphi technique was assessed by the degree of agreement among the panelists and analyzed using agreement and convergence metrics. Agreement was scored as 1 when the first (Q1) and third quartiles (Q3) were aligned, indicating a total agreement. Convergence indicates a value of zero when opinions converge to a single point.

A 5-point Likert scale was used to determine the modified Content Validity Ratio (CVR) was determined using a five-point Likert scale. Modified CVR is a widely used quantitative method for establishing content validity in instrument development and validation studies. When applied within the Delphi technique framework using a 5-point Likert scale, it provides a systematic approach to achieving consensus among expert panels while maintaining statistical rigor. Experts rated each item on a 5-point scale: Strongly Disagree, Disagree, Neither Agree nor Disagree, Agree, and Strongly Agree. The modified CVR calculation for 5-point Likert scale responses uses the following formula:

$$CVR = (ne - N/2)/(N/2)$$

where ne represents the number of experts indicating the essential rating (typically, ratings of four or five), and N represents the total number of experts. Ratings of 4 (Agree) and

5 (Strongly Agree) were considered "essential," while ratings of 1, 2, and 3 were considered "not essential".

Following the initial identification of the essential evaluation components, corresponding assessment measurement tools were constructed to systematically address each element. The expert panel subjected each questionnaire item to rigorous validation by examining the dimensions of clarity, relevance, and response accessibility. Upon completion of the second round, finalized assessment instruments were prepared for pilot testing.

2.2. Validation of the Measurement Tool and Pilot Testing

2.2.1. Measurement Tool Validation

A pilot survey was conducted to assess the questionnaire's preliminary validity and reliability. Among the 161 participants who completed the program by 2023, 100 were invited to participate in the survey, excluding 61 who participated in non-policy-related courses (clinical experts, biomedical engineers, and health professional education) that required different measures for impact evaluation. Given the time elapsed since graduation, a response rate of up to 53% was anticipated, excluding one invalid participant because of incorrect personal data, potentially yielding approximately 52 participants in the final survey.

The final survey instrument was administered electronically. Data were systematically captured and exported in comma-separated values (CSV) format to facilitate comprehensive

statistical analysis using appropriate analytical software. Preliminary analyses, including descriptive statistics, item analysis, and reliability testing (e.g., Cronbach's alpha), were performed to evaluate the internal consistency of the tool.

Data collected through the pilot survey were analyzed to determine the effectiveness and reliability of the tool. This structured methodology ensures that the evaluation tool is scientifically robust and consistent within valid, content-valuable, and real-world contexts. Dimensional analysis using inferential statistical techniques was used to uncover the latent construct patterns in the observed indicators. This approach allowed for the identification of fundamental relationship structures underlying surface-level measurements, which is particularly valuable in this research context, where the number of factors and their associated variables were not predetermined.

The suitability of the data was assessed using the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy, where values exceeding 0.7 indicate appropriateness for factor analysis, and Bartlett's Test of Sphericity, which confirmed statistically significant correlations between variables. For factor extraction, Principal Component Analysis (PCA) was selected from the available methods, including Principal Axis Factoring (PAF) and Maximum Likelihood (ML) estimation.

The Kaiser Criterion guided the retention decision, suggesting that factors with eigenvalues exceeding 1.0 be retained as they explain more variance than individual variables. To enhance interpretability, varimax rotation was applied to maximize the variance in loadings within the factors. Factor loadings exceeding 0.4 were considered meaningful for associating

variables with the factors. The analysis also examined potential cross-loadings, where items loaded significantly on multiple factors and communalities represented the portion of each variable's variance explained by the extracted factors.

The structural validation process was based on factor analysis, specifically focusing on establishing construct validity and reliability of the measurement tool. Exploratory factor analysis (EFA) was employed rather than confirmatory approaches, as it is more appropriate for developing an instrument without a predetermined theoretical structure. The EFA process included examination of eigenvalues and scree plots to determine the appropriate number of factors to be retained. For interpretational clarity, this study calculated the percentage of variance explained by the individual factors and their cumulative contributions. Cronbach's alpha coefficients were calculated for each distinct factor that emerged from the analysis to assess the internal consistency and reliability of the identified factors. This statistical approach provided evidence of whether the items within each factor consistently measured the same underlying constructs.

The validation process included computing corrected item-total correlations to determine how strongly each item was related to its respective factor. Values above 0.3 were established as the threshold for acceptable item performance, providing additional evidence of the cohesiveness of each identified dimension, were established as the threshold for acceptable item performance. This comprehensive methodological approach ensured a thorough assessment of both the underlying factor structure, and the psychometric properties of the measurement tool designed to capture the multidimensional impacts of the fellowship program.

The data analysis employed a mixed-methods approach combining quantitative and qualitative techniques to achieve a comprehensive and nuanced understanding of the evaluation results. This approach leverages the strengths of each method, allowing for both statistical rigor and contextual depth, and is widely recognized as the best practice in complex program evaluations, particularly in health and social science research.

A systematic two-step approach was designed to explore the relationships within the dataset and provide robust insights into key research questions. This approach utilized complementary software tools: Microsoft Excel for initial data organization and visualization, and IBM SPSS Statistics for advanced statistical testing. The methodology was structured to ensure a comprehensive analysis by addressing both descriptive and inferential aspects of the data.

The first phase focused on organizing the dataset and conducting a preliminary exploration using Microsoft Excel. This step is critical for ensuring data integrity, identifying potential anomalies, and establishing a foundation for a more in-depth statistical analysis. Tasks performed during this phase included data cleaning for missing values, inconsistencies, and outliers; data structuring by categorizing variables based on their type and relevance; preliminary visualization through basic charts; and calculation of descriptive statistics including mean, median, standard deviation, and frequency distributions.

2.2.2. Analysis of the Pilot Result

The second phase involved conducting inferential statistical tests using IBM SPSS Statistics software (version 27) to validate the hypotheses and quantify the relationships

between variables. The primary statistical methods included t-tests to compare the means of two independent groups to determine statistical significance, Analysis of Variance (ANOVA) to extend group comparisons to multiple categories simultaneously, correlation analysis to quantify the strength and direction of relationships between continuous variables, and regression analysis to quantify the effect of short-term outputs on long-term impacts.

2.2.3. Qualitative Analysis

Open-ended responses from the surveys were analyzed using thematic content analysis to identify recurring themes and patterns affecting the impact of the training. Qualitative analysis followed a systematic approach to ensure comprehensiveness of data interpretation.

All open-ended responses were systematically reviewed and coded using inductive thematic analysis. The analysis process involved multiple stages: initial familiarization with the data, generation of initial codes, identification of themes, review and refinement of themes, and final theme definition and naming.

Response frequency analysis was conducted to determine the participation rates in qualitative feedback across different participant groups. Of the 52 participants, 46 (88%) provided substantive open-ended responses, indicating a high engagement with the qualitative component of the evaluation. The frequencies of specific themes and sub-themes were systematically documented to identify the most commonly reported experiences and challenges.

Regional response patterns were analyzed to identify geographical variations in program implementation and outcomes. Responses were categorized by country of origin and

analyzed for region-specific themes. Regional contextual factors such as healthcare infrastructure, resource availability, and institutional support systems were considered when interpreting these variations.

The analysis also focused on identifying the major thematic categories that emerged from participant responses. A comprehensive thematic framework was developed to organize and present the qualitative findings. Representative quotes were selected for each theme and sub-theme to illustrate the key concepts and provide authentic participant voices. The selection criteria for quotes included clarity of expression, representativeness of the theme, and the ability to provide concrete examples of abstract concepts.

The qualitative findings were catalogued and interpreted to explain and contextualize the quantitative results. Cross-referencing of qualitative themes and quantitative survey responses was conducted to identify convergent and divergent patterns. This triangulation approach enhanced the validity of the findings and provided a more comprehensive understanding of the program's impact and challenges.

3. Literature Review

A comprehensive literature review was conducted using DBpia, PubMed, and Yonsei University Library databases to identify relevant studies published between 2010 and 2024. The selection process focused on literature related to Global Health Training, Training Impacts, Evaluation Methods, and Official Development Assistance (ODA) impacts. After the initial screening, 57 articles were shortlisted for comprehensive analysis.

3.1. Theoretical Background

The evaluation of training and educational programs has evolved significantly over the past several decades, with multiple theoretical frameworks emerging to assess program effectiveness across different contexts and objectives. Understanding these foundational models is essential for developing comprehensive evaluation approaches for contemporary ODA training programs, particularly for those targeting healthcare professionals in LMICs.

3.1.1. Kirkpatrick's Model (1959)

Kirkpatrick's four-level model, introduced in 1959, is a foundational framework for evaluating the effectiveness of training programs through a hierarchy of assessment levels. Level 1 (Reaction) gauges participants' satisfaction and engagement with the training content and delivery methods. Level 2 (Learning) measures the degree to which participants acquire

new knowledge, skills, or attitudes during the training sessions. Level 3 (Behavior) examines how well participants transfer and apply their newly acquired knowledge and skills to real-world settings. Level 4 (Results) evaluates broader organizational impacts such as improvements in productivity, efficiency, and measurable outcomes.

This foundational model has become the most widely recognized evaluation framework in training and development, influencing numerous subsequent evaluation approaches across various sectors. Its enduring popularity stems from its simplicity and practical applicability, making it accessible to practitioners, while providing a systematic approach to understanding training outcomes.

3.1.2. CIPP Model (1971)

Daniel Stufflebeam's Context, Input, Process, and Product (CIPP) model represents a decision-oriented evaluation framework designed to systematically assess programs and guide continuous improvement. Context Evaluation examines program alignment with organizational goals and missions. Input Evaluation assesses the adequacy of the plans, resources, and strategies required for effective implementation. Process Evaluation monitors implementation effectiveness and identifies gaps and challenges during execution. Product Evaluation measures the extent to which programs achieve their intended goals and desired outcomes.

3.1.3. Kaufman's Five Levels (2005)

Building on Kirkpatrick's foundation, Kaufman expanded the evaluation framework to include both organizational and societal impacts. The Kaufman model introduces five distinct levels: Level 1a (Input Assessment) evaluates the quality of training resources and delivery methods, while Level 1b (Process Assessment) examines the acceptability and efficiency of the learning experience. Levels 2 and 3 correspond to Kirkpatrick's learning and application stages, focusing on knowledge acquisition and behavioral change, respectively. Level 4 assesses organizational results, and Level 5 (Societal Outcomes) extends the evaluation to broader societal benefits such as economic growth, public well-being, and industry advancements.

Kaufman's model offers a more comprehensive approach by explicitly addressing the quality and appropriateness of training inputs and processes as well as the societal impacts that extend beyond organizational boundaries. This broader perspective allows evaluators to capture not only the immediate and organizational outcomes of training, but also its contribution to societal progress, making it particularly relevant for programs with wide-reaching objectives, such as those in international development and public health.

3.1.4. WHO Guidelines for Healthcare Professionals' Training (2010)

The WHO guidelines for healthcare professionals' training reflect the evolution of evaluation methodologies to address sector-specific needs, particularly in healthcare, where a distinctive approach to return on investment (ROI) is required. The WHO framework builds on Kirkpatrick's foundational model by incorporating five evaluation levels with a key focus

on public health benefits rather than solely financial returns. This adaptation is significant in global health workforce development, where outcomes are often intangible and influenced by diverse factors, such as training transfer and system performance.

Table 3 summarizes the five levels of the healthcare training evaluation framework suggested by WHO. Unlike traditional models, such as Kaufman, which emphasize financial metrics, the WHO guidelines prioritize broader health outcomes and societal impacts. While the WHO maintains its hierarchical relationship by addressing traditional evaluation levels (reaction, learning, behavior, and results), it recognizes the challenge of quantifying Level 5 (ROI) due to external variables and the complexity of measuring public health impacts. This approach underscores the importance of evaluating both the direct and indirect benefits of training while acknowledging the limitations inherent in assessing long-term societal outcomes.

Table 3. Five Levels of Healthcare Training Evaluation (WHO, 2010)

Level	Focus	Area of Impact	Evaluation Question
1	**Reaction**	Immediate reaction of trainees	Are the trainees satisfied?
2	**Learning**	Increase in trainees' knowledge and skills	What have the trainees learned?
3	**Behavior**	How trainees apply their new knowledge and skills	What do trainees do differently in the workplace
4	**Results**	How the training affects the trainees' broader area of work	What is the effect on the output of the trainee's team or department?
5	**Return on Investment**	Wider impact on the achievement of public health program objectives	How has the training contributed to achieving public health objectives?

3.1.5. New Evaluation Frameworks for LJWFP (2021)

Traditional evaluation methods primarily focus on short-term, individual-level outcomes and often overlook long-term, systemic impacts. However, recent studies emphasize the need for longitudinal evaluations conducted 2–10 years post-program to assess the sustainability of training impacts and to identify unintended consequences. Such evaluations offer critical insights into the long-term effectiveness of training programs, enabling stakeholders to refine their strategies for sustained capacity-building and policy improvements. By shifting towards longitudinal assessment frameworks, programs can ensure a more comprehensive understanding of their real-world impact over time.

New evaluation measurement tools were developed to enhance the evaluation practices of LJWFP. In 2021, new performance indicators were developed to evaluate the effectiveness of the program comprehensively. Since 2022, these indicators have been applied across the entire program cycle from curriculum development to post-program assessment.

Figure 2 shows the new LJWFP Evaluation Framework suggested by this study, which demonstrates the comprehensive integration of three established evaluation models to create a robust assessment system for legal and judicial workforce training programs. The framework strategically combines the four sequential phases of the CIPP Model (Context, Input, Process, and Product) with Kirkpatrick's hierarchical four-level evaluation pyramid (Reaction, Learning, Behaviors, and Result), and systematically incorporates the six OECD DAC criteria (relevance, coherence, effectiveness, efficiency, sustainability, and impact) throughout the evaluation process. This multidimensional approach ensures that evaluation occurs continuously across

all program phases, from the initial design through long-term impact assessment. The framework's structure emphasizes the product phase and result level, reflecting the study's focus on measuring sustained outcomes and systemic changes. By distributing the OECD DAC criteria across different program stages, the framework addresses the critical need for longitudinal evaluation mentioned in the study, providing stakeholders with the tools necessary to assess both immediate training effectiveness and long-term capacity-building impacts over the recommended 2-to-10-year post-program timeframe. However, the tools for measuring long-term impacts remain underdeveloped and require further research.
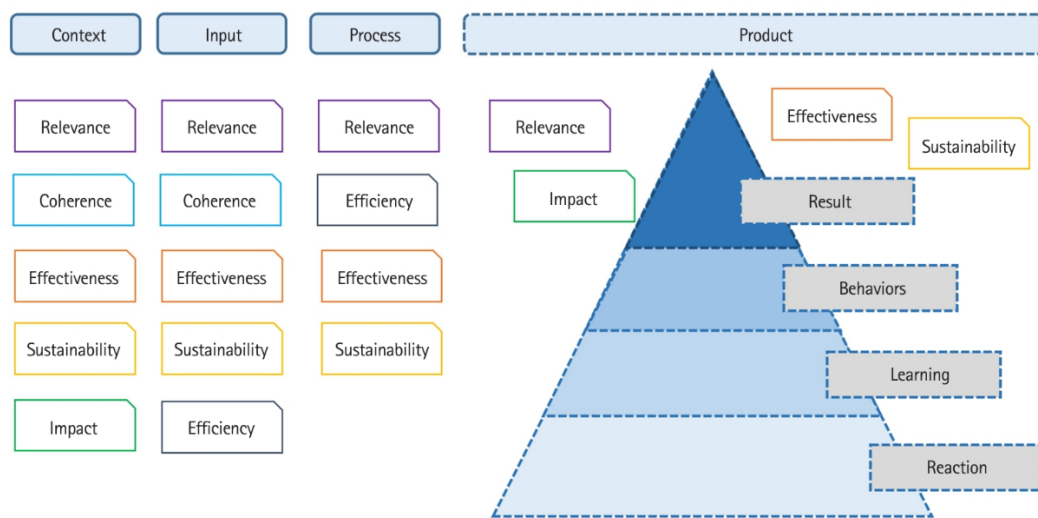
Figure 2. Evaluation Framework of the LJWFP (Oh & Yoon, 2024)

3.2. Current Evaluation Framework of LJWFP

Table 4 identifies three main categories of tools developed for the current evaluation framework of the LJWFP: Demographic Information, Short-term Change, and Mid-term Achievement. Demographic information was collected during the admission process and included basic participant data such as country, gender, and position. Short-term change was evaluated via immediate feedback following the training, focusing on metrics such as achievement, satisfaction, and subject of action plans. Mid-term achievement was assessed within six months of graduation through follow-up surveys, with metrics including workplace application, completion of follow-up checklists, and career promotions.

This highlights the need for further development of tools to assess the long-term impacts. This category involves analyzing outcomes through surveys conducted over two years post-graduation, with metrics focused on policy contribution, academic achievement, and the application of action plans. This distinction underscores a recognized gap in existing evaluation frameworks, where the long-term and sustained impacts of training programs are often under-measured or lack robust assessment tools. This reflects current trends in the literature, which advocate for comprehensive frameworks capable of capturing not only individual growth but also organizational and policy-level changes over extended periods. The development and implementation of such tools are essential for a more complete understanding of the sustained benefits of capacity-building initiatives in healthcare settings.

Table 4. Measurement tools and Specifics

Tool Development	Measurement Target	Assessment Time	Metrics
Measurement Tools developed, tested, and modified	Demographic Information	Assessed during the admission process	Country, Gender, Position
	Short-term Change	Evaluated via immediate feedback	Achievement, Satisfaction, Action Plan Subject
	Mid-term Achievement	Assessed within six months post-graduation using follow-up surveys	Workplace Application, Follow-up Checklist, Career Promotion
Measurement Tools yet to be developed	Long-term Impact	Analyzed through surveys conducted over two years post-graduation	Policy Contribution, Academic Achievement, Action Plan Application

3.3. Previous Studies on Existing Evaluation Models

Previous Studies have pointed to several significant limitations of the current evaluation models. Most evaluations were conducted during or within a few months of the completion of the training program, primarily by measuring participant satisfaction and immediate knowledge acquisition. While these metrics provide insights into trainees' learning experiences, they do not assess whether acquired skills and knowledge translate into tangible changes in workplace performance, institutional policies, or healthcare outcomes over time.

One of the most prominent issues is overreliance on short-term assessments that fail to capture long-term impacts. Many evaluations were conducted during or within a few months

of completion of the training program, primarily to measure participant satisfaction and immediate knowledge acquisition. While these metrics provide insight into trainees' learning experiences, they do not assess whether acquired skills and knowledge translate into tangible changes in workplace performance, institutional policies, or healthcare outcomes over time (Kim, 2020).

Another key limitation is the dominance of quantitative self-assessment tools, which often fail to provide a holistic understanding of the effectiveness of training. Surveys and numerical scales used to measure learning gains do not fully capture the complexities of behavioral and institutional changes. Moreover, trainees may overestimate or underestimate their progress, resulting in inconsistent data. Alternative methods, such as qualitative case studies and in-depth interviews, provide richer insights into how training affects individual careers and organizational practices. However, these approaches are less frequently utilized because of their resource-intensive nature (Vallejo, 2016).

The lack of comprehensive follow-up mechanisms further restricts the assessment of the impact of training programs. Many ODA training institutions, particularly those operating under short-term contracts, lack resources or mandates to track alumni's progress beyond the training period. Consequently, the long-term application of skills, career advancement, and institutional improvement remains largely unexplored. Without sustained engagement with graduates, policymakers and program designers struggle to refine and adapt training curricula to meet real-world needs (Oh & Yoon, 2024).

In addition, existing evaluation models often face difficulties in establishing direct causality between training and broader healthcare improvements. This makes it challenging to isolate the specific contributions of training programs. While frameworks such as Contribution Analysis attempt to address this issue by building plausible "performance stories" linking training to institutional benefits, such approaches require extensive stakeholder engagement and data collection efforts that are not always feasible in resource-limited settings (Rotem 2010).

Cultural and organizational barriers further complicate the evaluation process. In some recipient countries, hierarchical workplace structures and resistance to change limit the extent to which newly trained professionals apply their skills. Even when training programs successfully equip individuals with knowledge and leadership competencies, a lack of institutional support often prevents them from effectively applying these skills. This disconnect between training objectives and workplace realities is rarely captured in traditional evaluation metrics, leading to incomplete assessments of training outcomes (Mangwanya, 2022).

A major limitation of existing evaluation tools is their inability to measure long-term changes in workplace behavior and policy implementation. Many evaluations were conducted within six months of training completion, which was insufficient to observe substantial institutional transformations. Moreover, cultural, organizational, and political barriers often prevent trained professionals from fully utilizing their acquired skills, further complicating impact assessments (Rotem, 2010).

3.4. Improved Evaluation Model

To address the limitations found in previous studies, an improved model was suggested, as shown in Table 5. The desired evaluation model incorporates multiple follow-up observations at extended intervals such as 2, 5, and 10 years after completion. This allows the tracking of long-term outcomes and the ongoing relevance of training in changing institutional or system contexts. Finally, the traditional approach centers on the individual trainee as the sole evaluation target. The desired model broadens this by including supervisors, colleagues, affiliated organizations, and even national-level entities. This multilevel focus enables evaluators to assess how training outcomes are disseminated across networks and contribute to collective change.

Table 5. Existing Evaluation Models and Desired State

Category	Existing Evaluations Models	Desired State
Measurement	Output (e.g. number of trainees, number of countries, etc.)	Influence level (community contribution, academic achievement, service satisfaction)
Key Performance Indicators	Subjective individual competency (e.g. satisfaction, competency achievement, field application, etc.)	Objective organizational competency-based (e.g. manual establishment, policy improvement, national health key indicators, etc.)
Measurement Cycle	Once within 1 year after completion	Long-term follow-up observation (e.g. 2, 5, 10-year cycles)
Evaluation Participants	Individual trainee	Individual Trainee and others (e.g. superiors, colleagues, affiliated organization, country, etc.)

Figure 3 shows the integration of the multiple evaluation models applied at different stages of the LJWFP derived in this study. The timeline was segmented into five phases: during training, upon completion, six months after completion, mid-term impact (within two years), and long-term impact (beyond five years). This staged approach enables evaluators to assess the intervention's effectiveness not only in the short term, but also in its extended outcomes and impacts.

In the early phases, during and immediately after training, the evaluation focused on participants' immediate reactions, learning outcomes, and early behavior change. These are captured through models, such as Kirkpatrick's four-level model, which includes reaction, learning, behavior, and results. Kaufman's model is also applied during these stages, offering additional emphasis on input and process evaluation (1a), participant satisfaction (1b), and broader societal outcomes (L5). These models are well-suited for assessing the initial implementation fidelity and direct outcomes of the training.

As the program moves into mid- and long-term phases, the evaluation framework incorporates broader and more systemic models. The CIPP model, comprising the context, input, process, and product dimensions, supports the evaluation of design quality, implementation processes, and immediate outputs. This model bridges the gap between process-focused and result-focused evaluations by linking activities to their outputs and short-term effects. The OECD DAC evaluation criteria were employed at mid- and long-term points. These factors include relevance, effectiveness, efficiency, impact, sustainability, and consistency. DAC criteria are widely accepted in international development and provide a

structured basis for assessing the strategic value, operational performance, and enduring impact of interventions.
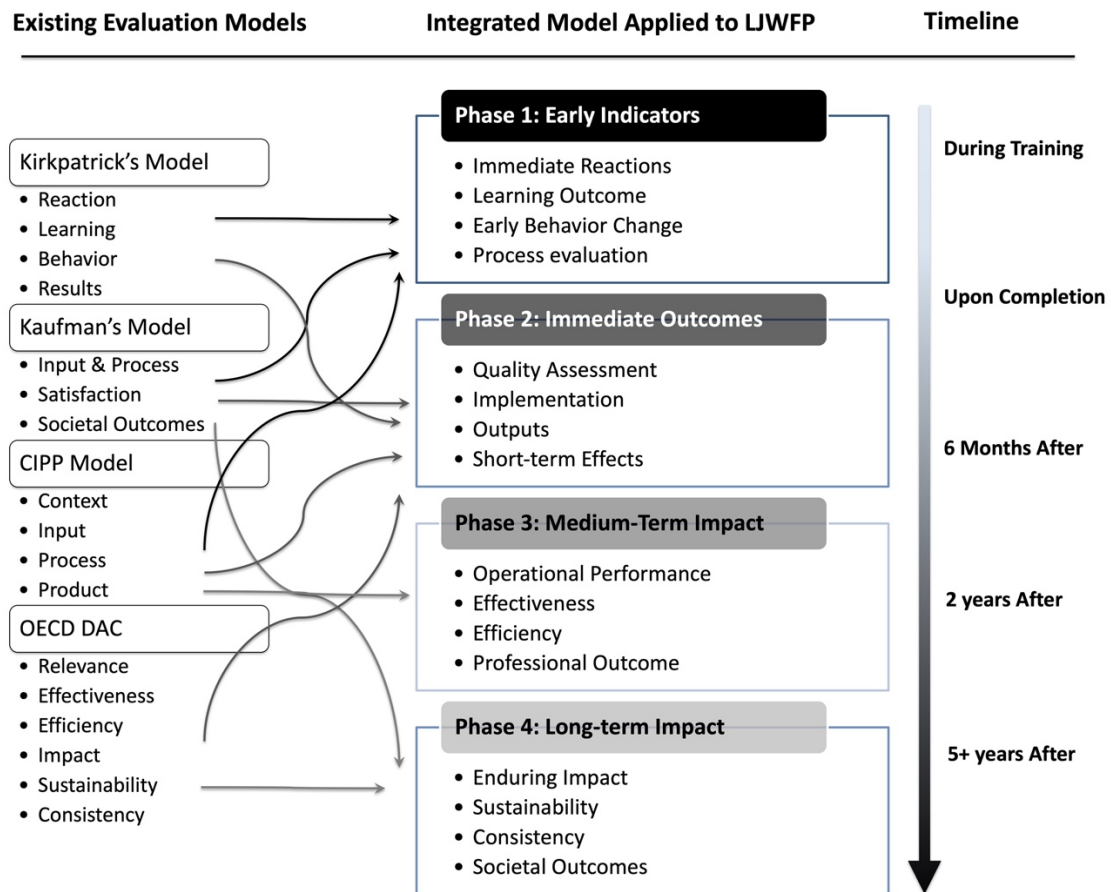
Figure 3. Integrated Application of Evaluation Models derived from this Study

4. Results

4.1. Development of the Impact Measurement Tool

Evaluating the long-term impacts of international fellowship programs requires a systematic and context-sensitive approach. The LJWFP aims to contribute to the development of health systems and leadership capacity in partner countries; however, assessing its extended influence remains complex because of the diverse professional trajectories and institutional environments of its participants. To address this challenge, a structured evaluation framework was designed to capture multiple dimensions of change at both the individual and organizational levels.

In constructing this framework, the development process was guided by both theoretical underpinnings and empirical inputs. Theories of capacity building, knowledge transfer, and professional development provide conceptual grounding, while qualitative and quantitative methodologies ensure practical validity. These steps ensured that the final tool was both analytically robust and contextually appropriate for evaluating the long-term outcomes of LJWFP.

4.1.1. Experts' Review and Refinement of the Evaluation Items

The FGI was conducted by five international experts in global health and medical education. The objective was to generate and refine items for evaluating the LJWFP based on three primary objectives: (1) Strengthening National Health Systems, (2) Promoting

Sustainable Development in Global Health, and (3) Expanding Korea's Role in International Health Diplomacy.

Table 6 outlines the initial questionnaire's structure. For each program objective, relevant evaluation categories, guiding questions, and corresponding sub-questions were defined. For example, under the objective of Strengthening National Health Systems, the evaluation focused on Career Achievement and Policy Contribution (national). Guiding questions included whether the participants had been promoted post-program and the current status of their action plans. Sub-questions explored job positions before and after participation, and levels of action plan implementation, from local sharing to full national dissemination.

Similarly, the objective of Promoting Sustainable Development in Global Health includes Academic and Policy Contributions (Global). Questions in this domain assessed academic outputs, such as presentations and publications, as well as contributions to policy at institutional and national levels. Notably, for the third objective–Expanding Korea's Role in International Health Diplomacy–the evaluation category was Collaboration Involvement, although specific questions and sub-questions were not developed at this stage.

Table 6. FGI Questionnaire Structure based on LJWFP Purposes

LJWFP Purpose	Evaluation Category	Questions	Sub-questions
Strengthening National Health Systems	Career Achievement	Have you been promoted to a higher position after completing LJWFP?	· How would you best describe your position at the time you participated in the LJWFP? · How would you best describe your current position?
	Policy Contribution (National)	What is the status of your action plan?	· The plan has been shared within the organization (locally), but no action has been taken. · The plan has been shared within the organization and some actions have been taken but not completed. · The action is completed and has achieved the outcomes and then shared with other organizations. · The action is implemented and has achieved the outcomes and has been spread out to other organizations and/or nationally.
Promoting Sustainable Development in Global Health	Academic Contribution	After completing LJWFP, what were your academic achievements?	· Presentations in Local/institutional Level Conferences, Seminars, or Webinars · Presentations in International Level Conferences, Seminars, or Webinars · Articles Published in Local or Non-SCIE Journals · Articles Published in Major (SCIE) JournalsEnrolled in advanced degree
	Policy Contribution (Global)	What were your contributions to health policy after completing LJWFP?	· Contributed to creating policy documents (manuals, guidelines, or programs) at the institutional level. · Involved in the implementation of policies at

	the institutional level
	· Contributed to creating policy documents (manuals, guidelines, or programs) at the national level
	· Involved in the implementation of policies at the national level
Expanding Korea's Role in International Health Diplomacy – N/A	

The panel members were presented with draft survey questions developed based on the purpose of the LJWFP and the long-term evaluation items.

Table 7 presents the categorized comments from the FGI members after reviewing the items. In the Terminology category, the comments recommended conducting a pilot survey to verify that terminology is universally understood across all participants' fields. They also suggested clarifying expressions related to policy implementation, as there may be confusion between policymaking, which is considered a higher-level activity, and implementation, which is considered a lower level.

Additionally, the comments proposed reframing questions to ask more clearly whether participants created policies or implemented them. In the Survey Structure category, feedback advises maintaining a 5-point scale for quantitative analysis purposes. Within the indicator category, the comments recommended developing separate indicators for policy-related and non-policy courses. They advise applying variable weighting to indicators, based on the characteristics of each course. For non-policy courses, feedback suggests developing

alternative indicators, as graduates of these courses may find it more difficult to contribute directly to policy changes. For policy-related courses, the comments propose broadening the "Academic Contributions" category to include policy reports, policy documents, conference presentations, and journal publications. Furthermore, it is recommended to include more items related to professional network expansion. In the Question Phrasing category, feedback suggests using the present perfect tense rather than the past tense in questions.

Table 7. Summary of FGI Feedback

Category	Comments
Terminology	· Conduct a pilot survey to verify that terminology is universally understood across all participants' fields · Clarify expressions related to policy implementation, as there may be confusion between policymaking (higher level) and implementation (lower level) · Reframe questions to more clearly ask whether participants created policies were implemented
Survey Structure	Maintain the 5-point scale for quantitative analysis purposes

Indicators	Develop separate indicators for policy-related courses and non-policy courses
	Apply variable weighting to indicators based on course characteristics
	· For non-policy courses: Develop alternative indicators, as graduates may find it more difficult to directly contribute to policy changes
	· For policy-related courses: Broaden the "Academic Contributions" category to include:
	   ✓ Policy reports
	   ✓ Policy documents
	   ✓ Conference presentations
	   ✓ Journal publications
	   ✓ Include more items regarding professional network expansion

Question Phrasing	Use present perfect tense rather than past tense in questions

As a result of the FGI, a final list of 30 evaluation items was established that integrated both qualitative and quantitative expert feedback, as shown in Table 7. Initially, a set of draft items was developed based on the core purposes of the LJWFP, as identified through a literature review and preliminary focus group interviews (FGI). These draft items reflected key domains, such as career achievement, policy contribution, academic contribution, organizational achievement, and professional network expansion.

The resulting 30 items were organized into five categories that reflected the purpose of LJWFP: career achievement, organizational achievements, academic contributions, policy contributions, and expansion of professional networks. Each item was designed to capture a specific aspect of program impact, such as increased work confidence, promotion to higher positions, sharing and implementation of action plans, academic output, policy involvement, and participation in professional networks.

Table 8 presents the final pool of 30 evaluation items categorized into five domains: career achievement, organizational achievements, academic contributions, policy contributions, and expansion of professional networks. Each item operationalizes specific constructs, such as increased professional confidence, action plan implementation, academic productivity (e.g., presentations and publications), engagement in policymaking, and participation in international collaborations. This multidimensional structure allows for the assessment of both individual- and institutional-level impacts and accommodates various forms of post-program engagement.

Table 8. Evaluation Items Identified

Category	Indicator Items
Career Achievements	Increased work confidence
	Assignment to desired position
	Recognition from colleagues and superiors
	Promotion to a higher position within the same organization
	Transfer to higher-level organizations such as central government ministries
	Transfer to overseas organizations such as international organizations
Organizational Achievements	Sharing training results within the organization
	Implementation of an action plan within the organization
	Achievement of action plan results
	Spreading action plan results to other organizations/regions
	Increasing awareness of the affiliated organization internally and externally
	Participation in government-supported projects of affiliated organization
Academic Contributions	Conducted dissemination training (ToT)
	Writing and Publishing papers/articles
	Advancing to higher-level schools
	Established ng new education curriculum
	Presenting in related forums
	Establishing new departments
Policy Contributions	Proposing manuals, guidelines, or programs within the organization
	Participation in developing or improving policies within the organization
	Writing national policy manuals, guidelines, or programs
	Proposing (inter)national policy development or improvement
	Participation in government policy projects
	Participation in International cooperation projects
Expansion of Professional Networks	Participation in cooperative projects with Korea
	Promotion of joint projects (research) with networks formed through training
	Participation in regular alumni meetings
	Participation in alumni social contribution activities
	Formation of expert groups
	Participation in related academic societies

4.1.2. Delphi Rounds and Derivation of the Final Evaluation Items

4.1.2.1. First Round of the Delphi

The first round of the Delphi study was based on preliminary thirty items in five categories identified through focus-group interviews. The evaluation process incorporated a rigorous quantitative assessment protocol utilizing three key methodological metrics: the Content Validity Rate (CVR), Consensus, and Convergence. These metrics were selected for their established reliability in determining both the relevance of the evaluation criteria and the level of expert agreement. To ensure methodological rigor, the research team established predetermined threshold values for each metric, as follows: Content Validity Rate (CVR) $\geq$ 0.62, consensus $\geq$ 0.75, and convergence $\leq$ 0.50. These threshold values were determined based on established methodological literature and implemented to maintain high standards of analytical validity throughout the evaluation process.

The results from the first round of evaluation indicated an average CVR of 0.493 across all items, with an average consensus of 0.772 and convergence of 0.471. Notably, the strongest validity was observed for items related to increased work confidence, dissemination of fellowship outcomes within organizations, participation in organizational policy development, cooperation with Korea, and joint projects initiated with networks formed during fellowships, resulting in perfect CVR scores.

Ten items demonstrated sufficient validity based on the established metrics ($\geq$ 0.62) and were selected for inclusion in the subsequent round. Additionally, five items that achieved

borderline validity with CVR scores of 0.6 were identified for re-evaluation in the subsequent round.

Table 9 reports the results of the first Delphi round using three metrics: Content Validity Ratio (CVR), Consensus, and Convergence. Items with CVR $\geq$ 0.62, consensus $\geq$ 0.75, and convergence $\leq$ 0.50 were considered valid. The table reveals variability in expert agreement, with only ten items meeting the minimum CVR threshold. Items such as "increased work confidence,' "sharing training results," and "participation in international cooperation" demonstrated high CVR and consensus scores. In contrast, items such as "Assignment to desired position" and "Participation in government-supported projects" underperformed, suggesting limited relevance or variability in interpretation.

Table 9. Delphi Round 1 - Content Validity Rate of the Items

	Items	CVR ($\geq$0.62)	Consensus ($\geq$0.75)	Convergence ($\leq$0.50)
1	Increased work confidence	1	0.85	0.38
2	Assignment to desired position	-0.4	0.75	0.38
3	Recognition from colleagues and superiors	0.4	0.81	0.38
4	Promotion to a higher position within the same organization	0.6	1.00	0.00
5	Transfer to higher-level organizations such as central government ministries	0.2	0.75	0.50

6	Transfer to overseas organizations such as international organizations	0.2	0.75	0.50
7	Sharing training results within the organization	1	0.85	0.38
8	Implementation of an action plan within the organization	0.8	0.85	0.38
9	Achievement of action plan results	0.8	0.75	0.50
10	Spreading action plan results to other organizations/regions	0.2	0.50	1.00
11	Increasing awareness of the affiliated organization internally and externally	0.2	0.75	0.50
12	Participation in government-supported projects of affiliated organization	0	0.71	0.50
13	Conducted dissemination training (ToT)	0.8	0.85	0.38
14	Writing and Publishing papers/articles	0.8	0.80	0.50
15	Advancing to higher-level schools	0.2	0.75	0.50
16	Established ng new education curriculum	0.2	0.75	0.50
17	Presenting in related forums	0.6	0.78	0.50
18	Establishing new departments	-0.2	0.67	0.50
19	Proposing manuals, guidelines, or programs within the organization	0.6	0.80	0.50
20	Participation in developing or improving policies within the organization	1	0.78	0.50
21	Writing national policy manuals, guidelines, or programs	0.4	0.56	0.88
22	Proposing (inter)national policy development or improvement	0.8	0.81	0.38
23	Participation in government policy projects	0.4	0.61	0.88
24	Participation in International cooperation projects	0.6	1.00	0.00

25	Participation in cooperative projects with Korea	1	0.78	0.50
26	Promotion of joint projects (research) with networks formed through training	1	0.78	0.50
27	Participation in regular alumni meetings	0.4	0.81	0.38
28	Participation in alumni social contribution activities	0.2	0.75	0.50
29	Formation of expert groups	0.4	0.56	0.88
30	Participation in related academic societies	0.6	1.00	0.00
Average		**0.493**	**0.772**	**0.471**

4.1.2.2. Second Round of the Delphi

Following the first round, the panel's feedback and comments informed the refinement of the evaluation items, which were subsequently presented in a clear questionnaire format during the second round. The same panel of ten experts re-evaluated the questions derived from the 15 items validated in the previous round. Evaluation criteria were expanded to include three additional dimensions: clarity, relevance, and ease of response. The same Delphi panel of ten experts assessed 15 items across the three dimensions using a five-point Likert scale.

The second Delphi round refined and re-evaluated 15 items that were assessed for clarity, relevance, and ease of response (Table 10). Improvements were observed across all validation metrics: the average CVR increased to 0.73, consensus was 0.82, and convergence improved to 0.42. Nine items exceeded the CVR threshold of 0.80 and were retained for the

final questionnaire. These included items assessing professional confidence, promotion, action plan implementation, academic publishing, and policy contribution.

However, the item on national policy development remained a challenge as it had the lowest validity (CVR = 0.40). Experts have noted that national policy development often exceeds the expected scope of impact for fellows, making universal assessment difficult. Such achievements, when present, may be better captured through qualitative success stories than standardized evaluations. Consequently, nine items that met or exceeded the 0.8 CVR threshold were selected for the final questionnaire, emphasizing the importance of a thorough validation process for constructing reliable evaluation tools.

Table 10. Delphi Round 2, Content Validity Rate of the Questionnaires

	Questions	CVR (≥0.62)	Consensus (≥0.75)	Convergence (≤0.5)
1	I have increased confidence in my professional capabilities.	0.60	0.75	0.50
2	I have been promoted to a higher position within my organization.	0.80	0.78	0.50
3	I have shared the fellowship outcomes within my organization.	0.60	0.79	0.50
4	I have implemented the action plan within my organization.	0.80	0.79	0.50
5	I have achieved tangible results from implementing the action plan.	0.80	0.78	0.46
6	I have conducted Training of Trainers (ToT) sessions.	0.80	0.85	0.38
7	I have written academic papers related to my fellowship work.	0.80	1.00	0.00
8	I have presented at academic forums.	0.80	0.83	0.42
9	I have proposed manuals, guidelines, or programs within my organization.	0.80	0.90	0.25
10	I have participated in organizational policy development or improvement.	0.60	0.80	0.50
11	I have proposed national policy development or improvements.	0.40	0.73	0.54
12	I have participated in international cooperation projects.	0.60	0.80	0.46
13	I have participated in cooperation projects with Korean or other international organizations	0.80	0.83	0.42
14	I have initiated joint projects or research with networks formed during the fellowship.	0.80	0.83	0.42
15	I have participated in relevant academic societies or conferences.	0.80	0.77	0.50
Average		**0.73**	**0.82**	**0.42**

4.1.2.3. Final Questionnaire to Measure Long-term Impact of the LJWFP

The final questionnaire employed a five-point Likert scale ranging from "Strongly disagree" to "Strongly agree" to assess various impact dimensions: career advancement, action plan implementation, knowledge sharing through Training of Trainers, academic publications, presentations at academic forums, development of organizational manuals and guidelines, engagement in international cooperation, initiation of joint projects with fellowship networks, and participation in relevant academic societies.

The measurement tool demonstrated high content validity through a systematic expert review. The two-round Delphi method with ten experts resulted in a strong consensus on item relevance, clarity, and comprehensiveness. The final nine items achieved Content Validity Ratios exceeding 0.8, well above the required threshold of 0.62 for ten experts, indicating robust content validity (Lawshe, 1975).

The final questionnaire, detailed in Table 11, consisted of nine items validated through the Delphi process. The questionnaire uses a five-point Likert scale. The items assess diverse areas, including career advancement, implementation and dissemination of action plans, Training of Trainers (ToT) activities, academic outputs, policy engagement, and participation in academic networks. This final form demonstrated high content validity, practical relevance, and comprehensive coverage of the program's intended outcomes. The use of plain language and behaviorally anchored response options enhances accessibility across diverse respondents' backgrounds.

Table 11. Final Questionnaire

Please rate the item below that best describes your situation after completing the LJWFP.

Scale: 1 = Strongly disagree 2 = Disagree 3 = Neither agree nor disagree 4 = Agree 5 = Strongly agree

	Question	Likert Scale
1	My career has progressed since LJWFP	□ 1 □ 2 □ 3 □ 4 □ 5
2	I have successfully implemented my action plan.	□ 1 □ 2 □ 3 □ 4 □ 5
3	I have been sharing my knowledge by conducting Training of Trainers (ToT) sessions.	□ 1 □ 2 □ 3 □ 4 □ 5
4	I have written academic papers and reports related to what I learned through LJWFP.	□ 1 □ 2 □ 3 □ 4 □ 5
5	I have presented my research in academic forums.	□ 1 □ 2 □ 3 □ 4 □ 5
6	I have proposed manuals, guidelines or programs within the organization.	□ 1 □ 2 □ 3 □ 4 □ 5
7	I have participated in collaborative projects with Korean and other international organizations	□ 1 □ 2 □ 3 □ 4 □ 5
8	I have started joint projects and/or research with the network formed through the LJWFP.	□ 1 □ 2 □ 3 □ 4 □ 5
9	I have participated in academic societies and conferences (other than KGA).	□ 1 □ 2 □ 3 □ 4 □ 5

4.1.3. Factor Derivation of the Evaluation Items

To further verify the validity and reliability of the measurement tool, the validity of the questionnaire was analyzed using exploratory factor analysis (EFA) to determine whether the nine items effectively measured distinct aspects of the impact of the fellowship program. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.705, which is above the recommended threshold of 0.6, indicating that the dataset was suitable for factor analysis. Additionally, Bartlett's test of sphericity was statistically significant (χ^2 = 135.47, df = 36, p < .001), confirming that the correlation matrix was not an identity matrix and that the items were sufficiently related to justify the factor analysis.

The results revealed four distinct factors, each representing a core component of the fellowship's impact: Policy implementation, academic contribution, collaborative networking, and project development. The cumulative variance explained by these four factors was 57.196%, indicating that the questionnaire effectively captured a broad range of fellowship-program-related outcomes. Hair et al. (2006) and Henson and Roberts (2006) established that a cumulative variance explained (CVE) of 50–60% is generally considered sufficient in exploratory factor analysis within social sciences and medical education, particularly when the constructs being measured are complex and multidimensional, as in this study. This standard is widely recognized because social and educational research often deals with latent variables that are inherently less precise than those in the natural sciences, where much higher thresholds are expected. Therefore, a CVE below 60% can be considered acceptable in exploratory studies evaluating healthcare education provided that the factor structure is theoretically coherent and interpretable.

Table 12 shows the Rotated Factor Matrix results, revealing robust factor loadings that supported the construct validity of the tools. Notably, action plan implementation exhibited a strong loading of 0.721 on the policy factor, while academic papers demonstrated an exceptionally high loading of 0.957 on the Academic Contribution factor, confirming the distinct construct representation.

However, certain items displayed relatively lower factor loadings or cross-loadings, particularly when differentiating between overlapping constructs, suggesting potential areas for instrument refinement in future applications.

The factor loadings from the analysis demonstrated that each item loaded well onto its respective factor, thus, supporting the instrument's construct validity. For example, items related to policy implementation (action plan implementation, proposed manuals/guidelines/programs, and ToT) exhibited strong loadings on the first factor, whereas items concerning academic contribution (academic papers and reports, and academic forums) formed a separate factor. Collaborative project participation and academic societies/conferences loaded highly on the third factor, whereas joint project proposals and career progress loaded on the fourth factor. This clear delineation of the factors suggests that the questionnaire can effectively measure distinct constructs. However, some items displayed relatively low factor loadings or cross-loadings, indicating potential areas for refinement, particularly in differentiating between overlapping constructs in the factor structure. For example, Q5. Academic forums loaded 0.391 on the policy factor and 0.458 on the academic factor, revealing a dual characteristic, whereby academic activities also translate into policy change.

Table 12. Factor Derivation of the Measurement Tool (Rotated Factor Matrix)

Questions	Factor			
	Policy	Academic	Network	Professional
Q2. Action plan Implementation	0.721	0.154	0.144	0.192
Q6. Proposed manuals, guidelines or programs	0.623	0.385	-0.135	0.085
Q3. Training of Trainers (ToT)	0.576	-0.027	0.129	0.219
Q4. Academic papers and reports	0.09	0.957	0.136	0.165
Q5. Academic forums	0.39	0.458	0.121	0.289
Q7. Collaborative projects Participation	0.22	0.046	0.752	-0.05
Q9. Academic societies and conferences (other than KGA)	-0.046	0.091	0.618	0.217
Q8. Joint projects Proposal	0.318	0.266	0.062	0.639
Q1. Career Progress	0.437	0.125	0.294	0.471
Eigenvalue	1.743	1.395	1.127	0.882
VE (%)	19.365	15.503	12.527	9.801
CVE (%)	19.365	34.868	47.395	57.196

KMO= .705, Barlett's $\chi 2$ = 135.474 (p<.001)

Table 13 presents the reliability analysis results for the four distinct factors identified through the exploratory factor analysis of the nine-item LJWFP impact measurement tool. This analysis employed Cronbach's alpha coefficients to assess internal consistency reliability, which measures how well items within each factor correlate with one another and consistently measures the same underlying construct. Four distinctive factors were identified using the nine survey items. The policy Implementation factor comprised items related to action plan implementation, proposed manuals/guidelines, and training of trainer programs, with a Cronbach's alpha of 0.707 indicating acceptable reliability. Academic Contribution factors included academic papers/reports and academic forums, demonstrating strong internal consistency, with an alpha of 0.731. The collaborative Networking factor encompassed collaborative project participation and academic societies/conferences with an alpha of 0.626, reflecting marginal but acceptable reliability. The career Development factor included career progress and joint project proposals, although with a lower alpha of 0.531, suggesting the need for further refinement.

Overall, the measurement tool demonstrated acceptable to good reliability for most factors except for Project Development, which requires further refinement. These results support the instrument's suitability for assessing the multifaceted outcomes of fellowship programs. However, the relatively small number of items per factor suggests that, while the structure is meaningful, future research should consider expanding the number of items to further stabilize each factor. Especially Q1. Career Advancement, and Q8. Project Initiation showed the lowest reliability, indicating that it is better to not group them under one factor. Career advancement is a personal achievement, and joint project proposals are more for public contribution in the long term, which might not be realized within a two-year timeframe.

Table 13. Reliability of the Survey Items

Factor	Representative Items	Key Characteristics	Cronbach α	Corrected Item-Total Correlation
Policy Implementation	Action plan implementation, ToT, Proposed manuals/guidelines	Organizational policy and capacity	0.707	0.469~0.610
Academic Contribution	Academic papers/reports, Academic forums	Research output, scholarly engagement	0.731	0.577
Network Utilization	Collaborative project participation, Academic Societies	Professional networking, collaboration	0.626	0.456
Career Development	Career progress, Joint projects proposal	Career advancement, project initiation	0.531	0.512

4.2. Results of the Pilot Tests

The pilot survey was administered to fellows who completed one of four relevant courses (Infectious Disease, Graduate Degrees, Health Policy, and Health Financing and Economics) in 2023, two years prior to the current study. Participation in this study was voluntary. The remaining four courses (Health Professional Education, Clinical Expertise, Biomedical Engineering, and Leadership) were excluded because they had distinct impact objectives that diverged from policy improvement and knowledge sharing, necessitating the development of separate assessment instruments better aligned with their specific purposes. Due to the nature of the courses, Health Policy courses and Health Financing and Economics courses were grouped together as Health Policy and Management.

This study was reviewed and approved by the Institutional Review Board (IRB) of Yonsei University Health System, Severance Hospital (4-2024-1636), ensuring adherence to ethical research standards. All participants provided informed consent before participating in the study. The research analyzed demographic variables, including gender, age, program attended, and geographic location. These variables were selected and managed carefully, considering their potential for individual identification. Demographic data were utilized exclusively for matching purposes with short-term indicator outcomes during the initial data-processing phase. Upon completion of the matching process, the dataset was structured to render reidentification practically impossible, thereby ensuring participant anonymity throughout the analysis phase.

4.2.1. General Characteristics of the Participants

The pilot survey involved 52 fellows from diverse geographical and professional backgrounds.

KOFIH has trained 1,672 healthcare professionals across 30 countries throughout LJWFP's 18-year operational history (2007–2024). The survey was intentionally limited to participants from three specific program tracks aligned with policy improvement and knowledge-sharing objectives. Four other program tracks were excluded from the survey because they had distinct impact objectives that required separate assessment instruments better aligned with their specific purposes.

Table 14 demonstrates the distribution of the participating program categories, indicating the program's successful recruitment of professionals across key healthcare specialization areas

The training Course for Infectious Disease Specialists is an educational program that encompasses epidemiology, clinical diagnosis and treatment, and disease diagnosis and research. It aims to contribute to the qualitative improvement of healthcare services, establish a foundation for self-sustained growth, and strengthen the capacity of health care professionals while building health care systems. It aims to improve the healthcare standards of partner countries and to foster partnerships to promote international cooperation. The study duration was 9 weeks. Every year, LJWFP welcomes around 35 participants to this course, comprising the 2nd greatest number of participants among all eight course categories, following the clinical

experts' course, which welcomes around 50 participants. There were 21 survey participants, comprising 40.4% of the total. (Korea Foundation for International Healthcare. (n.d.))

The graduate degree course encompasses basic medical science, medical education, healthcare policy, and biomedical engineering, and offers master's degree programs to healthcare professionals in the fields of basic medicine, medical education, and nursing. This program aims to support the qualitative improvement of healthcare services and the establishment of a self-sustaining growth foundation in partner countries, thereby enhancing and strengthening healthcare systems to promote sustainable development. The study duration was 2 years. Every year, LJWFP welcomes around 30 participants to this course, comprising the 3rd greatest number of participants among all eight course categories, next to the Infectious Disease Specialist course. The survey participants included 18 people, comprising 34.6% of the total.

The training Course for Health Policy and Management offers a variety of educational courses related to health policy and planning. It aims to enhance the skills for proposing, formulating, implementing, monitoring, and evaluating policies. It seeks to improve research and practical abilities in the field of health policy, ultimately enabling the development of healthcare policies tailored to the specific context of partner countries. The study duration was Three months. Every year, LJWFP welcomes approximately 14 participants in this course.

Health Financing and Economics courses, the latest sprout out of the Health Policy and Management course, provide theoretical learning and field experience related to health finance and economics. It aims to offer customized learning opportunities for trainees, including

mentoring tailored to the characteristics of partner countries' health systems, seminars, discussions, field studies, and action-plan development. It seeks to strengthen theoretical, administrative, and practical capabilities through country-specific case applications, ultimately aiming to achieve Universal Health Coverage (UHC) strategies. The duration of the course was 2 months. The survey participants in this study from the two courses comprised 13 people, comprising 25% of the total.

Table 14. Participated Program Category

Program Category	Number of Participants	Percentage (%)
Infectious Disease Specialists	21	40.4
Graduate Degrees	18	34.6
Health Policy Experts	13	25
Total	52	100

Note: This represents only participants from 4 of the 8 available program tracks, limited to 2023 graduates.

The regional representation of the participants was balanced between African (50%) and Asian (50%) countries. Table 15 illustrates the comprehensive geographical and gender distribution of the participants. The participants were recruited from 11 countries. The highest participation rate was found with Laos (21.2%), followed by Ghana (17.3%). This balanced regional representation ensured the applicability of the findings across diverse healthcare system contexts and cultural environments.

The program considers gender balance when selecting participants. As a result, the program could maintain over 40% of the female participants over the past five years. Table 16 reveals that 52 participants of the pilot survey achieved a perfect gender balance, with 26 males and 26 females, indicating no gender bias in the program's participation.

However, regionally, the gender composition diverged: 69.23% of male participants hailed from Africa compared to 30.77% from Asia, while 69.23% of female participants hailed from Asia versus 30.77% from Africa, indicating a potential bias due to gender variation among the regions. The majority (69.23%) of the male participants were from Africa and Majority (69.23%) were from Asia.

Table 15. Participants' Regional and Gender Distributions

Region	Number of Male Participants (%)	Number of Female Participants (%)	Total (%)
Africa	18 (69.23)	8 (30.77)	26 (100)
Asia	8 (30.77)	18 (69.23)	26 (100)
Total	26 (100)	26 (100)	52 (100)

*African countries include Ghana, Ethiopia, Tanzania, Uganda, and Mozambique, Asian countries include Laos, Mongolia, Vietnam, Cambodia, Indonesia, and Uzbekistan

Overall, 52 fellows from 11 countries equally split between Africa and Asia, with Laos contributing the highest share (21.2%), followed by Ghana the second highest (17.3%). Their professional backgrounds varied similarly, comprising 40.4% infectious disease specialists, 34.6% graduate degree holders, and 25% health policy experts. Gender distribution was

perfectly balanced at 50% males and 50% females overall, although regional dynamics differed: African participants were predominantly male (69.2%), while Asian participants were predominantly female (69.2%), highlighting important cultural and workforce factors to inform future recruitment strategies. Together, these demographic characteristics underscore a fellowship's commitment to diversity in geography, profession, and gender, thereby enhancing the applicability and equity of its capacity-building impact.

4.2.2. Descriptive Statistics of the Evaluation Items

The analysis of the pilot study results included descriptive statistics to evaluate responses to each questionnaire item. Table 16 lists the means and standard deviations for each question. The results indicated that the highest average scores and lowest variation were recorded for questions related to career advancement (mean = 4.5, SD = 0.61), indicating that most participants shared similar experiences regarding the fellowship program's impact on career advancement. Knowledge sharing through trainer sessions was also significantly high (mean = 4.08, SD = 0.93).

Conversely, the lowest mean score (3, SD = 1.508) and highest variation in responses were observed for the question related to academic forum presentations (Q5). The analysis revealed a notable disparity in participants' experiences regarding academic presentations, indicating that while some participants felt confident about their presentation skills, others encountered significant challenges.

Table 16. Descriptive Statistics of the Final Evaluation Items (N=52)

Items	Minimum	Maximum	Mean	SD values
Q1. Career Progress	3	5	4.50	.610
Q2. Action plan Implementation	1	5	3.42	1.177
Q3. Training of Trainers (ToT)	1	5	4.08	.926
Q4. Academic papers and reports	1	5	3.33	1.396
Q5 Academic forums	1	5	3.00	1.508
Q6. Proposed manuals, guidelines or programs	1	5	3.73	1.173
Q7. Collaborative projects Participation	1	5	3.48	1.336
Q8. Joint projects Proposal	1	5	3.02	1.475
Q9. Academic societies and conferences	1	5	3.87	1.299
Average			3.60	1.21

The immediate training outcomes of the LJWFP were assessed through two primary indicators: (1) Training Satisfaction and (2) application. Data collection for these indicators occurs at two time points: immediately upon training completion and within six months post-training, utilizing an online survey methodology. Both indicator surveys collect identical demographic information to that gathered in this study, thereby enabling participant tracking and matching of long-term and short-term outcome measures while maintaining participant anonymity and protecting personally identifiable information.

Table 17 presents the immediate training outcome indicators, including participant satisfaction upon program completion and workplace application within six months of graduation. These two indicators are the major evaluation criteria used for the LJWFP, which were collected every year for all participants. The satisfaction survey achieved a 98% response rate (51 out of 52 participants), whereas the application survey response rate declined slightly to 90.4% (47 out of 52 participants). The impact survey, conducted two years post-training, demonstrated a markedly low response rate of 52% (52 of 100 participants). This indicated a significant inverse relationship between the temporal distance from training completion and survey response rates, with participant engagement progressively diminishing.

The immediate training outcome indicators further reinforced the program's success, with a mean satisfaction score of 94.70, and an application score of 92.04 out of 100. These high scores reflect the participants' strong belief in their ability to apply what they learned, suggesting that the training was not only well-received but also practically relevant.

Table 17. Short-Term Training Evaluation Items

Items	Number of Participants (n=52)	Minimum (Out of 100)	Maximum	Mean	SD
Satisfaction	51	74.89	100	94.70	6.86
Application	47	77.78	100	92.04	7.57

4.2.3. Fellows' Positive Influence on Health System

Figure 4 illustrates the percentage of fellowship participants who gave positive ratings (4 or 5 on a 5-point Likert scale) across eight impact questions related to career progression, training activities, academic engagement, policy application, and action plan implementation. The highest positive response (96.23%) reflects participants' agreement that their careers have progressed since joining the program, followed by 75.47% who reported sharing knowledge through training-of-trainer sessions and 73.58% who participated in academic societies.
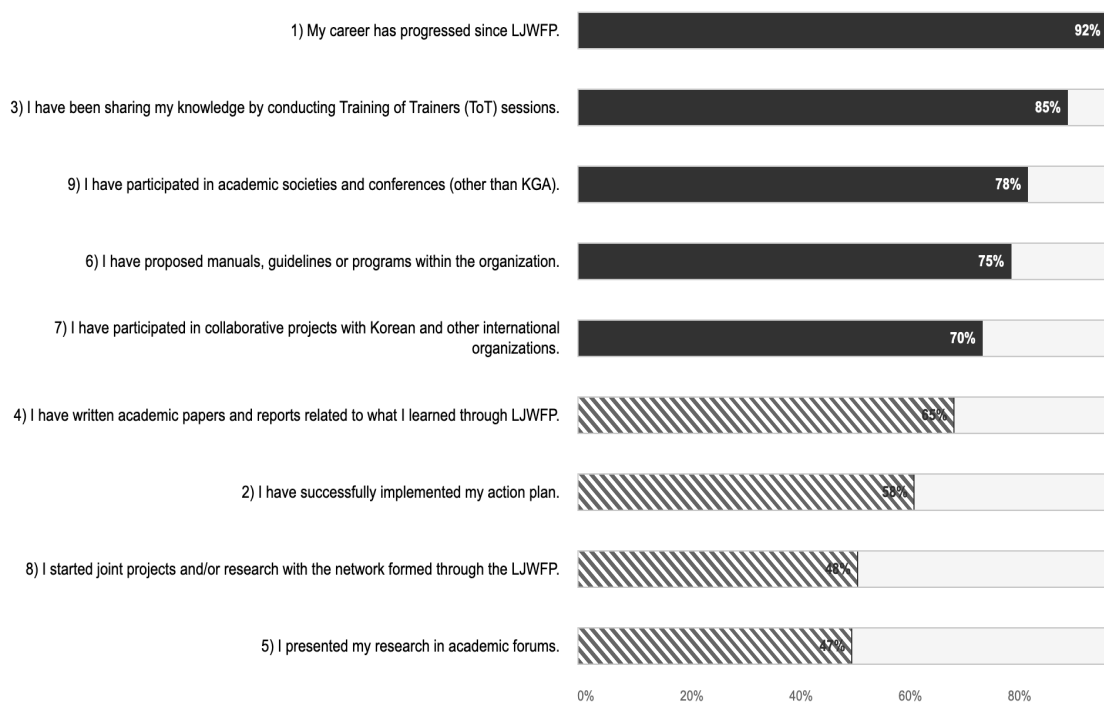
Figure 4. Percentage of Positive Responses on the Fellows' Influence

Approximately 66.04% of respondents indicated that they had proposed policy applications such as manuals and guidelines, and 58.49% took part in collaborative projects, demonstrating strong engagement in both policy development and teamwork activities. In contrast, only 47.17% of participants fully realized their action plans after training, with many citing financial difficulties and political bottlenecks as the main barriers to implementation.

The pilot test results provided quantitative evidence of the positive influence of fellows on health system improvement across the five key impact categories. Table 18 presents descriptive statistics of the identified impact factors. Career Progress achieved the highest mean score of 4.5 (SD = 0.61) on the five-point scale, indicating overwhelming participant reports of positive professional trajectory advancement following program completion. The relatively low standard deviation suggests a high consensus among respondents, with scores tightly clustered around the mean.

Policy Contribution (mean = 3.74, SD = 0.87) and Network Utilization (mean = 3.67, SD = 1.12) both received "Above Average" assessments. These results indicate substantial participant engagement in policy development activities and effective leveraging of professional networks established during the program. However, higher standard deviations, particularly for Network Utilization, indicated greater variability in the participant experience.

Academic Achievement (mean = 3.16, SD = 1.29) and Joint Projects Proposal (mean = 3.02, SD = 1.48) received "Moderate" ratings. Lower mean scores and higher standard deviations suggested greater outcome heterogeneity, with some participants achieving high

levels, whereas others reported minimal engagement, possibly due to differences in institutional support, individual motivation, or contextual barriers.

Table 18. Descriptive Statistics of the Identified Impact Factors

Factor	N	Mean	SD	Min	Max	Assessment
Career Progress	52	4.5	0.61	3	5	Excellent
Policy Contribution	52	3.74	0.87	1.67	5	Above Average
Network Utilization	52	3.67	1.12	1	5	Above Average
Academic Achievement	52	3.16	1.29	1	5	Moderate
Joint Projects Proposal	52	3.02	1.48	1	5	Moderate

4.2.4. Demographic Comparison of the Fellows' Impact

4.2.4.1. Regional Variations of the Fellows' Impact

A comparative analysis between Asian and African participants was conducted to provide valuable insights into the cross-cultural effectiveness of the fellowship program. Despite the potential differences in educational systems, healthcare infrastructure, and policy environments between these two regions, the study revealed consistent outcomes across all five impact domains.

Career Progress demonstrated the most striking uniformity, with both Asian and African participants achieving identical mean scores of 4.5 out of a possible scale. This perfect alignment (t = 0.00, p = 0.561) suggests that the fellowship program's career development components are equally effective, regardless of regional background, indicating a robust program design that transcends geographical and cultural boundaries. Policy Contribution showed the largest numerical difference between regions, with Asian participants achieving a higher mean score (3.92) than their African counterparts (3.56). However, this difference was not statistically significant (t = 1.5, p = 0.139) with a moderate effect size (Cohen's d = 0.42). This suggests that, while there may be practical differences in policy engagement opportunities between regions, the fellowship program provides sufficient foundational skills for meaningful policy contributions across both contexts.

Academic Achievement displayed a similar pattern, with Asian participants scoring higher (3.44) than African participants (2.88), yet remaining statistically non-significant (t = 1.58, p = 0.12). The moderate effect size (Cohen's d = 0.44) indicated a potentially meaningful practical difference that may warrant further investigation with larger sample sizes.

Interestingly, Network Utilization favored African participants slightly (3.75) over Asian participants (3.60), although this difference was minimal and non-significant (t = -0.49, p = 0.626). The small negative effect size (Cohen's d = -0.14) suggests that networking outcomes are remarkably consistent across regions. Joint Project Proposals showed marginal regional variation, with Asian participants (3.08) scoring marginally higher than African participants (2.96), resulting in negligible effect size (Cohen's d = 0.08).

Table 19 demonstrates that, despite varying healthcare systems, policy environments, and academic infrastructure between Asia and Africa, the fellowship program achieved consistent impacts across both regions. The absence of statistically significant differences (all p-values > 0.05) supports the program's universal applicability and suggests that the core competencies developed through fellowship translate effectively across diverse regional contexts.

Table 19. Regional Variations of Training Impact

Variable	Mean		t-value	p-value	Effect Size (Cohen's d)
	Asia	Africa			
Career Progress	4.5	4.5	0	1	0
Policy Contribution	3.92	3.56	1.5	0.139	0.42
Academic Achievement	3.44	2.88	1.58	0.12	0.44
Network Utilization	3.6	3.75	-0.49	0.626	-0.14
Joint Project Proposal	3.08	2.96	0.28	0.781	0.08

Note: all p > 0.05

4.2.4.2. Gender Variations of the Fellows' Impact

Table 20 shows the gender variations in the impact of the fellows. The gender-based analysis of the 26 male and 26 female participants represented a perfectly balanced sample that

provided insights into the program's gender inclusivity. The findings demonstrate gender parity across all measured domains, reflecting the program's success in creating an equitable learning environment that benefits participants, regardless of gender.

Career Progress showed similar outcomes between genders, with males achieving a mean score of 4.54 and females 4.46. The minimal difference ($t = 0.45$, $p = 0.654$) and small effect size (Cohen's $d = 0.13$) indicated that the fellowship program provided equal career advancement opportunities for both men and women. This finding is particularly significant, given the documented gender disparities in healthcare leadership and policy roles in partner countries.

Policy Contribution achieved perfect gender parity, with both male and female participants scoring identical means of 3.74. This complete alignment ($t = 0$, $p = 1$, Cohen's $d = 0$) demonstrates that the program successfully enables both genders to engage equally in policy development and implementation.

Academic Achievement displayed minimal gender variation, with males scoring slightly higher (3.25) than did females (3.08). However, this difference remained statistically insignificant ($t = 0.48$, $p = 0.633$), with a small effect size (Cohen's $d = 0.13$), indicating equivalent academic outcomes across genders. Network Utilization showed males achieved marginally higher scores (3.77) than females (3.58). Despite this numerical difference, the statistical analysis revealed no significant variation ($t = 0.61$, $p = 0.543$) with a small effect size (Cohen's $d = 0.17$). This suggests that both genders benefit equally from the program's networking opportunities and develop comparable professional connections.

Joint Project Proposals demonstrated males scored slightly higher (3.12) than females (2.92). However, this difference was not statistically significant (t = 0.47, p = 0.643) with a small effect size (Cohen's d = 0.13), indicating that collaborative project development skills were equally enhanced across genders.

Table 20. Gender Variations of the Fellows' Impact

Variable	Mean		t-value	p-value	Effect Size (Cohen's d)
	Male	Female			
Career Progress	4.54	4.46	0.45	0.654	0.13
Policy Contribution	3.74	3.74	0	1	0
Academic Achievement	3.25	3.08	0.48	0.633	0.13
Network Utilization	3.77	3.58	0.61	0.543	0.17
Joint Project Proposal	3.12	2.92	0.47	0.643	0.13

Note: all p > 0.05

4.2.4.3. Program-Type Variations of Fellows' Impact

The survey results revealed distinct performance patterns across the three program types when measured on a 5-point Likert scale across the five impact categories. Graduate Degree programs achieved the highest overall average performance at 4.00, followed by Infectious Disease programs at 3.57, and health policies and management programs at 3.50.

Figure 5 shows the variation in the impact of fellows between different program types for each impact category. Career Progress demonstrated the most consistent performance across all program types, with Graduate Degrees scoring 4.61, Health Policy & Management at 4.46, and Infectious Disease at 4.43, resulting in the smallest range of variation (0.18) among all measured categories. Policy Contribution showed the largest performance gap between programs, with Graduate Degrees achieving 4.24, Health Policy & Management at 3.87, and Infectious Disease at 2.95, creating a range of 1.00 points between the highest and lowest performing programs. Academic Achievement exhibited the greatest variation across program types, with a range of 1.39 points separating Graduate Degrees (4.06) from Infectious Disease programs (2.67), while Health Policy & Management programs scored 2.73. Network Utilization favored Infectious Disease programs with the highest score of 3.86, compared to Health Policy & Management at 3.73, and Graduate Degrees at 3.42, representing a moderate range of 0.44 points. Joint Project Development showed identical performance for Graduate Degrees and Infectious Disease programs at 3.67, while Health Policy & Management programs scored significantly lower at 2.71, creating a range of 0.96 points. The mean scores across all programs ranged from 4.50 for Career Progress to 3.35 for Joint Project Development, indicating that career advancement outcomes were most consistently achieved across all program types.

Overall, Career Progress showed minimal variation, Policy Contribution exhibited the largest gap, Academic Achievement varied markedly, Network Utilization favored infectious diseases, joint project development showed parity between graduate degrees and infectious diseases, and health policy and management trials.
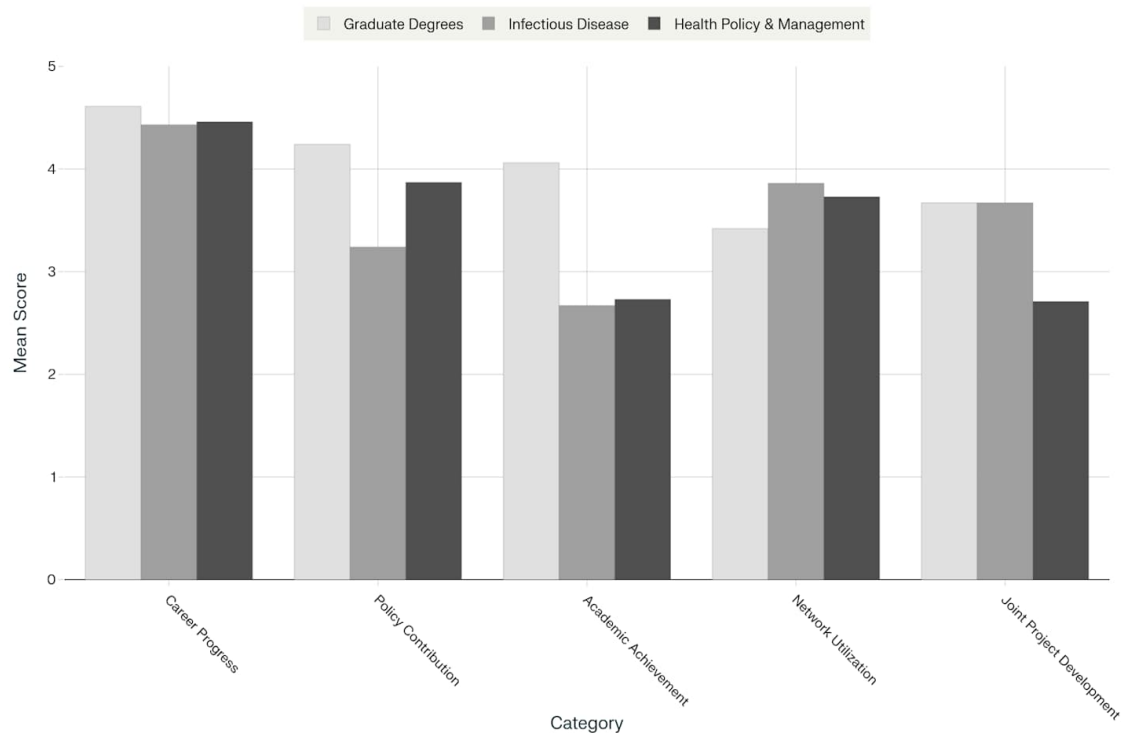
Figure 5. Impact Variation by Program Types

To determine the impact variations between the different program types, a One-way ANOVA was conducted. Table 21 shows the results of the impact according to program type. There were significant differences in Academic Achievement between the programs ($F_{(2, 49)}$ = 8.430, p = 0.001), with program type accounting for approximately 25.6% of the variance. Graduate Degree participants demonstrated markedly superior academic performance (mean = 4.06) compared with Infectious Disease (mean = 2.67) and Health Policy participants (mean = 2.73).

Policy Contributions also revealed significant between-program differences ($F(2,47)$ = 10.495, $p < 0.001$), with program type explaining approximately 30.9% of the variance1. Graduate Degrees (mean = 4.14) and Health Policy and Management programs (mean = 4.00) demonstrated substantially higher policy contributions than the Infectious Disease program (mean = 2.95).

Table 21. One-way ANOVA Result of Training Impacts by Program Type

Category	Program	Mean	F-value	p-value
Career Progress	Graduate Degrees	4.61	0.458	0.635
	Infectious Disease	4.43		
	Health Policy & Management	4.46		
Policy Contribution	Graduate Degrees	4.24	**8.539**	**0.001***
	Infectious Disease	3.24		
	Health Policy & Management	3.87		
Academic Achievement	Graduate Degrees	4.06	**8.544**	**0.001***
	Infectious Disease	2.67		
	Health Policy & Management	2.73		
Network Utilization	Graduate Degrees	3.42	0.760	0.473
	Infectious Disease	3.86		
	Health Policy & Management	3.73		
Joint Project Development	Graduate Degrees	3.67	2.865	0.067
	Infectious Disease	3.67		
	Health Policy & Management	2.71		

Significant differences (p < 0.05) based on Scheffe post hoc tests

4.2.5. Correlation between Personal Achievement and Policy Impact

Correlation analysis of personal achievement and policy impact relationships revealed interconnections between individual career advancement and broader policy contributions. Career progression emerged as a key predictor of policy engagement, with strong positive correlations indicating that, as individuals advance professionally, they are more likely to contribute meaningfully to policy development and implementation.

Table 22 displays the correlation coefficients (r) for the five impact variables. Career Progress showed strong positive correlations with Policy Contribution ($r = 0.466$, $p < 0.001$) and Joint Projects ($r = 0.512$, $p < 0.001$), indicating that career advancement was linked to greater policy engagement and collaboration. Its moderate correlation with Network Utilization ($r = 0.329$, $p = 0.017$) suggests that professional advancement enhances networking capabilities. These findings highlight the interconnectedness of these variables and suggest that fostering career development may lead to greater contributions to policies and collaborative projects.

Table 22. Correlation among Impact Variables

		Career Progress	Academic Achievement	Policy Contribution	Network Utilization	Joint Project Proposal
Career Progress	Correlation					
	P-Value					
Academic Achievement	Correlation	.405**	1			
	P-Value	0.003				
Policy Contribution	Correlation	.466**	.459**	1		
	P-Value	0.000	0.001			
Network Utilization	Correlation	.329*	0.220	0.169	1	
	P-Value	0.017	0.117	0.230		
Joint Projects Proposal	Correlation	.512**	.441**	.466**	0.175	1
	P-Value	0.000	0.001	0.000	0.214	

Note: Pearson's Correlation, *p < 0.05, **p < 0.005

The relationship between personal achievement and policy impact was further strengthened by the evidence that career advancement facilitated greater participation in joint projects and collaborative initiatives. This finding suggests that individual professional development serves as a catalyst for broader systemic contributions, supporting the program's theory of change, which invests in individual capacity building to achieve larger health system improvements.

However, the analysis also revealed that academic achievement, while valuable, did not demonstrate equally strong relationships with policy contributions. This finding indicates that different types of personal advancement may have varying effects on policy engagement, with career progression being more directly linked to policy influence than to academic output alone.

4.2.6. Short-term Evaluation Results and Long-term Impact

Regression analysis was conducted to examine the relationships between short-term indicators—satisfaction and application of learning—and long-term outcomes across the five impact factors.

Table 23 reports the standardized regression coefficients (β) linking short-term indicators—Satisfaction and Application of Learning—to five long-term impact domains. The application of learning significantly predicted Policy Contribution ($\beta = 0.382$, $p = 0.010$) and Career Progression ($\beta = 0.299$, $p = 0.049$), highlighting that practical application drove sustained policy engagement and professional advancement.

However, satisfaction did not significantly predict any long-term outcomes, suggesting that while participant satisfaction reflects positive training experiences, it does not necessarily translate into a sustained impact. Academic Achievement, Network Utilization, and Joint Project Proposals showed no significant relationships with either short-term indicator, indicating that these outcomes may be influenced by factors beyond immediate post-training measures.

Table 23. Regression analysis between Short-term and Long-term Indicators

Dependent Variable	Independent Variable	β	t	p
Policy Contribution	Cont.			0.416
	Satisfaction	0.109	0.762	0.45
	Application	**0.382***	2.674	**0.010***
Academic Achievement	Cont.		-0.827	0.412
	Satisfaction	0.288	1.909	**0.063**
	Application	0.033	0.219	0.827
Network Utilization	Cont.		0.489	0.627
	Satisfaction	-0.006	-0.036	0.972
	Application	0.191	1.23	0.225
Career Progression	Cont.		0.815	0.42
	Satisfaction	0.12	0.815	0.419
	Application	**0.299***	2.026	**0.049***
Joint Project Proposals	Cont.		-0.572	0.57
	Satisfaction	0.061	0.397	0.693
	Application	0.213	1.387	0.172

$p < 0.05$

Figure 6 is a visual representation that clearly demonstrates the relationship between the Application of Learning scores and various long-term impact variables. Solid lines represent statistically significant relationships between Application and Policy Contribution, and between Application and Career Progression, showing positive upward trends. Dotts

71

represent each participant's score. The steeper slope for Policy Contribution ($\beta = 0.382$) compared to Career Progression ($\beta = 0.299$) is visually apparent, reflecting a stronger predictive relationship.


Application vs. Career Development Regression Coefficient $\hat{I}^2 = 0.30$ (p = 0.049)

Application vs. Policy Contribution Regression Coefficient $\hat{I}^2 = 0.38$ (p = 0.010)
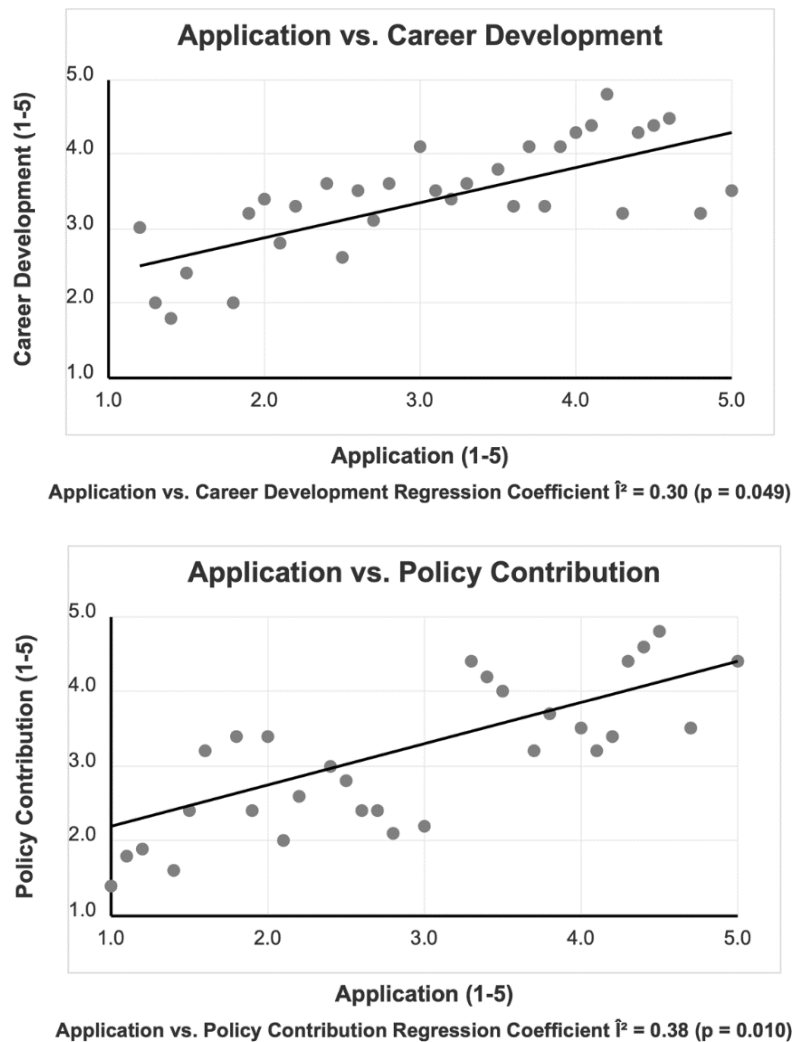
Figure 6. Application vs. Impact Variables

4.3. Qualitative Findings: Contextual Insights for Long-term Impact

In addition to the 5-point Likert scale quantitative survey results, most participants (46 out of 52, 88% of the total) provided open-ended responses during the pilot testing of the long-term impact assessment tool. While these qualitative responses were not the primary focus of this study, they provided valuable contextual insights that demonstrated the tool's capacity to capture supplementary information beyond quantitative metrics. Table 24 presents a thematic analysis of the responses, which offers important perspectives on how qualitative data can enhance the comprehensiveness of long-term impact assessments in fellowship programs.

Table 24. Qualitative Key Themes and Quotations from Pilot Testing

Theme	Sub-theme	Example Quote
Implementation Challenges	Financial constraints	"Due to limited funding in our institution, we have been unable to procure the necessary equipment to fully implement the diagnostic protocols learned during the fellowship."
	Resource availability	"In our medical education program, we lack essential learning materials such as updated textbooks and computers for students, which significantly impacts our ability to implement modern teaching methods."
	Stakeholder engagement	"The formation of Technical Working Groups has been essential in bringing together various stakeholders to contribute to our national immunization program improvements."
Positive Outcomes	Skill development	"The fellowship has fundamentally transformed my approach to nursing education. I have successfully established a new training curriculum that incorporates evidence-based practices learned during the program."

	Community Impact	"We have observed significant changes in community engagement with preventive health care services. Before implementing the action plan, attendance at educational sessions was minimal, but now we regularly have full participation from community members."
	Networking and collaboration	"The professional connections I made during the fellowship were invaluable. I am now collaborating with colleagues from three different countries on a joint research project addressing antimicrobial resistance."
Requests for Further Support and Future Directions	Funding and resources	"To sustain the program's impact, we need continued financial support specifically for laboratory equipment maintenance and staff training updates."
	Training opportunities	"Additional advanced training in health economics would help us better implement cost-effectiveness analyses in our healthcare decision-making processes."
	Sustainability	"We need to develop mechanisms that ensure the continuation of these improvements beyond initial implementation. This requires both institutional commitment and ongoing technical support."
	Regional/program variation	"The significantly higher policy contributions from Asian participants compared to their African counterparts suggest that regional contextual factors may influence policy implementation success."
	Academic/professional growth	"Graduate degree programs appear to be particularly effective in fostering academic output and policy contributions, possibly because of their comprehensive nature and longer duration."

4.3.1 Tool Performance: Capturing Implementation Context

The assessment tool successfully elicited detailed responses regarding implementation challenges, demonstrating its ability to provide a contextual understanding of the quantitative outcomes measured. The tool's open-ended questions revealed three primary contextual factors that influenced long-term impact measurements:

- Financial Constraints: The tool effectively captured how resource limitations affect program sustainability, with seven out of 44 respondents providing specific examples. These insights demonstrate that the assessment tool can identify external factors that may influence the measured outcomes.

"Due to limited funding in our institution, we have been unable to procure the necessary equipment to fully implement the diagnostic protocols learned during the fellowship." (Participant from Ghana)

- Resource Availability: The tool's capacity to identify systemic barriers was evident in response to the lack of essential resources. This contextual information helps to interpret quantitative impact scores by revealing the underlying implementation challenges.

"In our medical education program, we lack essential learning materials such as updated textbooks and computers for students, which significantly impacts our ability to implement modern teaching methods." (Participant from Laos)

- Stakeholder Engagement: The assessment tool captured the importance of collaborative networks in sustaining program impact, providing insights into factors that enhance or hinder long-term effectiveness.

"The formation of Technical Working Groups has been essential in bringing together various stakeholders to contribute to our national immunization program improvements." (Participant from Ethiopia)

4.3.2 Tool Validation: Confirming Positive Impact Measurement

Qualitative responses validated the tool's ability to detect and measure positive outcomes, providing narrative evidence that supports quantitative impact scores. This triangulation strengthens the tool's credibility for long-term impact assessment.

- Skill Development Documentation: This tool successfully captured detailed accounts of knowledge and skill improvements, providing a rich context for interpreting quantitative learning outcome scores.

  "The fellowship has fundamentally transformed my approach to nursing education. I have successfully established a new training curriculum that incorporates evidence-based practices learned during the program." (Participant from Mongolia)

- Community Impact Verification: Qualitative responses confirmed that the tool can effectively measure broader community-level impacts, extending beyond individual participant outcomes.

  "We have observed significant changes in community engagement with preventive health care services. Before implementing the action plan, attendance at educational sessions was minimal, but now we regularly have full participation from community members." (Participant from Ethiopia)

- Network Effect Measurement: The tool demonstrated its capacity to capture collaborative outcomes and professional networking benefits, which are often difficult to quantify but crucial for long-term program impacts.

"The professional connections I made during the fellowship were invaluable. I am now collaborating with colleagues from three different countries on a joint research project addressing antimicrobial resistance." (Participant from Cambodia)

4.3.3 Tool Refinement: Identifying Areas for Future Development

Pilot testing revealed that the assessment tool could be further enhanced to better capture sustainability factors and long-term support needs. The participants' feedback suggested areas where the tool could be refined.

- Sustainability Measurement: Responses indicated that the tool could benefit from more specific questions about long-term sustainability planning and institutional support mechanisms.

"We need to develop mechanisms that ensure the continuation of these improvements beyond initial implementation. This requires both institutional commitment and ongoing technical support." (Participant from Uzbekistan)

- Resource Need Assessment: The tool's ability to identify ongoing support requirements demonstrates its potential value for program administrators in planning continued assistance.

"To sustain the program's impact, we need continued financial support specifically for laboratory equipment maintenance and staff training updates." (Participant from Tanzania)

- Advanced Training Gap Identification: The tool effectively captured the need for continued professional development, suggesting its utility in identifying follow-up program requirements.

"Additional advanced training in health economics would help us better implement cost-effectiveness analyses in our healthcare decision-making processes." (Participant from Indonesia)

4.3.4 Implications for Long-term Impact Assessment Tool Development

The qualitative data obtained during pilot testing demonstrated that the assessment tool has significant potential for providing comprehensive long-term impact evaluations. The rich contextual information captured through open-ended responses serves as a valuable complement to quantitative metrics, offering several advantages for impact assessments.

- Contextual Understanding: The tool's ability to capture implementation challenges and success factors provides an essential context for interpreting quantitative impact scores.

- Validation of Quantitative Findings: Qualitative responses serve as triangulation data, confirming and elaborating on the quantitative outcomes measured by the tool.

- Identification of Moderating Factors: The tool successfully identified external factors (funding, resources, stakeholder engagement) that influence program impact, which is crucial for a comprehensive evaluation.

- Sustainability Planning: The assessment tool's capacity to identify ongoing support needs and sustainability challenges makes it valuable for program improvement and future planning.

These findings suggest that incorporating both quantitative and qualitative components into long-term impact assessment tools enhances their utility for comprehensive program evaluation and continuous improvement in fellowship programs.

5. Discussions and Conclusion

5.1. Discussions on the Findings

This study contributes to the training evaluation theory by demonstrating that a concise, nine-item instrument can validly measure the complex, multidimensional impacts of international capacity-building programs (Williams et al., 2010). The tool's construct validity, confirmed through exploratory factor analysis, supports the theoretical proposition that professional development, policy contribution, academic output, network utilization, and project initiation represent distinct, yet interrelated domains of impact.

The pilot test results provide compelling evidence of the effectiveness of the LJWFP in achieving its stated objectives of building healthcare leadership capacity in LMICs. The consistently high scores across impact domains, particularly in career progression where participants achieved near-universal positive outcomes, indicate that the program successfully translated training investments into tangible professional advancement. This finding is particularly significant, given the program's focus on developing change agents capable of driving health system improvements in their home countries.

The absence of significant demographic variations in most impact outcomes suggests that the program's design and implementation are equitable across gender lines. While not statistically significant, some interesting patterns emerged, where male participants showed slightly higher performance in academic and policy contributions, while female participants demonstrated somewhat stronger occupational achievements. This finding supports the

program's inclusive approach and indicates that benefits are consistently realized, although gender-specific support mechanisms may warrant further investigation.

However, significant regional variations were observed, with Asian participants demonstrating significantly higher policy contribution scores than their African counterparts. This finding highlights the importance of contextual factors in training and impact realization. Asian participants may benefit from more established institutional environments and clearer policy change pathways, while African participants may face greater resource constraints and infrastructure limitations that affect policy implementation capacity.

The significant variations observed between the different program tracks underscore the importance of tailoring evaluation approaches to specific training objectives and content areas. The degree programs showed the highest overall impact, which correlates with their extended duration (9-10 times longer than short-term programs) and correspondingly higher budget allocation, although cost-effectiveness analysis is needed to determine the optimal resource allocation.

The strong inter-factor correlations between career progression and policy engagement substantiate theories of capacity transfer, which posit that individual skill advancement catalyzes systemic changes in health systems (Frenk et al., 2010). Professional development has emerged as a critical mediating factor driving systemic change, with strong interconnections between career progression, policy contribution, and collaborative project participation. This supports the program's underlying theory of change that individual capacity-building leads to broader systemic contributions (Nakanjako et al., 2015).

The relationship between short-term indicators and long-term impacts provides valuable insights into program improvement (Phillips et al., 2012).

Training satisfaction was measured as part of the "Response" category within the product evaluation area. This aligns with Kirkpatrick's Level 1 evaluation, which measures whether learners find training engaging, favorable, and relevant to their jobs. The satisfaction indicators focused on participants' immediate reactions to the training experience, including their perception of relevance, training methods, trainers, and overall program quality (Yoon & Oh, 2024).

The program's satisfaction evaluation encompassed several key dimensions based on established training evaluation principles: overall satisfaction with the training experience, perceived relevance of training content to participants' professional needs, quality of instructional delivery and materials, appropriateness of training duration and pace, and motivation for continued learning and professional development.

Application indicators are designed to measure behavioral changes and the practical implementation of acquired knowledge and skills. These indicators align with Kirkpatrick's Level 3 evaluation, which assesses whether participants apply what they have learned during training when they return to their workplace. The application indicators are categorized under the "Behavior" evaluation area and include multiple components: Job Performance Improvement-Measures the degree to which participants apply trained performance in their work settings, Continuity of Trainees' Roles: Assesses whether participants maintain and

expand their professional responsibilities, Continuity of Exchange: Evaluates ongoing professional relationships and knowledge sharing

The program's application indicators specifically focus on sustainable behavioral changes that extend beyond the immediate training period. Research on healthcare training effectiveness emphasizes that the successful application of learning requires not only knowledge acquisition but also the practical ability to implement new skills in workplace contexts. The fellowship program addresses this through comprehensive post-training management programs that support participants in applying for their learning.

The significant positive relationship between the application of learning and both policy contribution and career progression underscores the importance of practical skill application in achieving a sustained impact. This finding suggests a clear pathway from knowledge application to professional advancement to policy influence, which has important implications for program design and post-training support.

Conversely, the lack of significant relationships between satisfaction and long-term outcomes challenges the common practice of relying heavily on post-training satisfaction surveys as the primary indicators of training success (김슬기, 2020). Instead, organizations should develop more sophisticated measures of learning applications to better predict and enhance long-term training effects.

Despite strong knowledge-sharing outcomes, this study identified a persistent knowledge-to-action gap that characterizes many capacity-building initiatives in LMIC

contexts (Turner et al., 2021). The findings regarding action plan implementation challenges, particularly financial constraints and political bottlenecks, align with the broader literature on barriers to training transfer in developing countries (Cho et al., 2013). These findings highlight the need for enhanced post-training support mechanisms and institutional engagement to maximize the realization of training benefits.

The analysis revealed several categories of implementation barriers that significantly affected program effectiveness. Structural constraints encompass governance issues such as centralized decision-making processes and rigid bureaucratic systems; resource limitations including insufficient human resources, budget constraints, and technological infrastructure gaps; and institutional resistance characterized by organizational inertia and lack of change management systems.

Environmental factors also play a crucial role in determining implementation success. Country-specific characteristics, such as Human Development Index (HDI) levels, governance quality, and the presence of international partnerships, significantly influence implementation success. Regional differences emerged as particularly important, with Asian countries showing advantages in policy pathway clarity, whereas African countries demonstrated strengths in community-based implementation approaches.

The findings of this study suggest a need for region-specific approaches (Kang et al., 2024). Asian regional characteristics reveal strengths in established institutional environments and clear policy change pathways, but face constraints from bureaucratic rigidity and hierarchical decision-making processes. The primary influence pathway follows a pattern of

policy to system to field implementation, suggesting that strategic focus should be on enhancing policy influence strategies.

In contrast, African regional characteristics demonstrate strengths in community-centered approaches and field application adaptability, while facing constraints from resource limitations and infrastructure deficiencies. The primary influence pathway follows a pattern from field to community to policy, indicating that strategic focus should emphasize resource-constraint mitigation strategies.

5.2. Discussions on the Methods

The development and validation of this comprehensive evaluation tool represent a significant advancement in the systematic assessment of international healthcare training programs. This study addresses a critical gap in the evaluation literature by providing a robust and theoretically grounded framework for measuring the long-term impacts of capacity-building initiatives in global health settings.

The refined measurement tool demonstrates theoretical contributions through the integration of established evaluation frameworks. While closely aligning with Kirkpatrick's four-level model, this study extends its scope to mid- and long-term outcomes (Levels 3-4), thereby addressing the common shortfalls in reaction and learning evaluations that dominate current practice (Brian, 2021). Cultural contextualization of LMICs represents a crucial advancement, as the original Kirkpatrick model was developed primarily for corporate training environments in developed countries.

The incorporation of Kaufman's five-level framework, particularly the societal outcome dimension (Level 5), enabled the tool to capture both organizational changes and broader public health impacts. However, this study reveals that the societal impact dimension requires careful adaptation to LMIC contexts, where structural constraints and governance challenges significantly influence the translation of individual capacity gains into systemic change. Stufflebeam's CIPP model informed the tool's contextual and input evaluation phases, ensuring that the evaluation items reflected the program design quality and implementation processes (Rahimzadeh, 2019). The post-training follow-up evaluation component, strengthened by this integration, provides crucial insights into the sustainability of training investments. Furthermore, by adopting the WHO guidelines for healthcare training ROI, the instrument emphasizes public health returns rather than purely financial metrics, thus aligning with sector-specific evaluation priorities (WHO, 2010).

The methodological process employed in this study, particularly the sequential application of focus group interviews and Delphi rounds, ensured both content validity and expert consensus in item selection (Oh & Yoon, 2024). The achievement of high Content Validity Ratios (CVR=0.73) and strong internal consistency reliability demonstrate that the resulting nine-item framework effectively captures the multidimensional nature of training program impacts while remaining practical for implementation across diverse contexts (Lawshe, 1975).

The mixed-methods approach proved particularly valuable for validating the multidimensional evaluation framework. The Analytic Hierarchy Process (AHP) analysis demonstrated strong consistency ratios (<0.1), while the exploratory factor analysis explained

57.2% of the total variance, confirming the robustness of the four core impact factors: academic achievement, policy contribution, network utilization, and professional development (Hair et al., 2019). This methodological triangulation represents a significant advancement in international health ODA evaluation frameworks (Vallejo, 2016).

5.3. Limitations

This study had several limitations. The relatively small sample size of 52 participants, which is appropriate for pilot testing, limits the generalizability of our findings (Tabachnick & Fidell, 2019). For a more robust statistical analysis and generalization of results, a minimum of three years of post-graduation cohorts should be surveyed to achieve an adequate sample size.

The self-reported nature of impact measures introduces potential bias, although this limitation is common in evaluation studies of this type (Rotem, 2010). Future research should incorporate external validation measures such as supervisor assessments or institutional outcome indicators to strengthen the evaluation framework and reduce reliance on self-reported measures. Direct observation or objective verification of achievements was not feasible because of budget and data access limitations.

The two-year timeframe for impact assessment, while longer than typical evaluation periods, may still be insufficient to capture the full scope of long-term impacts, particularly those related to policy implementation and systemic change (사업평가실, 2010). The declining response rates over time, from 99% immediately post-training to 86% at six months to 53% at two years, suggest the need for enhanced motivation strategies for long-term participation.

5.4. Implications

5.4.1. Long-term Impact Development Model

This study proposes a time-staged impact development model that recognizes the different timelines required to manifest various types of impacts. Short-term impacts occurring within the first year focus on individual capacity enhancement, including knowledge, skills, and attitude changes. The evaluation focuses during this period on competency development and immediate learning outcomes.

Medium-term impacts, spanning to 1-3 years, encompass organizational change and diffusion processes. During this phase, participants initiate organizational change initiatives, engage in peer knowledge diffusion, and make policy attempts. The evaluation focus shifts to knowledge dissemination through networks, academic activities, and educational initiatives, which demonstrate the spreading influence of acquired knowledge and skills.

Long-term impacts, occurring after three years, involve institutional impacts and system changes. This phase includes policy implementation, institutional reform, and standardization efforts, which represent the ultimate goals of capacity-building programs. The evaluation focus during this period emphasizes policy, institutional, and standardization impacts that demonstrate lasting systemic change.

This staged approach provides a framework for designing evaluation strategies that align with realistic timelines for different types of impacts in LMIC contexts, ensuring that evaluation efforts are appropriately timed and focused on the achievable outcomes at each stage.

5.4.2. Future research directions

The framework suggested by this study emphasizes long-term, multidimensional impact assessment, representing a paradigm shift from traditional, short-term evaluation approaches to more comprehensive, sustainable impact measurement. This shift is essential for demonstrating the true value of and return on investment in international development training programs in the health sector.

However, the lower reliability observed for Career Development suggests that this construct may require further refinement (Field, 2018). The grouping of career advancement with joint project proposals under a single factor may not be theoretically coherent, as these represent different types of outcomes with potentially different timelines for realization.

Program-specific evaluation items should be developed or differential weighting systems should be implemented to accurately measure the impact across different training objectives and content areas (허윤선 et al., 2017). A cost-effectiveness analysis incorporating return on investment (ROI) calculations should be systematically conducted rather than simple per-capita comparisons (이은석, 2022).

Future research should consider expanding the sample size and extending the evaluation timeframe to capture longer-term impacts more comprehensively (Kara DeCorby-Watson et al., 2018). A longitudinal study design tracking participants over 5-10 years would provide more definitive evidence of a sustained impact.

5.4.3. Strategic Recommendations

This study also provides several strategic recommendations: context-specific approaches, knowledge-to-action bridges, and customizations based on regions and programs.

Effective program implementation requires conducting a pre-program national context analysis to understand governance structures, resource constraints, and change pathways within each participating country. This analysis should inform the setting of objectives aligned with participants' authority levels and institutional positions, ensuring that training goals are realistic and achievable within the existing frameworks. In addition, developing integrated approaches that link education, application, and ongoing support creates a continuum of learning that extends beyond the formal training period.

Addressing the persistent knowledge-to-action gap requires implementing systematic post-training support mechanisms that provide ongoing assistance to participants, as they attempt to apply their learning in real-world contexts. Establishing mentoring relationships with experienced practitioners creates valuable guidance networks that can help to navigate implementation challenges. Creating communities of practice for ongoing knowledge exchange maintains connections among participants and facilitates continued learning and problem solving. Furthermore, providing small grants or resources for pilot implementation projects gives participants the means to test and refine their ideas in a practical setting.

Recognizing the significant regional variations identified in this study, programs should develop region-specific curriculum components that address the unique challenges and opportunities present in different geographic and cultural contexts. Establishing regional

alumni networks for peer support and collaboration creates sustainable support systems that extend beyond the formal program period. Additionally, adapting evaluation frameworks to reflect regional impact pathways and timelines ensures that assessment approaches are appropriate for the specific contexts in which the participants operate.

The significant variations observed across different program tracks necessitate program-specific customization strategies to optimize the outcomes for each training modality. For infectious disease specialist programs, customization should focus on strengthening the integration between academic research and practical field applications while developing robust alumni engagement mechanisms to sustain long-term network connections. This includes creating specialized mentorship programs that pair recent graduates with experienced practitioners and establishing research collaboration platforms that facilitate ongoing academic partnerships.

5.5. Conclusion

This study successfully developed and validated a comprehensive nine-item evaluation tool for international health policy expert training programs through systematic expert consensus and empirical validation, demonstrating its effectiveness in building sustainable capacity across multiple dimensions. This research aims to contribute to international development evaluation by providing a replicable, evidence-based methodology for assessing long-term training impacts that extend beyond traditional short-term metrics.

Pilot testing revealed excellent program performance, with participants showing significant career advancement and meaningful contributions to their national health systems. These findings align with established research demonstrating that ODA training programs have substantial effects on career development, knowledge sharing, and policy contributions (Bastos et al., 2013; Nakanjako et al., 2015). The analysis revealed that participants achieved substantial performance across five interconnected factors: academic achievement, policy contribution, network utilization, professional development, and project collaboration.

The identification of professional development as a key mediating factor provides valuable strategic insights for designing effective capacity-building interventions. The program's underlying theory of change that individual capacity-building leads to broader systemic contributions was empirically validated through the strong interconnections observed between career progression, policy contribution, and collaborative project participation.

Furthermore, the program's equitable impact across demographic groups validates its inclusive design approach, though regional variations highlight the importance of contextual

adaptation. Performance variations across program tracks and regions underscore the need for differentiated evaluation approaches and targeted support strategies tailored to specific contexts and training objectives.

Despite strong knowledge-sharing outcomes, practical implementation barriers revealed a persistent knowledge-to-action gap that characterizes many capacity-building initiatives. Resource constraints and institutional resistance emerged as primary obstacles to translating acquired knowledge into practice. The findings suggest that practical skill application serves as a critical factor in achieving sustained impact, particularly for policy contributions and career advancement.

The multidimensional approach successfully captured the complex nature of training program outcomes in global health contexts. The validated framework offers a practical tool for ongoing program assessment and improvement, enabling systematic tracking of sustained outcomes essential for demonstrating return on investment in global health workforce development.

This research aims to contribute to the ODA training evaluation theory by demonstrating that comprehensive, long-term impact assessment is both feasible and essential for understanding the true value of international capacity-building investments. The integration of multiple theoretical frameworks and mixed-method validation provides a robust foundation for future evaluation efforts in similar contexts.

Reference

Allen, J. T., Laks, S., Zahler-Miller, C., Rungruang, B. J., Braun, K., Goldstein, S. R., & Schnatz, P. F. (2023). Needs assessment of menopause education in United States obstetrics and gynecology residency training programs. Menopause, 30(10), 1002–1005.

Baek, S.-G., Li, H., & Yun, G. (2024). Development and validation of a global competence scale for schoolteachers. The SNU Journal of Education Research 33(3): 1–30.

Bastos, L. F., Ciampone, M. H., & Mira, V. L. (2013). Assessment of evaluation of transference support and training impact on the work of nurses. Revista Latino-Americana de Enfermagem, 21(6), 1274–1281.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246.

Brian, W. (2021). What's your formula?: Combine learning elements for impactful training. Association for Talent Development.

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd ed.). Guilford Press.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 136–162). Sage.

Buttenheim, A. M. (2023). Training health policy researchers on policy engagement and research translation for greater impact: Evaluation of the Amplify@LDI Program. Journal of General Internal Medicine, 38(13), 3073.

Cho, W. (2023). 국제개발협력사업 수행 주체로서의 대학의 역할 고찰: 국내 학술지 논문을 중심으로 [A study on the roles of universities as actors in international development cooperation projects: Focusing on domestic research]. Global Creative Leader: Education & Learning, 13(3), 185.

Cho, Y., Kalomba, D., Mobarak, A. M., & Orozco-Olvera, V. (2013). Gender differences in the effects of vocational training: Constraints on women and drop-out behavior. World Bank Policy Research Working Paper (6545).

CIPP 모형을 활용한 R&D 과제기획 선정평가 도구 개발 연구: K 연구원 중심으로 [Development of selection evaluation tool for R&D project planning using CIPP model: Focusing on K Research Institute]. (2023). 기술혁신연구, 31(3), 171.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research & Evaluation, 10(7), 1–9.

Czabanowska, K., & Rodriguez Feria, P. (2024). Training needs assessment tools for the public health workforce at an institutional and individual level: A review. European Journal of Public Health, 34(1), 59–68.

Damalie Nakanjako, Namagala, E., Semeere, A., Kigozi, J., Sempa, J. B., Ddamulira, J. B., Katamba, A., Biraro, S., Naikoba, S., Mashalla, Y., Farquhar, C., Afya Bora Consortium members, & Sewankambo, N. (2015). Global health leadership training in resource-limited settings: A collaborative approach by academic institutions and local health care programs in Uganda. Human Resources for Health, 13, 1.

Endale Hadgu Gebregzabher, F. T., Waqtola Cheneke, A. E., Negesso, A. E., & Kedida, G. (2023). Continuing professional development (CPD) training needs assessment for medical laboratory professionals in Ethiopia. Human Resources for Health.

Rodríguez DC, Jessani NS, Zunt J, Ardila-Gómez S, Muwanguzi PA, Atanga SN, Sunguya B, Farquhar C, Nasuuna E. Experiential Learning and Mentorship in Global Health Leadership Programs: Capturing Lessons from Across the Globe. Ann Glob Health. 2021 July 12;87(1):61.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299.

Field, A. (2018). Discovering statistics using IBM SPSS statistics (5th ed.). SAGE Publications.

Frantz, J. M. (2015). The impact of a faculty development programme for health professions educators in sub-Saharan Africa: An archival study. BMC Medical Education, 15, 28.

Frenk, J., Chen, L., Bhutta, A. Z., Cohen, J., Crisp, N., Evans, T., & Zurayk, H. (2010). Health professionals for a new century: Transforming education to strengthen health systems in an interdependent world. The Lancet, 376(9756), 1923–1958.

Gebregzabher, E. H., Tesfaye, F., Cheneke, W., Negesso, A. E., & Kedida, G. (2023). Continuing professional development (CPD) training needs assessment for medical laboratory professionals in Ethiopia. Human Resources for Health, 21(1), 47.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage Learning.

Hansoti, B., Hahn, E., Dolive, C., Akridge, A., Atwell, M., Mishra, A., & Schleiff, M. (2021). Training global health leaders: A critical review of competency gaps. Annals of Global Health, 87(1), 65.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. Educational and Psychological Measurement, 66(3), 393–416.

Hosseini, S. E., Jafari, M., Nemati, A., Rahmani, K., Mahmoudian, P., & Ferdosi, M. (2023). Training needs assessment of hospital CEOs in a developing country: The example of Iran. BMC Medical Education, 23(1), 602.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Structural Equation Modeling, 6(1), 1–55.

Kang, K. (2017). The impact of training developing country ministerial officials at Korean universities: Evidence from the KOICA-KU Master's Degree Program. 글로벌교육연구, 9(3), 163.

Kang, K. (2019). Are the procedures of recruiting and selecting developing countries' officials for scholarship programs fair? The examination of the KOICA Master's degree program. KEDI Journal of Educational Policy, 16(2), 43.

Kang, K., Kim, J., & Kwak, D. W. Satisfying national and individual interests: Impact evaluation of the KOICA master's degree scholarship program. Journal of Contemporary East Asia Studies.

Kara DeCorby-Watson, G. M., Kim Bergeron, S. A., Abdi, S., Rempel, B., & Manson, H. (2018). Effectiveness of capacity building interventions relevant to public health practice: A systematic review. BMC Public Health, 18(1), 684.

Kim, S. Y., Kim, S. H., Kim, S. H., & Kim, S. Y. (2020). Analysis of the effects of a health policy capacity development education program as a public-private partnership model in official development assistance for health policy administrators. Osong Public Health and Research Perspectives, 11(5), 303–312.

Kline, R. B. (2016). Principles and practice of structural equation modeling (4th ed.). Guilford Press.

Korea Foundation for International Healthcare. (n.d.). Main programs of 5 agencies. Integrated Management System of Healthcare Invitational Training. https://imsohit.kofih.org/uat/uia/actionMain.do

Lanza, T. E., Paladini, A., Marziali, E., Gianfredi, V., Blandi, L., Signorelli, C., & Cadeddu, C. (2023). Training needs assessment of European frontline health care workers on vaccinology and vaccine acceptance: A systematic review. European Journal of Public Health, 33(4), 591–595.

Lawshe, H. C. (1975). A quantitative approach to content validity. Personnel Psychology, 28(4), 563–575.

Mangwanya, M. (2022). Evaluating the impacts of foreign aid on low-income countries in Sub-Saharan Africa. International Journal of Research in Business & Social Science, 11(6), 370.

Nakanjako, D., Kendall, D., Sewankambo, N. K., Razak, M. H., Oduor, B., Odero, T., & Farquhar, C. (2021). Building and sustaining effective partnerships for training the next generation of global health leaders. Annals of Global Health, 87(1), 66.

Navarro, S. M., Stewart, K., Tessier, K. M., Berhane, A., Alvarado, S. P., Tafirei, T., & Rickard, J. (2023). Medical students' perceptions of clinical and research training: An international needs assessment of 26 countries. International Journal of Translational Medical Research and Public Health, 7(2).

Njah, J., Hansoti, B., Adeyami, A., Bruce, K., O'Malley, G., Gugerty, M. K., & Berman, E. (2021). Measuring for success: Evaluating leadership training programs for sustainable impact. Annals of Global Health, 87(1).

Oh, M., & Yoon, Y. B. (2024). A new performance evaluation indicator for the LEE Jong-wook Fellowship Program of Korea Foundation for International Healthcare to better assess its long-term educational impacts: A Delphi study. Journal of Educational Evaluation for Health Professions, 21, 27.

Park, M., Choi, E., Jeong, M., Seo, H. J., Kim, J., & Seo, E. (2024). Interprofessional educational needs for shared governance of integrated care. International Journal of Integrated Care, 24(1), 15.

Phillips, J. J., Patti, R., & Ray, R. (2012). Measuring leadership development: Quantify your program's impact and ROI on organizational performance.

Rahimzadeh, M. (2019). Exploring learners' viewpoint regarding training program of health care providers based on CIPP model. Future of Medical Education Journal, 9(3), 3.

Rotem, A. (2010). A framework for evaluating the impact of the United Nations fellowship programmes. Human Resources for Health, 8.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. Methods of Psychological Research Online, 8(2), 23–74.

Tabachnick, B. G., & Fidell, L. S. (2019). Using multivariate statistics (7th ed.). Pearson.

Thompson, B. (2004). Exploratory and confirmatory factor analysis: Understanding concepts and applications. American Psychological Association.

Turner, M. W., Bogdewic, S., Agha, E., Blanchard, C., Sturke, R., Pettifor, A., & Ramaswamy, R. (2021). Learning needs assessment for multi-stakeholder implementation science training in LMIC settings: Findings and recommendations. Implementation Science Communications, 2(1), 134.

Vallejo, B. (2016). Capacity development evaluation: The challenge of the results agenda and measuring return on investment in the global South. World Development, 79, 1.

WHO. (2010). Evaluating training in WHO.

Williams, B., Onsman, A., & Brown, T. (2010). Exploratory factor analysis: A five-step guide for novices. Journal of Emergency Primary Health Care, 8(3), 1–13.

Yusoff, B. S. M. (2019). ABC of content validation and content validity index calculation. Education in Medicine Journal, 11(2), 49–54.

Zemmel, D. J., Kulik, P. K. G., Leider, J. P., & Power, L. E. (2022). Public health workforce development during and beyond the COVID-19 pandemic: Findings from a qualitative training needs assessment. Journal of Public Health Management and Practice, 28(5 Suppl 5), S263–S270.

김미영. (2024). 한국어판 R.I.G.H.T. leadership 측정 도구의 신뢰도와 타당도 검증 [Reliability and validity verification of the Korean version of the R.I.G.H.T. leadership measurement tool] (Publication Number 국내석사학위논문) [Master's thesis, Yonsei University Graduate School of Health Sciences]. [Wonju].

김슬기. (2020). ODA 초청연수사업의 성과지표와 현업적용도 측정 개선방향 연구 [A study on improving performance indicators and field applicability measurement of ODA invitational training projects] (Publication Number 국내석사학위논문) [Master's thesis, Pukyong National University Graduate School]. Busan.

사업평가실. (2010). 연수생 초청사업 효과성 평가 [Effectiveness evaluation of trainee invitation project]. 국제개발협력(2010-3), 216.

이은석. (2022). ODA 사업의 유형별 평가방법 연구: 기술협력 사업을 중심으로 [A study on evaluation methods by type of ODA project: Focusing on technology cooperation projects] (KIEP 연구보고서).

허윤선, 이주영, 유애라, 윤상철, & 이종욱. (2017). ODA 성과평가 개선방안과 정책과제 [Improvement plans and policy tasks for ODA performance evaluation]. 연구보고서, 2017(1), 1.

Appendix 1. About Dr LEE Jong-wook Fellowship Program

The Dr. LEE Jong-wook Fellowship Program (LJWFP), operating under the Korea Foundation for International Healthcare (KOFIH), represents a significant implementation of the Republic of Korea's commitment to global health diplomacy and development cooperation. As a public institution under the Republic of Korea's Ministry of Health and Welfare, KOFIH's implementation of this program exemplifies the strategic integration of healthcare capacity-building with international development objectives. This approach aligns with the current global health governance frameworks that emphasize sustainable capacity development rather than traditional aid models.

The program's reach of 1,672 healthcare professionals across 30 countries (as of December 2024) demonstrates its substantial contribution to the development of the global health workforce. This extensive geographical and numerical coverage suggests a systematic approach to address the critical shortage of healthcare professionals in developing regions, as identified by the World Health Organization's Global Strategy on Human Resources for Health (2016-2030). The program's implementation strategy reflects an understanding of the multifaceted nature of strengthening the healthcare system, as evidenced by its eight specialized tracks.

The LJWFP curriculum represents a comprehensive approach to healthcare capacity-building, addressing both the technical and systemic aspects of healthcare development.

1. Health Professional Education: This track addresses the crucial need for medical education expertise, recognizing that sustainable healthcare system development requires local capacity building in professional education. The focus on educating educators creates a multiplier effect, potentially impacting healthcare delivery systems beyond the direct program participants.

2. Health Policy and Management: By emphasizing governance capabilities, this component acknowledges the critical role of effective policy frameworks in healthcare system development. This aligns with the principles of global health governance, emphasizing the importance of strong institutional frameworks.

3. Clinical Experts: The focus on clinical competencies reflects the understanding that service delivery quality is fundamental to the effectiveness of the healthcare system. This track directly addresses the need for skilled practitioners in partner countries.

4. Biomedical Engineering: This technical training component recognizes the growing importance of medical technology in modern healthcare systems, addressing an often-overlooked aspect of healthcare infrastructure development.

5. Leadership: The incorporation of leadership training coupled with exposure to the Korean healthcare system represents a strategic approach to fostering long-term international partnerships and developing future healthcare leaders.

6. Health Financing and Economics: This track acknowledges the critical role of sustainable financing in healthcare system development and addresses one of the most challenging aspects of healthcare system sustainability in developing countries.

7. Infectious Disease Specialists: The inclusion of this track demonstrates foresight into addressing global health security concerns, which are particularly relevant in the post-pandemic context.

8. Graduate degree: The provision of advanced degree opportunities represents a long-term investment in partner countries' academic and research capabilities, contributing to sustainable healthcare system development.

The comprehensive framework of the LJWFP represents more than just a training initiative; it embodies a strategic approach to the global health workforce development. The program's design reflects an understanding of the interconnected nature of the healthcare system components and the need for capacity building at multiple levels. This holistic approach aligns with the current global health development paradigms that emphasize sustainable, system-wide capacity building over isolated interventions

LJWFP represents a strategic approach to global health capacity building, integrating education, research, and international cooperation. By training healthcare leaders and fostering cross-border partnerships, the program advances individual professional development and drives systemic improvements in healthcare systems and delivery worldwide. As the program continues to evolve, its long-term impact is expected to contribute to a more resilient, efficient, and inclusive global health system. LJWFP was designed as a capacity-building initiative to foster leadership among healthcare professionals from different countries. It serves as a global training platform, equipping participants with the necessary knowledge, skills, and networks to improve their healthcare systems. Furthermore, it functions as a strategic diplomatic effort by the Korean government to

enhance international recognition of Korea's healthcare system and position the country as a leader in global health.

The goal of LJWFP is to build a sustainable and healthy global community through the exchange of knowledge, expertise, and best practices in public health and medical care. The program emphasizes the following core objectives:

1. **Capacity Building for Global Health Leaders**

   - Fellows should be provided with advanced training in medical and healthcare management to enhance their leadership and technical competencies.

   - Encouraging innovative problem-solving and strategic planning to address healthcare challenges in their home countries.

2. **Strengthening International Healthcare Networks**

   - Establishing collaborative networks among Korean medical institutions, academia, and global health professionals is essential.

   - Facilitating international cooperation through research, policy discussions, and the sharing of best practices.

3. **Enhancing the Global Recognition of the Korean Healthcare System**

   - Showcasing Korea's medical advancements and policy frameworks as models for other nations.

- Increasing the competitiveness of Korea's healthcare sector by fostering international collaboration.

The participants can contribute to meeting the objectives in three key phases: pre-program preparation, on-site training, and post-program impact assessment. Each phase was designed to ensure that fellows actively engaged in learning, knowledge sharing, and implementing their acquired expertise in their respective countries.

1. **Pre-training Phase:** Identifying Health Issues and Preparing Solutions - Before commencing the fellowship, the selected participants were required to conduct a situational analysis of the health sector in their home countries or affiliated organizations. This involves:

   - Identifying key healthcare challenges affecting communities.

   - Conducting needs assessments to determine the gaps in medical services, infrastructure, and policies is essential.

   - Preparing a strategic plan to address these challenges using the knowledge and expertise gained during the fellowship.

2. **Training Phase:** Learning, Collaboration, and Action Plan Development: During the program, fellows participated in a series of lectures, workshops, field visits, and interactive discussions that covered essential aspects of healthcare system management and policy implementation. The program includes:

- Knowledge Sharing: Fellows present and analyze health system structures, policies, and challenges in their home countries.

- Expert Training: Korean healthcare professionals, academics, and policymakers provide training in healthcare management, digital health, universal health coverage, and other key areas.

- Collaborative Problem-Solving Fellows engage in discussions to develop evidence-based solutions tailored to their healthcare environment.

- Action Plan Development: Each fellow formulates a detailed action plan, integrating insights from the program into a strategic framework for implementation in their country.

3. **Post-training Phase:** Implementation, Knowledge Dissemination, and International Cooperation. After completing the fellowship, the participants were expected to apply their acquired knowledge and skills to their local health systems. The post-program activities included the following.

- Institutional Knowledge Sharing: Fellows disseminated their training outcomes within their organizations, health ministries, and broader healthcare communities.

- Implementation of the Action Plan: Participants took the initiative to implement their action plans in their respective institutions, aiming for measurable improvements in healthcare services.

- Engagement in the KOFIH Global Alumni Network (KGA): Graduates of the program become members of the KOFIH Global Alumni (KGA), an international platform for continued collaboration, research, and exchange of best practices.

- Development of new collaborative projects: Fellows are encouraged to propose collaborative projects (PCP – Project Concept Paper) involving ministries of health, national governments, and KOFIH country offices to formalize long-term partnerships and institutional cooperation.

The impact of the program extends beyond individual participants, fostering sustainable improvements in healthcare systems worldwide. This program contributed to the following:

1. **Strengthening the National Health System.**

   - Equipping healthcare leaders with policy-driven evidence-based approaches to reforming health services.

   - Encouraging interdisciplinary and cross-border collaboration to address emerging global health challenges is essential.

2. **Promoting Sustainable Development in Global Health**

   - This supports the United Nations' Sustainable Development Goals (SDGs), particularly Goal 3, which aims to ensure healthy lives and promote well-being in all ages.

- Enhancing healthcare access and equity in low- and middle-income countries through knowledge transfer and skill-building.

3. **Expanding Korea's Role in International Health Diplomacy**

- Establishing Korea as a key player in global health leadership by demonstrating its advanced healthcare system and policy expertise.

- Strengthening international partnerships through joint projects, training programs, and research collaboration.

As the program's goal is to support change agents from Low- and Middle-Income Countries (LMICs), who can drive policy improvements in the healthcare sector within their countries and beyond, it is necessary to ensure that participants align with this objective. The KOFIH utilizes a structured selection process.

KOFIH issues a recommendation request through diplomatic channels, inviting national governments to nominate suitable candidates. Only individuals officially recommended by their respective governments are eligible to apply for the program.

As part of the application process, candidates are required to state their Objectives and Goals related to the program in their application, and present and elaborate on their objectives during in-person or online interviews to assess their alignment with the program's mission. This rigorous selection process ensures that the program benefits individuals with the capacity and commitment to implement meaningful healthcare policy improvements in their home country.

The LJWFP was developed based on a comprehensive needs assessment conducted two years prior to training. This ensures alignment with the requirements of global health capacity-building. Partner countries recommend candidates through diplomatic channels, and the selection is based on interviews and review of application materials. A detailed curriculum was designed to address the training needs of the selected participants. Participants' workplace applications were assessed within a year of graduation as a post-training evaluation. The training impact is expected to be assessed two years post-training, but tools and indicators are yet to be developed.

Needs Assessment (N-2)

| Global Health Issue analysis | National Health Issue analysis of partner countries | Capacity Building Needs Assessment of partner countries |

Pre-training (N-1)

| Recommendation Request | Selection of Candidates | Training Needs Assessment of selected candidates | Curriculum Development |

Training (N)

| Understanding of Korean Health System and LEE Jong-wook spirit | Action plan development | Skills training | Team-work Development |

Post Training (N+1)

| Action plan application | KGA(KOFIH Global Alumni) Enrollment | Academic publications |

Training impact (N+2)

| Career Promotion | Spread out the Knowledge | Policy Application |

Process of LJWFP

Appendix 2. Survey Questionnaire – Tool Development

중-저소득국 보건정책 전문가(Health policy experts) 초청연수(Invitational training) 영향력 평가 지표

건강에 대한 권리는 최소한의 건강 기준과 건강 보호를 요구할 권리를 보장하는 필수적인 인권입니다. 이 권리는 세계 인권 선언, 경제적, 사회적, 문화적 권리에 관한 국제 규약, 장애인 권리 협약을 포함한 국제적 틀에 명시되어 있습니다.

공중 보건 시스템을 강화하는 것은 국민의 건강권을 보장하기 위해 매우 중요하며, 이러한 시스템을 관리하는 보건 정책 전문가의 교육 및 훈련도 마찬가지로 중요합니다.

세계보건기구(WHO)는 의료인력 교육 지침(2010)을 통해 Kirkpatrick 의 4 단계 모델에서 수익사업에서 사용하던 투자 수익률(ROI)을 교육에 투입되는 비용 대비 영향력을 측정하는 추가 단계로 제시하였습니다.

보건복지부 산하 한국국제보건의료재단은 2007 년부터 ODA 교육훈련 사업인 이종욱펠로우십을 통해 중저소득국 보건정책 전문가를 양성하고 있습니다. 본 사업의 효과를 종합적으로 평가하기 위해 교육 결과평가에 주로 사용되는 Kirkpatrick(1959)의 4 단계(반응, 학습, 행동, 결과) 및 교육의 과정 평가에 주로 사용되는 Stufflebeam(1960)의 CIPP(상황, 투입, 과정, 산출) 모델에 OECD DAC 의 여섯 가지 평가기준(일관성, 지속 가능성, 효율성, 관련성, 영향 및 효과성)을 더하여 1~4 단계 까지의 측정도구들을 개발하여 2022 년부터 본격 적용하고 있습니다.

하지만 5 단계인 영향력에 대한 평가는 사례조사에 그치고 계량화 가능한 측정도구가 개발되지는 못하였기 때문에 객관적인 분석 및 장기적인 추적조사에 한계가 있습니다. 본 연구를 통해 중-저소득국 보건정책 전문가의 연수 후 중장기 영향력에 해당하는 지표에 대해 정의하고, 이를 측정하기 위한 문항을 개발하는 것을 목표로 하고 있습니다.

선행 연구 및 유사기관의 사례를 검토, 초청연수 사후 조사 문항에 주로 사용되는 질문의

유형을 목표에 따라 '1) 직업적 성취도 2) 조직적 성취도 3) 학술적 기여도 4) 정책적 기여도 5) 전문 네트워크 확산' 등 5개 카테고리로 구분하여 총 20개의 본지표와 10개의 유사지표(variations or similar indicators)를 제시합니다.

※참고- 아래의 내용을 확인 후 번호에 표기하여 주십시오

1= 이지표는「보건정책전문가 초청연수의 중장기 영향력 측정 평가지표」로 보기에 적합성이 매우 낮은 수준임

3= 이지표는「보건정책전문가 초청연수의 중장기 영향력 측정 평가지표」로 보기에 산출하기에 적합성이 중간정도의 수준임

5= 이지표는「보건정책전문가가 초청연수의 중장기 영향력 측정 평가지표」로로 보기에 산출하기에 적합성이 매우 높은 수준임

아래의 항목은 선행연구 및 이종욱펠로우십 프로그램의 사례조사 및 인터뷰를 통해 확인된 중장기영향력에 해당하는 요소입니다. 중저소득국 보건정책 전문가 초청연수의 영향력을 평가하기 위한 목록으로 보는데 대한 적합성 정도를 해당번호에 표기(○)하여 주십시오.

직업적 성취도 관련 질문	

연번	세부지표	적합성 낮음 ←--------------→ 매우 적합함				
1	업무 자신감 증대	1	2	3	4	5
2	원하는 부서에 배치	1	2	3	4	5
3	동료와 상급자들의 인정	1	2	3	4	5
4	동일 조직 내 상위 직급으로 승진	1	2	3	4	5
5	중앙 부처 등 국내 상급 기관으로 이직	1	2	3	4	5
6	국제기구 등 해외기관으로 이직	1	2	3	4	5

조직적 성취도 관련 질문	
연번	세부지표	적합성 낮음 ←------------→ 매우 적합함

연번	세부지표	적합성 낮음 ←-----------→ 매우 적합함				
1	연수성과 조직 내 공유	1	2	3	4	5
2	액션플랜 조직 내 이행	1	2	3	4	5
3	액션플랜 성과 창출	1	2	3	4	5
4	액션플랜 성과 타 기관/ 지역 확산	1	2	3	4	5
5	소속기관의 대내외 인지도 상승	1	2	3	4	5
6	소속기관의 정부 지원사업 참여	1	2	3	4	5

학술적 기여도 관련 질문		

연번	세부지표	적합성 낮음 ←-----------→ 매우 적합함				
1	전파교육(ToT) 추진	1	2	3	4	5
2	학술논문 작성	1	2	3	4	5
3	상급 학교 진학	1	2	3	4	5
4	교육 커리큘럼 신설	1	2	3	4	5
5	관련 포럼 발표	1	2	3	4	5
6	관련 학과 신설	1	2	3	4	5

정책적 기여도 관련 질문		

연번	세부지표	적합성 낮음 ←-----------→ 매우 적합함				
1	조직 내 매뉴얼, 가이드라인 또는 프로그램 제안	1	2	3	4	5
2	조직 내 정책 개발 또는 개선에 참여	1	2	3	4	5
3	국가 정책매뉴얼, 가이드라인 또는 프로그램 작성	1	2	3	4	5
4	국가 정책 개발 또는 개선 제안	1	2	3	4	5
5	정부 정책사업 참여					

6	국제 협력사업 참여	1	2	3	4	5
전문 네트워크 확산						
연번	세부지표	적합성 낮음 ←------------→ 매우 적합함				
1	한국과 협력사업 참여	1	2	3	4	5
2	연수를 통해 형성된 네트워크와 공동 사업(연구) 추진	1	2	3	4	5
3	동문 정기모임 개최	1	2	3	4	5
4	동문 사회환원 활동 참여	1	2	3	4	5
5	전문가 그룹 형성	1	2	3	4	5
6	관련 학회 참여	1	2	3	4	5

Abstract in Korean (국문 초록)

제목:

중저소득국 보건 정책 전문가 초청연수의 중·장기 성과 측정을 위한 도구 개발 및 시범적용

배경:

국제 보건 연수 프로그램은 중저소득국가(LMICs) 보건의료 전문가의 역량 강화를 목표로 운영되고 있으나, 이들의 중·장기적 효과를 체계적으로 측정하기 위한 도구는 부족한 실정이다. 한국국제보건의료재단(KOFIH)이 운영하는 이종욱 펠로우십 프로그램(LJWFP)은 보건 정책 및 거버넌스 분야의 리더 양성을 목표로 하나, 이에 대한 장기적인 성과 평가가 제한적이었다.

목표:

본 연구는 LJWFP 의 중·장기 성과를 체계적으로 측정할 수 있는 평가 도구를 개발하고, 그 타당성과 신뢰성을 검증하는 것을 목적으로 하였다.

방법:

본 연구는 국제 보건정책 연수 프로그램의 중장기적 성과를 평가하기 위한 도구를 개발하고 타당화하기 위해 세 단계에 걸쳐 연구 방법론을 설계하였다. 첫째, 기존의 연수 평가 모델(예: Kirkpatrick, Kaufman, CIPP, OECD DAC)에 대한 광범위한 문헌 고찰과 프로그램 수료생 대상 인터뷰 및 자유서술형 응답에 대한 질적 분석을 통해 초기 평가 항목 풀을 도출하였다. 이후 글로벌 헬스 및 의학교육 분야 전문가 5 인을

대상으로 포커스 그룹 인터뷰를 실시하여 항목을 검토하고 개선 방향을 논의하였으며, 이어서 10 명의 전문가를 대상으로 2 차례 델파이 조사를 진행하였다. 델파이 조사에서는 5 점 척도 리커트 스케일을 사용하여 각 항목의 관련성, 명확성, 포괄성을 평가하였고, 내용타당도비(content validity ratio, CVR)와 전문가 합의 기준에 따라 최종 항목을 선정하였다. 이를 통해 경력 발전, 실행 계획 이행, 지식 공유, 학술 성과, 조직 기여, 국제 협력, 공동 프로젝트, 전문가 네트워킹 등 9 개 항목으로 구성된 설문지가 완성되었다.

자료 수집은 2023 년에 연수를 수료한 이종욱 펠로우십 프로그램(Dr. LEE Jong-wook Fellowship Program) 졸업생 52 명을 대상으로 진행된 파일럿 테스트를 통해 이루어졌다. 설문조사는 전자 방식으로 실시되었으며, 9 개의 폐쇄형 리커트 척도 문항과 함께 실행 경험 및 도전 과제에 대한 정성적 의견을 수집하기 위한 자유서술형 문항도 포함되었다. 양적 자료는 기술통계 분석, 탐색적 요인분석(EFA)을 통한 도구 구조 확인, Cronbach's α를 이용한 신뢰도 분석을 통해 검토되었다. 추가적으로 지역, 성별, 프로그램 유형에 따른 결과 비교를 위해 t-검정과 ANOVA 를 실시하였으며, 단기 및 장기 지표 간의 관계를 분석하기 위해 회귀분석 및 상관분석도 수행되었다. 또한 각 문항에 개방형 응답을 더하여 주제별 분석을 통해 구현 도전 과제, 맥락적 영향, 지속가능성 요구사항을 식별할 수 있도록 하였다.

결과:
시범적용 분석 결과 평가 도구는 커리어, 학술, 정책, 네트워크, 프로젝트 등 5 개 핵심 요인을 기반으로 높은 타당도와 신뢰도를 보였다. 참가자들은 직무역량 향상과 지식 확산 활동에서 높은 성과를 보였으며, 지역 간 성별간 차이에 대한 성과의 차이는

통계적 의미가 있을 정도는 아니었다. 다만, 참여 프로그램 간 성과의 차이는 있었으며, 석사과정 참가자의 경우 학술성과가 두드러졌다. 연수 직후 측정한 현업 적용도는 장기성과와 유의한 상관관계를 보였다. 또한 연수생 개인적인 성취가 국가 정책이나 커뮤니티에 미치는 영향에도 강한 긍정적인 관계가 있는 것으로 확인되었다. 한편, 본 설문 내에 개방형 질적 구성요소를 포함하여 도전 과제, 지속가능성 요인, 영향력 확산을 위한 개별 맥락 등을 파악하는 것을 시도하였다. 이러한 통찰은 향후 맞춤형 교육 및 후속지원 개선, 국제보건 연수 프로그램의 전략적 방향수립 등에 기여할 것으로 기대된다.

결론:

본 연구에서 개발된 평가 도구는 국제 보건 연수 프로그램의 중·장기 성과를 측정하는 데 있어 효과적인 수단이 될 수 있으며, 프로그램 개선 및 정책 결정에 실질적인 자료로 활용될 수 있다. 향후 지속적인 사후평가와 맞춤형 지원, 연속적 추적조사가 필요함을 강조한다.

_____