



**USAID**  
FROM THE AMERICAN PEOPLE

HEALTH CARE  
IMPROVEMENT  
PROJECT

## **Review of Selected Literature: HIV Training Evaluation & Performance Measurement**

Barton Burkhalter  
USAID Health Care Improvement Project  
University Research Co., LLC  
7200 Wisconsin Avenue  
Bethesda, MD 20814  
[www.hciproject.org](http://www.hciproject.org)

Prepared July 10, 2009  
Revised August 4, 2010

This literature review was prepared by University Research Co., LLC (URC) for review by the United States Agency for International Development (USAID), under funding from the U.S. President's Emergency Plan for AIDS Relief through USAID. It was authored by Barton Burkhalter of URC. The USAID Health Care Improvement Project (HCI) is made possible by the support of the American people and is managed by URC under Contract No. GHN-I-01-07-00003-00. This review is solely the responsibility of the author and does not reflect the official policy of USAID or of the United States Government.

## Table of Contents

---

	<u>Page</u>
<b>1. Introduction</b>	1
<b>2. Training Evaluation Literature</b>	1
Family Health International. 2007. <i>Evaluation of Uganda continuing medical education (CME).</i>	1
Felderman-Taylor J, et al. 2007. <i>A structured interview approach to evaluate HIV training for medical care providers.</i>	2
Flores R, et al. 2002. <i>Distance education with tutoring improves diarrhea case management in Guatemala.</i>	2
Garcia ML, et al. 2000. <i>Training of obstetric nurses and HIV testing in pregnancy at a managed care organization.</i>	3
Hamblin AC. 1974. <i>Evaluation and control of training.</i>	3
Katz K, et al. 2005. <i>Evaluating the effectiveness of decentralizing national reproductive health training and supervision.</i>	3
Kirkpatrick DL, et al. 2005. <i>Evaluating training programs: The four levels (3<sup>rd</sup> edition).</i>	4
Li Li, et al. 2007. <i>Diffusion among positive AIDS care messages among service providers in China.</i>	4
Murphy C, et al. 2006. <i>Supporting existing health cadres in learning new skills: Tools and approaches.</i>	5
Reynolds HW, et al. 2005. <i>Evaluation of a training program for onsite, in-charge supervisors in Kenya to improve reproductive health quality of care.</i>	5
Stein J, et al. 2008. <i>Building capacity for antiretroviral delivery in South Africa: A qualitative evaluation of the PALSA PLUS nurse training programme.</i>	5
Stiernborg M. 1996. <i>Impact evaluation of an international training course on HIV/AIDS.</i>	6
Weaver MR, et al. 2006. <i>Measuring the outcomes of a comprehensive HIV care course.</i>	6
<b>3. Measuring Practices Literature</b>	
Burkhalter BR. 1995. <i>Preliminary literature review: CME evaluation.</i>	7
Franco LM, et al. 2002. <i>Methods for assessing quality of provider performance in developing countries.</i>	7
Franco LM, et al. 1997. <i>Quality of case management of sexually transmitted diseases: Comparison of the methods for assessing the performance of providers.</i>	8
Hermida J, et al. 1999. <i>Comparative validity of three methods for assessment of the quality of primary health care.</i>	8
<b>4. Conclusions and Discussion</b>	9



## 1. Introduction

A limited literature review was undertaken to inform the HIV Training Evaluation Activity. Documents about HIV and other related training evaluations were sought using a standard search of several databases such as Google Scholar, USAID DEC, the University Research Co. database, as well as individual project staff knowledge of useful documents. Initially the search focused on evaluations of specific HIV-related training activities, and on more general assessments of training evaluation methodology. Training programs were defined broadly, including such components as on and off-site training, individual and whole-facility training, follow-up counseling, and job aids. Two summarizing reports (Hamblin; Kirkpatrick) led us to focus primarily on the effect of training on individual and organizational practices over time, rather than trainee reactions, attitudes, knowledge and skills, or on outcomes (because of the difficulty of causal attribution). Thirteen articles and reports with useful information are summarized below. Later we included four reviews on methodologies for measuring the quality of client-provider interaction when it became clear that our evaluations were likely to require assessment of client-provider interaction. This review served the purpose of providing insights into evaluation designs and also the special opportunities and challenges associated with HIV training evaluation.

## 2. Training Evaluation Literature

**Family Health International (Team: Wesson J, Katz K, Keyes E, Kirunda A, Kyakulaga J, Nyago R). 2007. "Evaluation of Uganda continuing medical education (CME), Final report." FHI, August 2007.**

Training program. This study evaluated continuing medical education (CME) workshops held by the Uganda MOH in six districts in 2005 and 2006. Each workshop was attended by health care providers, health care managers, and community representatives, and consisted of five presentations that provided the latest information on family planning trends in Uganda, the reintroduction of IUDs in Kenya, and WHO medical eligibility criteria, as well as introducing five job aids for the trainees to use in their work, and providing information about an addendum to the MOH reproductive health guidelines. The workshops also included group work and role play activity. The training program was slightly different in the 2005 workshops than in the 2006 ones.

Preliminary assessment. A preliminary assessment was done immediately following the completion of each workshop. Trainees completed a short knowledge test just before and just after the training, which showed significant increases in knowledge. Nearly 100% of the trainees rated the workshop as good or excellent. Short-term pre-to-post change in trainee practice was not measured. The preliminary assessment did not attempt to measure long-term change in knowledge or practice.

Evaluation methodology. In order to assess if the workshops had long-term effects on trainee knowledge and/or practices, a sample of trainees was interviewed 1-2 years after the workshops. The interviews collected information on trainee recall of key workshop messages, how the trainees have used the information obtained in the workshops such as dissemination of key messages to colleagues and changes resulting from the workshop, and the availability, current use, and perceived usefulness of the job aids introduced at the workshops. Two trained research assistants traveled to each study district, where a total of 53 interviews were carried out with 18 providers (17 nurses and midwives, one lab tech), 20 managers, and 15 community representatives.

Results of Long-term Evaluation. Nearly all 53 trainees interviewed said the workshop was very useful, and all said it was very or somewhat useful. The session on family planning trends was said to be very useful by all three types of trainees. Health care providers and managers found the IUD session very useful. On average each job aid has been used by about one-third of the providers and managers, all but one of whom reported the job aid had been very helpful. Nearly 80% of the providers and managers remember being trained about the job aids at the workshop, most of whom said the training had been very useful. About two-thirds of providers and managers felt very confident that the information contained in the job aids was correct. The report also presents findings related to individual job aids and on some specific information presented at the workshop. All those interviewed gave suggestions about how to improve the workshops.

Comment. The preliminary assessment measured immediate trainee reactions and knowledge gains; both were positive. This longer term evaluation used interviews to obtain trainee recollections about the workshops and opinions about the usefulness of the workshop experience. It did not measure sustainability of nor changes in

knowledge or practice. One interesting finding was that many providers and managers did not use the job aids because they did not receive copies of them, which the authors note seems to be easily correctable.

**Felderman-Taylor J, Valverde M. 2007. “A structured interview approach to evaluate HIV training for medical care providers.” *J Assoc Nurses in AIDS Care* 18(4):12-21.**

Training program. This study evaluated a training program for rural nurses and other non-doctors providing direct HIV care in rural New Mexico in low density HIV patient settings. The training used different modalities, ranging from on-site lectures to teleconferencing to three-day mini-sabbatical clinical experience in Denver.

Evaluation methodology. Of 112 providers trained, 27 provided direct HIV care, and 24 of these 27 agreed to participate in the study. Each of the 24 participants self-reported their answers to a 15-minute structured interview asked by an interviewer, covering eight topics:

1. Knowledge of medicine and side-effects
2. Provide sensitive, appropriate care
3. Willingness to treat HIV patients
4. Dealing with patient adherence
5. Co-infections
6. Screening frequency
7. Documentation and charting
8. Early detection and intervention

Results. Of the different modalities, participants liked the lunch meeting mode best, and the teleconferencing mode least. On the positive side, most participants said they gained knowledge (medication, co-infections) and became more willing to care for HIV patients as a result of the training. On the negative side, they said they are NOT likely to identify or treat HIV patients earlier, to chart or document better, or to implement risk-prevention strategies rather than treatment strategies. In fact, respondents said since the training they were more interested in treatment, less in prevention.

Comment. Based on their review of the HIV training literature, the authors conclude that most published HIV training evaluations have been at the 1<sup>st</sup> and 2<sup>nd</sup> Kirkpatrick levels, and a few at the 3<sup>rd</sup> level. They say that the results reported in the HIV training literature are not consistent: negative results claim that little or no changes in provider attitudes were observed, whereas positive results claim increases were observed in confidence to treat, in knowledge and in practice. This evaluation applied the 2<sup>nd</sup> Kirkpatrick level of rigor – change in trainee attitudes.

**Flores R, Robles J, Burkhalter BR. 2002. “Distance education with tutoring improves diarrhea case management in Guatemala.” *Intl J Quality HC* 14 S1:47-56.**

Training program. This in-service course provided information on diarrhea and cholera case management to doctors and nurses. After an initial one-day meeting, trainees returned to their home districts where they received ten written course modules in the mail, one each month for ten months. The trainees studied the received module, wrote responses to questions that tested their understanding, and mailed their work back to the course tutors. The experienced tutor-physicians reviewed the answers and mailed back written, individualized feedback to each trainee, along with the next module. Throughout the course, tutors monitored progress via phone conversations with the trainees, and also by visiting each trainee once or twice at their workplace. The tutors held a mid-course practice session at a local clinic for about half the trainees. A final graduation ceremony for all trainees was held. 1550 people started the course and 1381 completed (89%) it in four countries (El Salvador, Guatemala, Honduras, Nicaragua), and used 87 tutors, for an average of 15.9 trainees per tutor.

Evaluation methodology. The study used a pre-post panel, control and program group design applied in six districts of Guatemala. The program group consisted of 66 course graduates, and the control had 66 doctors and nurses who were interested in taking the course but worked in areas where the course was not offered. Trained physician observers were present in provider-patient encounters at the trainee workplace and observed whether the trainees complied with predetermined case management standards, using a pre-designed checklist. Pre-observations occurred during the two months before the course and post-observations during the two months after course completion. Audience bias was analyzed using multiple observations over one week of a sub-sample of trainees, and the validity of observed data compared patient exit interview data to the observed data.

Results. Correct assessment increased by 25% more in the program than in the control cases, although post-course assessment was still only about 60% correct in the program group. Rehydration treatment did not improve, and counseling only insignificantly. Exit interviews suggest that the observations of the counseling performance may not be valid. No audience bias was found. The program cost US\$60 per trainee.

Comment. This is a large program at the national level. Although the study measured change in practice relative to a control, the timing of the post measurement was not measured precisely, stretching out over a 2-month period after the training, and so the sustainability of the change was not estimated.

**Garcia ML, Grimes RM. 2000. “Training of obstetric nurses and HIV testing in pregnancy at a managed care organization.” *AIDS Care* 12(2):137-147.**

Training program. All 26 obstetric nurses employed by a private, multi-specialty group in Houston, Texas received didactic instruction in lectures, including information about the infection, its course with and without treatment, psychological and family adjustment issues, maternal-fetal transmission, patient rights, and Texas law, and was supplemented by role-playing of HIV counseling scenarios.

Evaluation methodology. The control group included 100 pregnant women counseled by nurses in their first two trimesters *before* the introduction of the training program; the program group included 100 pregnant women counseled by the same 26 nurses in their first two trimesters two months *after* introduction of the training program. The *nurse-counselors* completed a 33 item self-administered questionnaire that measured their knowledge at about the time they counseled the control women, and completed a similar questionnaire later at about the time they counseled the program mothers. Self-administered survey questionnaires were distributed to 300-400 *women* counseled before the training program, and the first 100 who responded (after removing 3<sup>rd</sup> trimester women) were included in the control group, and a similar procedure was used to select 100 women in the program group who had been counseled 2 months after the training program was introduced. The survey asked mothers about HIV/AIDS knowledge, about testing for HIV infection (including what they were advised by the nurses), and social-demographic information.

Results. More program women said they were advised by the nurse-counselors to get tested for HIV than control women, but program women did not have more HIV knowledge nor did program nurses give more information about HIV to their mothers than the controls. However the program mothers did get tested more frequently when the nurses told them to, even though they did not know more about HIV. Thus, although program nurses increased their knowledge about HIV as a result of the training, they were not effective in passing this knowledge on to the mothers.

Comment. The training program resulted in better outcomes (more pregnant women tested for HIV) because the nurses advised them to do so, and not because the women gained more knowledge about HIV. This evaluation used the 4th Kirkpatrick level of rigor – change in final results (more pregnant women tested for HIV), although the long term impact of the training on this final result was not measured (Did the program nurse-counselors continue to tell women to be tested for months or years after the training?)

**Hamblin AC. 1974. “Evaluation and control of training.” London: McGraw-Hill.**

Training model. This book extends the four-level model of Kirkpatrick to five levels of training effects, linked by a cause and effect chain, in which:

- Training ..... leads to (1) Reaction
- which leads to (2) Learning (in attitudes, knowledge, skills)
- which leads to (3) Changes in trainee practices
- which leads to (4) Changes in organization
- which leads to (5) Changes in achievement of ultimate goals.

Comment. The five-step, causal model posited by Hamblin is an interesting variation of the Kirkpatrick training evaluation model. Step 4 in the Hamblin model (“changes in organization”) is an important addition that applies in some but not all situations. The one-way causal linkage of the Hamblin model is not necessarily true and should be investigated before it is accepted in any evaluation.

**Katz K, Toroitich-Roto C, Cuthbertson C, Family Health International. 2005. “Evaluating the effectiveness of decentralizing national reproductive health training and supervision.” Family Health International.**

Training program. AMKENI is the national decentralized Kenya system for training and supervising reproductive health providers. It has several integrated components, including training, supervision, and job descriptions. It reaches staff working primarily in facilities and staff providing health services outside facilities.

Evaluation methodology. The evaluation compared 17 program health facilities where AMKENI was operating to 17 matched control facilities without AMKENI, located in several Kenyan districts. Data were obtained by 34 data collectors (one in each facility) from 34 facility audits, 50 supervisor interviews, 107 provider interviews, observation of 151 client-provider interactions, and 151 client exit interviews. Types of data collected included characteristics of the study participants, performance management by providers and supervisors, and quality of care at the facility.

Results. The program and control groups were similar in many ways, but despite being matched were not the same in some ways: clients were similar in both groups (age, education, marital status), but program facilities had more clients than the controls; most supervisors in both groups were nurses or midwives (not doctors), but program facilities had more trained staff with more experience (especially supervisors) than the controls. The success of the AMKENI program was judged by several performance indicators. The program facilities scored higher on three indicators: program supervisors gave more feedback to providers, client satisfaction was higher (90% satisfied in program vs 67% in controls), and more program facility-focused supervisors had job descriptions than facility-focused controls, although the frequency of job descriptions varied widely across facilities and did not differ for non-facility-focused supervisors. Program facilities scored worse on one indicator: waiting time was longer. No significant difference between program and control facilities was found in numerous indicators: number of supervisory visits, quality of facility infrastructure, infection prevention (wash hands, use gloves) which was low in both groups, client-provider interaction (good in both groups), provider job satisfaction (low in both groups), and providers in both groups thought their skills were inadequate.

Comment. The authors note that the results may be suspect because of the differences between the program and control groups, especially the difference in supervisor characteristics. This evaluation attempted to measure trainee practices (Kirkpatrick level 3) and system performance (Kirkpatrick level 4), but did not measure or control for other potential confounding factors that might have influenced the results, beyond the use of the “matching” methodology.

**Kirkpatrick DL, Kirkpatrick JD. 2005. “Evaluating training programs: The four levels (3<sup>rd</sup> edition).” San Francisco: Berrett-Koehler Publishers, Inc.)**

This book gives a conceptual framework and analyzes methodologies for evaluating training programs. The four levels referred to in the title and typical methodologies used to measure them are:

<u>Levels of evaluative rigor</u>	<u>Typical methodologies</u>
1. Reaction of trainees	1. Participant satisfaction survey at end of training
2. Changes in attitudes, knowledge, skills of trainees	2. Interviews or observations to assess learning, etc.
3. Changes in trainee behavior resulting from training	3. Interviews or observations over time
4. Final results (outcome changes) resulting from training	4. (a) interview patients or (b) chart reviews

The book notes that valid measurements become increasingly difficult to obtain as the level increases from 1 to 4, but that higher level evaluations (e.g., 3 and 4) should be accompanied by lower level measurements (e.g., 1 and 2). A wealth of examples and sample data collection forms are provided, mostly from U.S. business training programs.

**Li Li, Haijun Cao, Zunyou Wu, Sheng Wu, Lin Xiao. 2007. “Diffusion among positive AIDS care messages among service providers in China.” *AIDS Ed & Prev* 19(6):511-518.**

Training program. China has a broad program to train health providers about HIV. This study was not an evaluation of the Chinese HIV training programs, but rather a study of whether providers who took an HIV training program were more likely to diffuse positive messages about HIV than providers who had not taken one of the HIV training programs. Previous research found that popular opinion leaders with certain characteristics (gender, ethnicity, medical education, contact with PLWHA, HIV training) were more likely than others to successfully diffuse positive messages about HIV to co-workers and friends. Thus, one result of HIV training may be that it causes more diffusion of HIV messages to co-workers and others. This studied addressed this possibility.

Evaluation methodology. Survey questionnaires with 172 items were sent to 1,101 providers (doctors, nurses, lab techs) at 89 randomly selected health facilities, stratified by region and care level. The questionnaires contained self-reported questions on: (1) diffusion of positive messages, (2) discriminatory attitude, (3) knowledge of HIV/AIDS, (4) demographic and training information. About diffusion it asked whether they discussed the following with co-workers: How to protect themselves from infection? If HIV/AIDS patients deserve the same quality of care as other patients? If AIDS/AIDS patients should receive lesser care? Importance of confidentiality.

In addition, the questionnaire obtained information about the respondent's gender, ethnicity, medical education, contact with PLWHA, and HIV training.

**Results.** HIV training had by far the highest correlation with diffusion in the one-variable-at-a-time analysis:  $F=43.9$  for HIV training vs  $16.8$  for medical education (the 2<sup>nd</sup> highest correlation), and also in the multi-variate analysis:  $t=5.56$  for HIV training vs  $2.62$  for contact with PLWHA (the 2<sup>nd</sup> highest correlation).

**Murphy C, Millonzi K. 2006. "Supporting existing health cadres in learning new skills: Tools and approaches." Published by The Capacity Project, IntraHealth International, Inc. for USAID: Chapel Hill, NC.**

This report identifies training programs and resources related to HIV/AIDS, including:

- **Short courses and curricula.** Appendix 1 lists 11 organizations in Africa that provide HIV/AIDS related short courses. Appendix 2 lists six providers of training materials and their websites (such as WHO and FHI). Appendix 3 lists nine training resource databases and their websites (such as I-Tech and AETC).
- **Guides for course design and curriculum development.** Appendix 4 lists six organizations providing such guides, and their websites (such as Jhpiego, IPAS, and IntraHealth).
- **Alternative training approaches.** This includes: structured on-the-job (OTJ) Training (such as the IntraHealth Hareg project in Ethiopia focusing on VCT and PMTCT, the Jhpiego OTJ work in Kenya and Zimbabwe, and the clinical attachments mentor approach in Malawi and Rwanda); E-learning for knowledge transfer using email; and blended learning in which the knowledge transfer program consists of several "blended" components such as group sessions, self-study, learning partners, distance learning, mentoring, clinical practica, OJT, job aids, facilitator consults, problem-based learning, action planning. Appendix 5 lists 11 sources of alternative training approaches and their websites.
- **Tools for performance support.** Includes job aids, help lines, and websites.

**Reynolds HW, Toroitich C. 2005. "Evaluation of a training program for onsite, in-charge supervisors in Kenya to improve reproductive health quality of care. Final report." Family Health International.**

**Training program.** The training program was designed and implemented by Jhpiego for Kenya nurse-supervisors in MCH and family planning in 2002. It consisted of a one-week training, with follow-up mailings, work site visits, and a second two-day meeting about a year later. However, the report does not describe the training program in very much detail.

**Evaluation methodology.** The evaluation methodology used a true experimental design with pre and post measures of program and control groups. 60 health facilities were paired on several variables and then for each matched pair, one facility was randomly assigned to the program and one to the control. Supervisors, providers and clients were interviewed pre in April 2002, and post a year later. Not all individuals were interviewed pre and post because of personnel changeover. Training program cost data were also obtained.

**Results.** The report discussed findings related to the interview topics, including use of motivation-communication techniques by the supervisors, supervisor problem identification, facility amenities and equipment availability, stock-outs, job satisfaction, supervisor observation of provider-client interaction, and performance quality (primarily hand-washing, instructions about dosage, etc.), waiting time, and client satisfaction. The study reported that the program group was better than the control in hand-washing, client communication, and confidentiality, but not in facility improvement or client satisfaction. The report said it did not address outcomes such as contraceptive use, or quality of care directly related to desirable outcomes. Cost was estimated to be \$2,113 per supervisor trained.

**Comment.** This evaluation measured changes in trainee practices (Kirkpatrick level 3) and system performance (Kirkpatrick level 4) a year following the program start. However, the interview data did not investigate many important performance indicators, and the validity of the interview data was not mentioned in the report. The large sample size and random assignment of facilities to program and control is a reasonable control of potentially confounding factors that might influence the results.

**Stein J, Lewin S, Fairall L, Mayers P, English R, Bheekie A, Bateman E, Zwarenstein M. 2008. "Building capacity for antiretroviral delivery in South Africa: A qualitative evaluation of the**

**PALSA PLUS nurse training programme. *BMC Health Services Research* 8:240. (Published online Nov 18, 2008 doi: 10.1186/1472 6963 8 240.)**

Training program. The PALSA PLUS program supports primary health care nurses (PHC nurses) in the management of lung diseases and HIV/AIDS, including ART, in South Africa. It provides both guidelines and training for the clinical management of lung diseases and HIV/AIDS. The point of the training is to facilitate guideline use. The traditional provincial training program for PHC nurse was off-site, lasted one week, and used a didactic, information dissemination approach. The new on-site training was conducted in weekly two-hour sessions over three-four months, based on evidence that the alternation of learning and practice is best for application of new knowledge, and included all PHC nurses at the clinic, not just those providing lung diseases and HIV/AIDS care.

Evaluation methodology. The first 15 clinics chosen to receive the PALSA PLUS program in the Free State (of South Africa) were randomly assigned to the control (seven clinics) or program (eight clinics). The control received the PALSA PLUS guidelines and traditional 1-week off-site provincial training course, while the program received the guidelines and one-week off-site traditional provincial training course *plus* the additional new experimental on-site PALSA PLUS training. The qualitative evaluation collected subjective data via 14 semi-structured interviews with doctors and nurses, three nurse focus groups, and unstructured participant observation of the trainings themselves.

Results. Authors report the uptake of the PALSA PLUS program by nurses was very high, both for lung health and ART-specific components. The program nurses perceived several advantages of the new program: better on-going support and thereby learning, both emotional and supervisory; encourages better managerial oversight of system-level changes that facilitate individual efforts, and better integration of AIDS care in the clinic due to incorporation of all nurses in the program, not just ART nurses.

Comment. This study was limited to nurse perceptions, and did not attempt to observe practices, or estimate change in practice.

**Stiernborg M. 1996. “Impact evaluation of an international training course on HIV/AIDS.” *AIDS Care* 8(3):311-319.**

Training program. WHO international course on HIV/AIDS conducted at U. New South Wales, Sydney for physicians and nurses from WHO East Med, SE Asia, and West Pacific regions. In 1988-1990 the University conducted four six-week classroom training courses that included testing, ethical, legal and psychosocial as well as health and medical aspects of the disease for 45 physicians and 39 nurses from 28 countries. Course objectives were to enable trainees to diagnose, case manage, and counsel HIV patients, and to assume leadership as resource persons and trainers back home.

Evaluation methodology. Citing the Kirkpatrick and Hamblin evaluation frameworks, this study evaluated level 3 (change in trainee practices). It mailed a pre-tested questionnaire to the 84 trainees 2.5-4.5 years after completing the course, with reminders to respond at four and seven months. Response rate was 72.6% (61 of 84). The self-administered questionnaire contained both open and close-ended items about the trainee’s own level of knowledge, skills and attitudes, about caring for AIDS patients, and fear of contracting HIV at work, at the time of the course and when they received the questionnaire.

Results. Most respondents seemed to have been appropriate selections for the training. They reported: high commitment to HIV/AIDS care, duties related to AIDS care (physicians 94%, nurses 79%), had cared for at least one AID patient since returning home (physicians 75%, nurses 48%), most had organized HIV/AIDS in-service training (physicians 75%, nurses 55%) and/or had taught HIV/AIDS courses organized by others, and were involved in other HIV/AIDS activities such as mass media information dissemination, conference presentations, national/international policy or program.

**Weaver MR, Nakitto C, Schneider G, Kamya MR, Kambugu A, Lukwago R, Bpharm, Ronald A, McAdam K, Sande MA. 2006. “Measuring the outcomes of a comprehensive HIV care course.” *Acquir Immune Defic Syndr* 43(3):293-303.**

Training program. Since 2004, Makerere University Medical School in Uganda has operated a four-week in-service course for doctors in the management of HIV, with half the time devoted to classroom sessions and half to clinical sessions. Its aim is to enhance clinical leadership, by increasing clinical skills and fostering increased communication of HIV/AIDS information in their home hospital.

Evaluation methodology. The study used a pre-post panel of trainees with no control. The panel included 46 doctors who completed the training program in 2004 and 2005, all who worked in Uganda. Four measurements of trainee performance were made: (1) observed performance in six patient exams at the beginning of the program, (2) observed performance in six patient exams at the end of the four-week program, (3) a telephone interview one month after the end of the program, and (4) a 1½ day follow-up session three-four months after the end of the program, at which another patient exam was observed and trainees were interviewed. The observed performance in patient exams was done by a trained physician using a checklist. However, actual measurements were obtained on less than all 46 study trainees -- both pre and post clinical tests were obtained for 32 trainees, telephone interviews were held with 45 of the 46 trainees, and follow-up session information was obtained for only 14 trainees.

Results. 17 indicators of performance were defined and observed during the patient exams. All 17 had higher average immediate post-scores than pre-scores, and 11 of these were significantly higher ( $p < .03$ ). Results were less clear-cut for performance at the patient exam in the follow-up session three-four months after the end of the program. While 14 of the 17 indicators had higher average scores for the follow-up exam than the immediate post exam, only three of these were significant. The results of the telephone interviews provided a call to action regarding some issues; for example, during their last HIV patient encounter 53% of the trainees did not know if the patient's weight had changed since the previous encounter, most had not asked the patient, and some did not know if the patient's appetite had changed. However, there was no data against which the telephone data could be compared.

Comment. The small follow-up sample makes the results less convincing. The article has a very useful list of references.

### 3. Measuring Practices Literature

#### **Burkhalter BR. 1995. "Preliminary literature review: CME evaluation." (for BASICS Project).**

This preliminary review provides information on studies published in the 1980s and before, mostly U.S. based. Four comprehensive reviews of the CME evaluation literature are summarized (Bertram and Brooks-Bertram, 1977, reviews 65 studies; Lloyd and Abrahamson, 1979, reviews 47 studies; Haynes et al., 1984, reviews 248 studies; Raymond, 1986, meta-analysis of 58 studies). These four reviews found very few "rigorous" studies, and contradictory results, some studies reporting positive impacts of CME on provider performance and others negative. The Raymond meta-analysis found that a CME's duration was not related to its impact on performance. This report's review of nine more recent articles concluded that: (1) few rigorous studies, (2) some knowledge but little practice improvement, except for prescription practice, (4) inconsistent findings that are not understood or analyzed.

The review identifies eleven different techniques that have been used to measure physician behavior: (1) record reviews, (2) direct observation, (3) audio and camera "observation," (4) face-to-face physician after interviews, (5) telephone physician after interview, (6) written after questionnaire to physician, (7) face-to-face patient exit interview, (8) telephone patient after interview, (9) simulated trained actor-patients who report in writing, (10) standardized patients, and (11) structured skill exam of physician. Each has strengths and weaknesses with respect to reliability and bias (validity). Observer bias (also termed observer effect, audience effect, and Hawthorn effect) is one important potential issue. Nine reviewed studies do not agree, with some finding audience bias and others not. Three studies of self-reporting bias begin to unravel this phenomenon, investigating factors causing this bias (e.g., memory loss, motivation to appear better), variation by type of performance, self-reporting before the fact.

#### **Franco LM, Franco C, Kumwenda N, Nkhoma W. 2002. Methods for assessing quality of provider performance in developing countries. *Intl J Quality HC* 14,S1:17-24.**

This study compared four methods of assessing the provider performance – direct observation, patient exit interviews, record review, and provider interviews. Data were obtained on 436 provider-patient (children under five years) encounters for 30 providers in 14 health facilities in Malawi. 52 different performance indicators were measured, spread among four functions (history taking, physical exam, diagnosis and treatment, and counseling) for cough, diarrhea and fever. Direct observation by trained nurses with research experience was considered the standard against which the other data collection methods were judged, using the Kappa statistic. The study also looked at the cost of each method.

**Results: *Data validity.*** Patient exit interviews generally provide the best agreement with observation data, while record reviews and provider interviews are worse. Nevertheless, each method was strong for some indicators and weak for others. For example, patient exit interviews provided reliable information on many history-taking tasks, easily discernable physical exam tasks, and many counseling tasks, but less reliable data on certain assessment tasks. (The article presents statistical results of the agreement between observation and exit interviews for 43 different performance indicators.) The authors note that although inter-rater reliability of observers was high, the observer scores may not be a reliable gold standard, partly because the presence of observers may influence the performance of the providers and because some of the observer data in the study was aggregated over all patients receiving care from the provider.

**Results: *Cost.*** The cost of each method has two components – time to obtain the data for a single case once the provider-patient encounter has started, plus the time waiting for the next case. Direct observation of a patient-patient encounter averaged 3-5 minutes; each patient exit interview averaged 5-6 minutes; provider interviews averaged 30-35 minutes but included multiple cases and situations and thus are not comparable to the time per case for direct observation and exit interviews. (Record review time per case was not provided.) The second cost component (time waiting for the next case) depends on the density of the patient load for the direct observation and exit interview methods, but has no effect on the provider interview or record review methods. Thus, substantial additional costs are involved for the direct observation and exit interview methods in low patient load situations. Severe cases are usually very low frequency and therefore the total cost to assess performance in severe cases is much less for record review and provider interview than for direct observation or exit interviews.

**Recommendation:** The authors conclude that best data collection strategy depends on the purpose and conditions of the study. To make improvements, triangulation of data collection methods is probably best. If the purpose is to support the supervisory process, then limited observation combined with provider interviews is probably best. If the purpose is to spot a faulty provider for sanction, then exit interviews possibly combined with a mystery patient might be best. Provider interviews can substitute for observation when an assessment is trying to understand the decision processes of a provider or when direct observation is difficult or impossible (e.g., for ethical reasons). Record reviews for some information may be useful for assessing low frequency severe cases.

**Franco LM, Daly CC, Chilongozi D, Dallabetta G. 1997. “Quality of case management of sexually transmitted diseases: comparison of the methods for assessing the performance of providers. *WHO Bull* 75(6):523-532.**

Three different methods for assessing the quality of STD case management in Malawi were investigated: direct observation of patient-provider encounters, provider interviews, and simulated patients. Provider performance was observed in 150 patient-provider encounters with 54 different providers in 39 health facilities; 103 different providers were interviewed, including 49 of the 54 observed; 20 providers received a visit from a simulated patient.

**Results.** Simulated patient data are probably the best measurement of normal performance, but may require multiple simulated patients because the multiple observation data indicated that a provider’s behavior is not consistent across patients. Direct observation data are probably best for some tasks. Data from provider interviews are of concern because they may reflect provider knowledge rather than practice. Agreement was poor between direct observation and provider interview, and between direct observation and simulated patient data, but often high between direct observation and provider interview. Providers perform better when they know they are being observed, even after 3 or more days of being observed. When interviewed, providers said they did more than they actually did. Performance as measured by direct observation is similar or slightly higher than performance assessed by simulated patients. Although assessments by simulated patients are probably the most accurate for normal tasks, this method is very resource intensive.

**Hermida J, Nicholas DD, Blumenfeld SN. 1999. “Comparative validity of three methods for assessment of the quality of primary health care. *Intl J Quality HC* 11(5):429-433.**

This study compared the sensitivity and specificity of three methods for assessing health worker performance in three Guatemala health facilities. The performance of doctors, nurses, and auxiliary staff during patient encounters for respiratory infections, acute diarrhea, and family planning was assessed by (1) a checklist-based non-expert observation of the consultation, (2) exit interview with the mother following the consultation, and (3) review of the patient’s clinical record. The assessed performance of seven key tasks was then compared to an expert “gold standard” observer who used the checklist to assess the same seven tasks during each encounter. The sensitivity (percentage of cases actually performed incorrectly that were assessed as incorrect) and specificity

(percentage of cases actually performed correctly that were assessed as correct) were calculated for each task for each of the three methods.

**Results.** The article concludes that sensitivity and specificity was highest for the non-expert observer and the mother interview methods, and lowest for the record review method. A table reports sensitivity and specificity values for the three methods for eight performance indicators (ask if blood in feces, evaluate mouth dryness, count respiratory rate, prescribe ORS, prescribe antibiotic, and advise mother when to continue breastfeeding, about pneumonia danger signs, and about family planning).

**Recommendations.** The authors recommend using the mother interview method because direct observation may influence provider behavior during an encounter. They also note that the difference between the expert observer and non-expert observer assessments may reflect inter-rater reliability rather than incorrect assessments by the non-expert observers, which is a second reason for using mother interviews.

#### 4. Conclusions and Discussion

1. This evaluation activity should address the impact of training on trainee and organizational practices over time, not the impact on trainee reactions, attitudes, knowledge, or skills, or the impact on outcomes. Many training assessments measure, and find, immediate impact of training on trainee reactions, attitudes, knowledge and skills, but most of these studies do not measure impact on practices. Fewer assess the impact of training on trainee and organizational practices, especially over a longer period of time. The few that measure both reaction-knowledge-attitudes-skills and practice often do not find that impact on reaction-knowledge-attitude-skills do not lead to a positive impact on practice. Even fewer attempt to measure impact on outcomes, especially with careful attention to causal attribution. Since the primary focus of this study is the impact of training on practices over time, this review of selected literature includes mostly reports of evaluations that assess practice, not reports assessing only immediate reactions, attitudes, knowledge or skills.

Some situations enable evaluations to assess the impact of the training on trainee practices and on outcomes, if outcomes are relatively easily measurable and closely linked to trainee practices. The Garcia and Grimes (2000) study is such an evaluation. However, other evaluations that attempt to measure outcomes encounter difficulties that extend the duration (and increase the cost) of the evaluation to many months or years. Most evaluations do not attempt to measure outcomes, or do not control adequately for alternative causes of outcome change.

The Kirkpatrick four-level model of training evaluation [Reaction of trainees; Changes in trainee attitudes, knowledge and skills; Changes in trainee behavior resulting from the training; Final results (outcomes) resulting from the training, adjusted to include Hamblin's fifth level (Changes in organizational practices)], provides an appropriate conceptual framework for the present evaluation.

2. Literature provides evidence that multi-component training programs can achieve better results that are longer-lasting than single-component programs, and therefore should be the primary focus of this evaluation program. Several evaluations of in-service training programs with an on-the-job focus (Stein et al. 2008, Weaver et al. 2006) and which incorporated components in addition to classroom training such as job aids and follow-up communications, consultations and meetings, concluded that multi-component programs had positive results (Family Health International 2007, Felderrman-Taylor 2007, Flores et al. 2002, Katz et al. 2005, Reynolds and Toroitich 2005). Several opinioned that the multi-component nature of the training probably led to longer-lasting improvements. This literature thus provided support for the decision to include multi-component training programs in the HCI HIV training evaluation activity.

3. Evidence from the literature indicates that each training program may improve some practices but not others, suggesting the need for more fine-tuned training program designs and evaluations. Several studies and reviews reported that specific training programs improved some provider practices but not others

(Flores et al. 2002, Franco et al. 2002, Reynolds and Toroitich 2005). This finding suggests that training programs aimed at improving practices must correctly match training program substance and method to specific practices and conditions. (However, we did not actually encounter studies in our limited literature review that did this.)

4. Evaluations should measure the extent to which the training program was fully implemented? One study noted the importance of measuring the degree to which the program was actually implemented (Family Health International, (2007)). This evaluation program should incorporate this clearly sensible information into the evaluations carried out by the HCI HIV training evaluation activity.

5. The validity and cost of measuring provider practices is difficult, especially the quality of client-provider interaction, indicating that this evaluation needs to take special care in defining and measuring practices. A wide variety of methods exist for measuring the quality of client-provider interaction, including among others: direct observation, visual and audio taped recording, standardized patients, client interviews, provider interviews, and record review. Validity and cost differ substantially for the different methods used to measure provider practices, each with its strengths and weaknesses. For example, direct observation is both expensive and may result in significant observer bias. Furthermore, some providers and some types of client conditions may have highly variable validity from provider-to-provider client-to-client (Burkhalter 1995, Franco et al. 1997 and 2002, Hermida et al. 2002).