International Journal of Research Studies in Education
2018 Volume 7 Number 2, 41-51

Measuring training participants' changing performance using self-reporting methods and their implication in the Grameen Bank training evaluation

Rouf, Kazi Abdur ✉
York Center for Asian Research (YCAR), York University
Visiting Scholar, Faculty of Environment, York University
Noble International University (MIU), Canada (Kaziabdur56@hotmail.com)

ISSN: 2243-7703
Online ISSN: 2243-7711

OPEN ACCESS

Received: 5 February 2017 **Revised**: 14 April 2017 **Accepted**: 15 April 2017
Available Online: 17 April 2017 **DOI**: 10.5861/ijrse.2017.1756

Abstract

Transfer of learning through training contributes to clients, managers, and organizations' success in achieving organizational goals. Therefore, learning gained evaluation is very important to measures transfer of learning and skills development of employees. However, to avoid training evaluation costs, usually institutions are not interested to evaluate its training performance by outside evaluators and consultants. Hence, less expensive self-reporting training evaluation methods are popular to evaluate training participants' changing performance and their contributions in achieving institutional goals. This paper reviews self-reporting different training evaluation methods and their suitability of using Grameen Bank's training evaluation in order to measure learning achievement and skills development of the Grameen Bank's officials. The paper also discusses prospective and retrospective self-reporting methods and compares them with other training measurement methods to find an appropriate training evaluation method for Grameen Bank in Bangladesh.

Keywords: Grameen Bank; response bias; self-assessment; self-reporting; and training evaluation

Measuring training participants' changing performance using self-reporting methods and their implication in the Grameen Bank training evaluation

1. Introduction

Self-reporting is a widely-used form of training evaluation, because of its simplicity, low–cost and convenience. It is one of the main approaches that can be used to measure training outcomes, especially learning gains. Mezzof (1981) reported that at least 80% of evaluation programs employ self-reporting as the measure of choice due to its ease and low-cost. Although self-assessment training evaluations can be measured in several ways with several benefits, self-assessment still has response-shift-bias that contaminates training evaluation test scores (Mezzof, 1981; Bray & Howard, 1980; Stufflebeam & Wingate, 2005; Lam & Bengo, 2003; Taylor & Taylor 2009; and Hill & Betz, 2005). Moreover, other questions arise in the use of self-reporting methods such as: retrospective design problems in pre- and post-test design, validity concerns, subject response style effect and recall bias. Another question centers on whether the measurement of all of Kirkpatrick's four levels of training outcomes is necessary to accurately understand learning outcomes at the end of a training course. Therefore, the paper looks at the issues arising from using self-assessment in measuring Kirkpatrick's four levels of training outcomes in practice; discuss different methods of self-assessment and as an example, the suitability of using Grameen Bank's training evaluation.

1.1 Implications of learning through trainings

Transfer of learning through training contributes to everyone's success: clients, managers, and organizations. Therefore, training learning gained in evaluations is very important, because training measures learning and skills developed through training. Also, it measures the effectiveness of training to determine how much the trainees have learned. However, if training evaluation results are unreliable or invalid because of response-shift bias, or other errors, one cannot say how or if a program achieved its desired outcomes. Hence, evaluators are serious about finding appropriate training evaluation methods and are placing an emphasis on the validity of retrospective tests (Hill & Betz, 2005, p. 501) to minimize response-shift-bias and other errors. According to Bray and Howard (1980) response-shift bias, a methodological source for contamination that confuses results of studies including self-report measures are looked at by training evaluation studies (p. 62). In this regard, the Pre-Then-Post-Tests (retrospective test) can improve training accuracy and can legitimately document the benefits of training. In retrospective tests, participants are asked to take a test of their skills and knowledge in order to judge their knowledge and ability in specific areas of evaluation prior to and following a certain (weeks/months) instructional period. This retrospective method helps training participants to reduce the responses that engender response-shift bias as compared to other evaluation methods like control groups and experimental groups, prospective tests, and many others.

1.2 Objectives of the paper

There are several models and methods that exist in training evaluation. This paper is not arguing different training evaluation models; rather it discusses prospective, retrospective self-reporting methods and compares them with other training measurement methods to find an appropriate training evaluation method for Grameen Bank.

1.3 Issues of self-reporting

Some common problems have been found with respect to subjective self-reporting strategies. One example is that pre-test scores can cause frustration, nervousness and trauma for trainees if they are asked to be tested on items that they have not yet learned. Training; however, provides subjects with more accurate information that

can improve their understanding and can change their perception from their initial level of functioning. This weakens pre-and-post treatment comparisons and is the source of response-shift-bias, a source of contamination of the training outcomes. In order to minimize response-shift-bias, several training evaluation experts like Mezzof (1981), Stufflebeam and Wingate (2005), Darling and Gallagher (2003), Bray and Howard (1980), Taylor and Taylor (2009), Hill and Betz (2005), and Lam and Bengo (2003) used different prospective and retrospective self-reporting methods. Below sections discuss different self-reporting issues, self-reporting methods and their pros-and-cons in training assessment.

Mezzof (1981) identified a major issue of self-reporting: The issue of response-shift bias (p. 3). Mezzof (1981) challenged the traditional prospective self-report training method and argued that a learning evaluation calculation based on a pre-test and post-test measurement is not able to provide accurate learning outcomes, because trainees reported using the traditional pre- and post-testing methods often underestimated or overestimated training benefits. Therefore, Mezzof proposed a "then-test" to minimize what he described as response-shift-bias. The test process is called "Pre-Then-Post" testing where participants reflect back after training to their level of functioning prior to the training and re-state themselves in addition to the pre-post measure of traditional self-report. (p. 4). Mezzof uses a one-way analysis of variance (ANOVA) strategy to compare pre-then tests and then-posttests. His empirical study finds the pre-then-post (retrospective) testing eliminates response shift bias (Ibid, p. 5). A fictitious code number is used when testing participants in the then-posttest, which allows analysts to anonymously compare the then-posttests with the pre-then tests. Although this method is easy to administer; it is important to develop appropriate self-reporting questionnaires to eliminate the instrumental error-response-shift bias. Moreover, Lam (2009), through their findings, suggest that self-assessment is not be the only measure that is paid attention to in self-assessments (in Lam, 2009, pp. 4-5).

Stufflebeam and Wingate (2005) also uses a self-reporting assessment method for evaluating learning outcomes by using the Self-Assessment of Program Evaluation Expertise (SAPEE) method. His study has three types of participants: Novice evaluators, experienced evaluators and experts. He compares these three groups through learning gains. Participants' *Self-Assessment Mean Pretest* and *Post-test* Scores are taken by novice participants and experienced participants (Stufflebeam & Wingate, 2005, p. 14). This in-depth study was designed for three weeks training for the novices and two weeks for experienced participants. The self-assessments included needs assessments so that the instructor could understand which areas needed more attention and to help participants know what they learned (Stufflebeam & Wingate, 2005, p. 3). The results were converted to an average score and this result was compared with participant from other years' pre-and-post instruction results (p. 9). The study showed that the average gains in perceived evaluation expertise made by novice participants were substantially larger than those of the experienced participants. This is because (a) the novice participants received more instruction than the experienced participants, and (b) the novice participants had more room for improvement. This indicates that the extra week of training may be effective in closing the gap between experienced and novice participants (pp.13- 15). However, there were several limitations found in this study. An example is that participants inflated their rating of learning gains because of a desire to please the institution's staff members. It needs further validation of this method. There are many other training evaluation methods encapsulate below that are identified by different training evaluation experts.

2. Different training evaluation methods

- Self-reporting (Solely) Method Self-assessment in combination with other methods other than self-reporting
- Mezzof (1981) Retrospective testing, identified response-shift bias Darling and Gallagher (2003) CSPD, Need Assessment, Multivariate analysis (MONOVA) of Bamberger (2004): Shoestring + project design, contextual factor, No control and Experienced. Groups
- Stufflebeam and Wingate (2005) method SAPEE, pre- and post- test Taylor and Taylor (2009): Have

clear wording, conventional and retrospective test. Eckert's (2001) method is considered best training design that talks about situational factors analysis with checklists, use pre-and post-tests, No control and experienced group

- One training evaluation method called "Multilevel self-assessment training evaluation", Lam and Bengo (2003) method called HLM, compare conventional, retrospective and perceived change. This also has multilevel assessments + direct observation + Need assessment+ perceived change + program design. Haccoun (2008) method contains "One group-pretest and posttest design, ANOVA, IRS, Identify Type -1 Error and Type-2 Error."

- Hill and Betz (2005) training evaluation method compares prospective and retrospective test, effect size, prospective and retrospective test, Subjective self-report measure, Bray and Howard (1980) retrospective test, IDEA, social content analysis to look for subject response style effects Mayne (1999) Performance measurement analysis, analysis of existing files, secondary data and case study

- Blanchard (2009) training valuation is about self–reporting, theory and practice; Brinkerhoff (2003) model is called "Success Case Method (SCM)", Hoogstraten (1982) training evaluation method is called "Seeing Problems Strategy (SPS), conventional and retrospective test, look at subject response style effect:, Aurther, Edens, Bell and Bennett (2003) training evaluation focus on need assessment + Meta-Analysis of design + paper –Pencil test

- Jenkins and Curtin (2001) theory centered on "Job analysis method."

- Kaupins (1997) method called "Live Case Analysis, internship, and chalk board-display paper."

- Darling and Gallagher's (2003) study looked at the requirements for Comprehensive System of Personnel Development (CSPD) pre-service and in-service training for those living with disabilities (p. 2). This CSPD method has three stages of self-assessment: SA1: administered at the beginning of training and reviewed responses, SA2: overseen at the end of the training to self-reflect (p. 4) and in SA3 assessment questionnaires are mailed out with their certificates after three months. (p. 4). A repeated measure MANOVA (multivariate analysis of variance) with SPSS program uses in this experiment to determine the improvement of participants over time (p. 5). This CSPD method also requires a needs assessment to verify and clarify issues like strengths and weakness for participants (Darling & Gallagher, 2003, p. 2). During a 5-day training module a participant completes 15 self-assessments. Although the CSPD method is an in-depth study; it is time-consuming. SA3 should be avoided because it is administered too late to assess learning and provide feedback. It is unreliable, as the evaluator is subject to the trainee mailing back the information. Below the paper explores a retrospective self-reporting study that covers different types of subjects like instructor and student self-rating for self-reporting assessments.

- Bray and Howard's (1980) retrospective self-reporting test has five parts: instructor, student-rated progress, course evaluation, student self-rating and student demographic data (p. 65).

Table 1

Different training evaluation methods

Self-reporting (Solely) Method	Self-assessment in combination with other methods	Methods other than self-reporting
Mezzof (1981) Retrospective testing, identified response-shift bias	Darling & Gallagher (2003) CSPD, Need Assessment, Multivariate analysis (MONOVA)	Bomberger, Church, and Fort (2004) Shoestring + project design, contextual factor, No control and Experienced. Groups

Table 1 … continued

Self-reporting (Solely) Method	Self-assessment in combination with other methods	Methods other than self-reporting
Stufflebeam and Wingate (2005) SAPEE, pre- and post- test	Taylor and Taylor (2009): clear wording, conventional and retrospective test	Eckert (2001) Best training design, situational factors, Checklists, use pre-and post-tests, No control and experienced group
Multilevel self-assessment	Lam and Bengo (2003) HLM, compare conventional, retrospective and perceived change, Multilevel assessments + direct observation + Need assessment+ perceived change + program design	One group-pretest and posttest design, ANOVA, IRS, Identify Type -1 Error and Type-2 Error
Hill and Betz (2005) compare prospective and retrospective test, effect size, prospective and retrospective test, Subjective self-report measure,	Bray and Howard (1980) retrospective test, IDEA, social content analysis to look for subject response style effects	Mayne (1999) Performance measurement analysis, analysis of existing files, secondary data and case study
Blanchard (2009) Self –reporting, theory and practice		Brinkerhoff (2003) Success Case Method (SCM)
Hoogstraten (1982) Seeing Problems Strategy (SPS), conventional and retrospective test, look at subject response style effect		Aurther, Edens, Bell, and Bennett (2003) Need assessment + Meta-Analysis of design + paper –Pencil test
		Jenkins and Curtin (2001) Job analysis method
		Kaupins (1997) Live Case Analysis, internship, chalk board-display paper

3. Descriptions of different training methods

Bray and Howard's (1980) retrospective self-reporting test has five parts: instructor, student-rated progress, course evaluation, student self-rating and student demographic data (p. 65). The method uses multivariate analysis for collected data analysis (p. 66). Participants rate themselves with regard to their knowledge just after training and as they were before the training. Then each is scored separately (p. 64). Howard and his colleagues and many other training evaluators aimed at demonstrating the response–shift-bias phenomenon and the superiority of retrospective pre-tests over traditional pre-tests (Bray & Howard, 1980; Hoogstraten, 1982). Bray and Howard's (1980) study looks at whether *response-shift bias* is a threat to internal or construct validity (p 63), finding a solution in using retrospective pre-tests, comparing test data and then determining the self-rating impact assessment on training measurement. This method is a little different than Mezzof's. Bray and Howard (1980) use the instructional development and effectiveness assessment (IDEA) for student rating. Training was successful in the areas of teaching assistants' (TA's) self-reports of teaching, actual teaching and student ratings of instruction (p. 66). Although the findings indicate several benefits such as teacher training programs as a way to improve teaching ability and effectiveness; the program has subject response style effects (memory distortion, social desirability, compliance with implicit demand characteristics, and many others) on retrospective pre-test ratings which can be a high contaminator (Bray & Howard, 1980, p. 64). However, the perception of learning knowledge does not reflect an exact measurement.

Howard et al. (1979) also conducted a study on the validity of retrospective pretest and concluded that retrospective pre-test-post-test comparisons yield more valid results than conventional pre-test comparisons, but informed pre-tests do not improve self-reporting accuracy. They found two possible causes of response-shift bias: (a) memory distortion (forgetting), (b) subject response-style effects (social desirability, subject acquiescence). To combat this, they recommended collecting self-reporting ratings of pre-intervention performance retrospectively rather than prospectively; as well as suggesting uniform standard measurement for both then-test and post-test ratings. Although several benefits can be found in this method, there was no information collected

about the student achievement measure information. Therefore, the actual effectiveness of the trainers is not completely known and thus needs further study (Bray & Howard, 1980, p. 69).

Hill and Betz (2005) look at different training evaluation methods including self-reporting measure especially examining and critiquing of self-reporting training evaluation. Mayne's (2001) performance-based measurement for training analysis although Shadish, Cook, and Campbell, (2002) suggest for objective performance assessments that can considered superior to self-reporting (in Hill and Betz 2005 p. 502). Hill and Betz (2005) assert that performance-based measurement may be supplemented or replaced by subjective self-reporting measures. They (Hill and Betz) conduct a comparison study on retrospective assessment with prospective rating bias and show that in a prospective test program effect is underestimated, where it is overestimated in retrospective tests, which is different from the Stufflebeam and Wingate (2005) study.

Moreover, Hill and Betz (2005) have concerns about the validity of the retrospective method, which is a primary threat to program evaluation because there are some degrees of *recall bias* (distortion or degradation of memory) in all retrospective ratings. Hill and Betz (2005) place emotional biases under the category *socially desirable or impression management responses* (Hill & Betz 2005, p. 504). This finding is relevant to other program evaluation ratings; including many others (King & Bruner, 2000 in Hill & Betz, 2005, p. 504, Lam & Bengo, 2003). All of these studies find other cognitive biases in retrospective rating that are *implicit theories of change*. Aronson and Mills (1959) call this cognitive dissonance bias *effort justification bias* (Hill & Betz, 2005, p. 505). On retrospective pre-tests, clients may be able to provide more accurate estimates of pretreatment behaviors; however, in traditional pretests this may yield an underestimate of treatment effects or, even worse negative effects.

Hill and Betz (2005) also study *Effect size* (ES), the standardized measure of different scores between then-pre-test-posttest outcome variable. The findings say the pretest-posttest ES is slightly larger .52 than ES reported from the program's experimental trial (p. 510). However, here the primary problem of the study is there are no objective criteria to which pre-test and post-test results can be measured against. They suggest for future research for incorporating objective criteria for more fine-grained analysis of pre-test-then-test differences. A well-designed research comparing both types of pre-test across multiple studies would provide a useful benchmark for exploration of pre-test–then-test differences. In this respect, Lam and Bengo's (2003) study is prominent.

Lam and Bengo's (2003) self-reporting data collection system includes a needs assessment, service utilization, and program processes (design) and uses them in different evaluation projects to determine pre-intervention status of knowledge and post –intervention status. Their study analysis has three self –reporting methods: the post and retrospective pre-test method, the post and perceived changed method, and perceived change method. Their study shows that teachers in the post and retrospective pre-test condition reported least change, but in the perceived change condition, participants reported the greatest change. Schwaz and Oyserman (2001) say that this significant in change scores is the intervention-related change (in Lam & Bengo, 2003, p. 66). However, Eckert (2000) concludes that these tests designs have inherent validity threats, which are caused by self-reporting in the pretest because a pretest can have a lower internal validity by introducing a carry-over or practice effect. This results in post-test participants recalling their responses and inflating their performance on the post-test. Lam and Bengo (2003) call it the *carry-over-effect* that can minimize internal validity threats.

Lam (2003) identified other drawbacks in pre-test models where trainees depreciated their pretest scores to inflate their treatment-counterfactual estimation is also a concern in training evaluation. Therefore, Lam and Bengo (2003) suggest not using pre-tests prior to intervention in measuring self-reporting change. Within the literature, there is an indication that response-shift-change measurement obtained from post and retrospective pre-test methods are often more accurate estimates of change than those obtained from traditional, pretest-posttest design (Bray & Howard, 1980; Hoogstraten, 1982; Howard & Dailey, 1979; Howard et al., 1979a, 1979b in Lam & Bengo, 2003, p. 69). However, Lam and Bengo's comparative study on three methods

report that all have response-shift-bias differences. Therefore, Lam and Bengo ask for the use of multiple methods including social desirability measures in future research to avoid the false response in testing (p. 78). They suggest for the post- method for self-reporting estimation (2003, p. 69). However, Lam and Bengo do not explain the design of their retrospective self-reporting method. They do suggest further research that can substantiate this approach (2003, p. 66).

Hoogstraten (1982) identifies three causes of response-shift bias: initial lack of information, memory-effects and subject response style. He investigates the relative validity of self-reporting measures using two conventional pre-test-post-test designs and retrospective pre-test-post-test designs. Here subjects receive training in Seeing Problems Strategy (SPS) and are assigned to three conditions at random. The findings are the then-post scores reflect actual performance changes while conventional pre-post scores do not. Therefore, Hoogstraten (1982) suggested that the retrospective pre-test is a valid means to control for response-shift bias. However, this experiment did not allow researchers to determine the impact of these three causes on response shift bias.

Lam (2009) challenges self-reporting assessment method and argues with evidence that aggregated self-assessments results both at the single level and multilevel analysis [Hierarchical Linear Models (HLM)] are not generalizable due to individual differences, contextual differences, assessment content and procedural factors (p 12). Lam says generalizability based on over- or under-estimates for both the self-assessment and criterion assessment is flawed (p. 12). To fully capture both over- and underestimation biases, he suggests evaluators should use both difference-in-means and correlation indices to determine the validity and subsequently the usability of self-assessments for measuring training outcomes and effects (Lam, 2009, p. 4). Lam says that by only using self-assessment, can only determine a participant's performance not overall training effectiveness (Lam, 2009, p. 6). With regards to this, Lam proposes for further research on aggregated self-assessment for its validity effectiveness and to make it efficient and more professional (p. 12). He offers nine points to use self-assessment to determine workshop success (p. 12). He also suggests for multilevel analysis or Hierarchical Linear Models (HLM) a special regression analysis procedure that looks at the aggregation effect by using information from all levels. This multivariate analysis is more efficient and precise, but cumbersome. Moreover, the HLM method needs statistical advanced knowledge and is expensive for measuring learning gains.

Taylor, Russ-Eft, and Taylor's (2009) research contains trained supervisors and untrained subordinate participants in tests. Conventional and retrospective pretest self-reporting scores are compared using a correlated *t* test. The conventional pre-test ratings are typically found to be significantly higher than retrospective pre-test ratings and lead researchers to conclude that a response-shift bias has occurred. Training effects based on self-ratings are substantially larger (Taylor, Russ-Eft, & Taylor 2009, p. 40). Lam and Bengo (2003) also found the same results. Participants indicated even greater changes when asked about the degree to which they have changed than when asked separately for retrospective pre-test and post-test ratings (in Taylor et al., 2009, p. 41). Although retrospective pretests may prevent response-shift bias, they may introduce other biases that inflate intervention effects. For example, individuals may be motivated to exaggerate their improvement to reflect favorably on themselves. Similarly, they may be motivated to show a substantial improvement regardless of their actual improvement to justify the effort that they have expended in completing the program (Taylor et al., 2009, p. 32). Therefore, the problem is more likely in a retrospective pretest design because individuals complete both pre-test and post-test ratings at the same time and are thus better to manipulate their pretest ratings to show an improvement. Hence the use of retrospective pretests for any rating source could be suspect (Taylor et al., 2009, p. 34). To avoid this response-shift- bias, Taylor et al. (2009) recommend it should be addressed through careful construction of clearly worded measures rather than using retrospective pretests.

The Canadian Society for Training and Development (CSTD 2009) uses three methods for training measurement: (1) The Immediate Impact Questionnaire (IIQ), (2) The Job Impact Questionnaire (JIQ), and (3) The Effective Practices Audit (EPA) of training evaluation to evaluate different trainings outcomes. The Job Impact Questionnaire (JIQ) evaluates the impact of a training program on job performance following the participants' self-report in light of four key questions: applying the knowledge and skills; application of the

learning for improving job performance, improved performance impacting business results, and difficulties faced by the participants to apply their learning on-the-job. This self-reporting training evaluation method is very simple and is widely-used in Canada. However, it is important that CSTD include 'evaluators' direct observation' while collecting participants self-reporting data as suggested by Lam.

Below the paper look at other alternative training evaluation methods to compare status within self-reporting training evaluation methods. Bamberger, Church and Fort (2004), Eckert (2001), Haccoun and Hamtiaux (1994), and Mayne (1999) have different training evaluation methods, which do not use self-assessment training evaluation methods. For example, Haccoun and Hamtiaux (1994) use a simple procedure for estimating the effectiveness of training on trainee knowledge through the Internal Referencing Strategy (IRS) as compared to a traditional experimental evaluation (p. 593). Haccoun and Hamitaux (1994) identify two types of errors: Type 1 and Type 2 errors in the study. To avoid these errors, Haccoun suggests parallel pre- and post-tests (Haccoun & Hamitaux (1994), p. 597). However, IRS is still susceptible to type II errors and should not be used to replace more complex and rigorous designs (Haccoun & Hamitaux, 1994, p. 603). IRS may not be used for behavioral or higher learning levels (p. 603). While, the shoestring evaluation approach collects training data without using a control group and baseline data for the project group, but it has six intensive steps. This method is used when working within time and budgetary constraints and limitations on data accesses. Evaluators collected secondary data using participatory methods and used checklist for quantitative data. Although Bamberger's training evaluation finds solutions to the varied threats to validity and adequacy of evaluation designs; this method is very detailed, intensive, complicated, and expensive.

Eckert (2001) and Mayne (1999) do not believe in using the pre-test and post-test control groups and experimental groups for training learning and change behaviors. Rather, they emphasize on the best training evaluation designs, which can identify internal threats and address them (p. 186). Eckert (2001) places emphasis on the right setting or situational factors, regression analysis and uses a checklist to determine plausible threats to validity. Here Eckert's method's advantage is that the work offers a better application of designs not better designs (p. 192) and is less expensive. However, regression analysis cannot provide case-to-case participants variation results. Mayne's (1999) performance measurement attempts to program contribution analysis, which explores and demonstrates performance measures. Existing program files, secondary analysis, and case study evidence are used for program measurement. The features of the contribution analysis are: acknowledging the problem, presenting the logic of the program, identifying and documenting behavioral changes, and many others. Mayne's logic model chart encourages programs to be more precise in program designs (1999, p. 9). However, here expert opinions and a structured survey need to give support evidence about the contribution of the program.

Throughout the study, it is found that several training evaluation models and methods are developed by different training evaluators. The classical one is Kirkpatrick and his four levels of measurement. Others are Philips and Stone's (2007) five levels of training outcomes, Alliger's (1997) Augmented Framework Training Taxonomy Model (learning sub-levels), Eseryel's (2002) ADAPT–IT instructional design tool, Kraiser, Ford and Sales' (2002) primary classification scheme of learning outcomes-Cognitive, skill-based and effective learning outcomes. The Context, Input, Process and Product (CIPP) model, and Alvarez et al.'s (2004) integrated model are used for training evaluation. All of these models suggest measuring more than one training outcome level like reaction, learning outcomes, behavior change and results. However, Blanchard, Thacker and Ways (2000) share that not all companies need follow all these steps of training evaluation. The measurement of all training steps depends upon organizational demand, executives' interest, budget, time and manpower. Blanchard, Thacker and Ways (2000) argue that there is a difference for training evaluation academia versus practitioners' practices. Academicians are proponents of evaluating Kirkpatrick's four-level training (p. 2) to justify the worth in constantly improving training and prefer follow all levels of training evaluation, but practitioners tend not to follow all of Kirkpatrick's levels of training evaluation (p. 3). Blanchard suggests there should be a balance among academics and practitioners. They should communicate with each other to discuss current training issues (p. 9). Blanchard's study shows that reaction is the most used for both management-employee and

non-management-employee training (p. 4). More than half of companies are not evaluating training based on behavior or results levels. (p. 5). According to Blanchard, Thacker and Ways (2000) levels of evaluation should be conducted depending upon the client's objectives, despite the perceived need for all four.

4. Grameen Bank training evaluation system in Bangladesh

Below the section narrates how self-reporting can be used in Grameen Bank (GB) to better inform training practices and looks at how this model may apply to other microfinance institutions' training evaluation. In the GB context, measuring knowledge of learning gains on loan delinquency depends on various factors such as cliental situations, budget and organizational policies. Even employees can perceive the problems from different perspectives. Therefore, trainees' responses in pre-then-post testing self-assessments will vary. Self-assessment is undoubtedly the most efficient data-collection method in training evaluation (Lam, 2009, p. 12). However, invalid self-assessment cannot be used to gauge training effectiveness and self-assessment data with equal underestimates and overestimates are invalid. In this situation, Hill and Betz (2005) concisely recommend retrospective pre-tests for the examination of subjective experiences of program-rated change that can be used in GB to measure trainee learning gains.

4.1 Implications of self-reporting evaluation training tools to skills development of Grameen Bank employees in Bangladesh

Grameen Bank follows a minimalist decentralized training evaluation approach. GB's training evaluation data collection uses the pre-test and post-test control group and experimental group (Eckert method) and uses IRS with pre-post single group training evaluation design (Haccoun & Hamitaux, 1994), which involves many stages. Moreover, the use of Mayne's training performance validity measurement and Bamberger, Church and Fort's (2004) shoestring method incorporating checklists for quantitative secondary data collection will be complicated, expensive, and cumbersome, and grassroots employees may feel uneasy and nervous to go through all the steps of training evaluation. Hence, asking questions of trainees about learning gains following the retrospective method could be one method in the GB training measurement.

GB has conducted several delinquent loan recovery workshops for its employees and informally asked them about their perception of the delinquency problems and possible solutions for reducing loan delinquencies. Some participants realized the issue(s) after the workshop and shifted their self-assessment in comparison to how they did before the workshop (different mental yardstick). Now, the author of this paper can anticipate a response-shift-bias with Grameen Bank (GB) trainees, employees and managers in their in-service pre-training and post-workshops, especially in measuring knowledge of learning gained on loan delinquency, which is dependent on various others factors such as cliental socio-economic situations. Hence, training levels and trainee responses in pre-then-post testing self-assessments vary from place to place and varied social contexts. Blanchard et al. (2000) identify this and provide an example where some respondents, using Kirkpatrick's levels in different ways, like an integrated strategic human resource system where the performance review process assesses employee behavior and organizational effectiveness throughout the process. However, respondents using this system may not be able to answer all review questions (Blanchard, Thacker, & Ways, 2000, p. 7), but in the perceived-change method, questions are asked directly to trainees-how participants think they have changed as a result of an intervention (Lam & Bengo, 2003).

Moreover, the distinction of uses of GB's relevant and irrelevant training content can identify training gained learning and gaps with respective to various items. In this context, Darling and Gallagher's CSP training evaluation strategy, Stufflebeam and Wingate (2005) shoestring SAPEE method, Bray and Howard's (1980) IDEA methods are all centralized training evaluation that have huge costs, require a lot of staffing and other resources for evaluating 30,000 employees at 2575 branches in Grameen Bank in Bangladesh. The above methods cannot give answers to GB's relevant and irrelevant items. Compared to all methods discussed, Mezzof's, simple framework for measuring the response-shift bias can help GB in a formal way to determine

response-shift biases with some limitations. Internships that include job shadowing strategies provide regular direct feedback to GB's relevant and irrelevant items. Although there are many models and methods developed for the treatment of evaluation, to minimize counterfactual estimation, response-shift biases and effect size, Grameen Bank job shadowing could incorporate intensive observation while directly asking questions of interns using retrospective method. This can impact upon trainees and managers so that they can think intensively about the problems of loan delinquencies and their solutions. Hence, the paper recommends GB branch managers would prefer a decentralized, self-reporting training evaluation data collection method by directly asking questions of trainees at the end of their internships using the retrospective self-reporting method at the branch level until a refined training gains measurement that is tested may be published.

5. References

Aronson, E., & Mills, J. (1959). The effect of severity of initiation on liking for a group. *Journal of Abnormal and Social Psychology, 59*, 177-181. https://doi.org/10.1037/h0047195

Arthur Jr., W., Edens, P. S., Bell, S., & Bennett Jr., W. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, *88*(2), 234-245. https://doi.org/10.1037/0021-9010.88.2.234

Birkenbach, X. C. (1986). Self-report evaluations of training effectiveness: Measuring alpha, beta, and gamma change. *South African Journal of Psychology, 16*(1), 1-7. https://doi.org/10.1177/008124638601600101

Blanchard, J. (2009). *Teaching, learning and assessment,* UK: McGraw-Hill Education.

Blanchard, P. N., Thacker, J. W., & Way, S. A. (2000). Training evaluation: Perspectives and evidence from Canada. *International Journal of Training and Development*, *4(*4), 295-304. https://doi.org/10.1111/1468-2419.00115

Bomberger, M., Church, M., & Fort, L. (2004): Shoestring evaluation: Designing impact evaluations under budget, time and data constraints. *American Journal of Evaluation*, *25*(1), 67-87. https://doi.org/10.1177/109821400402500102

Bray, J. H., & Howard, G. S. (1980). Methodological considerations in the evaluation of a teacher-training program. *Journal of Educational Psychology*, *72*(1), 62-70. https://doi.org/10.1037/0022-0663.72.1.62

Brinkerhoff, R. O. (2003). *Using the success case impact evaluation method to enhance training value and impact*. MI: The Learning Alliance Inc.

Darling, S. M., & Gallagher, P. A. (2003). Using self-assessments in early intervention training. *Journal of Early Intervention*, *25*(3), 219-227. https://doi.org/10.1177/105381510302500306

Eckert, W. A. (2000). Situational enhancement of design validity: The case of training evaluation at the World Bank Institute. *American Journal of Evaluation*, *21*(2), 185-193. https://doi.org/10.1177/109821400002100205

Eseryel, D. (2002). Approaches to evaluation of training: Theory and practice. *Educational Technology & Society, 5*(2), 1-10.

Gilovich, T., Griffin, D. W., & Kahneman, D. (2002). *Heuristic and biases*. UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

Haccoun, R. R., & Hamtiaux, T. (1994). Optimizing knowledge tests for inferring learning acquisition levels in single group training evaluation designs: The internal referencing strategy. *Personnel Psychology*, *47*, 593-604. https://doi.org/10.1111/j.1744-6570.1994.tb01739.x

Hill, L. G., & Betz, D. L. (2005). Revisiting the retrospective pre-test. *American Journal of Evaluation*, *26*(4), 501-517. https://doi.org/10.1177/1098214005281356

Hoogstraten, J. (1982). The retrospective pre-test in an educational training context. *Journal of Experimental education*, *50*(4), 200-204. https://doi.org/10.1080/00220973.1982.11011824

Howard, G. S. (1980). Response-shift bias, a problem in evaluating interventions with pre/post self-reports. *Evaluation Review*, *4*(1), 93-106. https://doi.org/10.1177/0193841X8000400105

Howard, G. S., & Dailey, P. R. (1979). Response-shift bias: A source of contamination of self-report measures. *Journal of Applied Psychology*, *64*, 144-150. https://doi.org/10.1037/0021-9010.64.2.144

Howard, G. S., Dailey P. R., & Galvanick, N. A. (1979). The feasibility of informed pretests in attenuating response shift bias. *Applied Psychological Measurement*, *3*, 481-494. https://doi.org/10.1177/014662167900300406

Howard, G. S., Ralph K. M., Gulanick, N. A., Maxwell, S. E., Nance, D.W., & Gerber, S. K. (1979). Internal invalidity in pretest-posttest self-report evaluations and a re-evaluation of retrospective pretests. *Applied Psychological Measurement, 3*(1), 1-23. https://doi.org/10.1177/014662167900300101

Howard, G. S., Schmeck, R. R., & Bray, J. H. J. (1979). Internal validity in studies employing self-report instruments: A suggested remedy. *Journal of Educational Measurement*, *16*(2), 129-135. https://doi.org/10.1111/j.1745-3984.1979.tb00094.x

Jenkins, S. M., & Curtin, P. (2006). Adapting job analysis methodology to improve evaluation practice. *American Journal of Evaluation*, *27*(4), 485-494. https://doi.org/10.1177/1098214006294303

Kaupins, G. (1997). Trainer opinions of popular corporate training methods. *Journal of Education for Business*, *73*(1), 5-8. https://doi.org/10.1080/08832329709601607

King, M. F., & Bruner, G. C. (2000). Social desirability bias: A neglected aspect of validity testing. *Psychology and Marketing, 17*(2), 79-103. https://doi.org/10.1002/(SICI)1520-6793(200002)17:2<79::AID-MAR2>3.0.CO;2-0

Kirkpatrick, D. L. (1977). Determining training needs: Four simple and effective approaches. *Training and Development Journal*, *31*(2), 22-25.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, *78*, 311-328. https://doi.org/10.1037/0021-9010.78.2.311

Lam, T. (2009). Do self-assessments work to detect workshop success? An analysis of argument and recommendation by D'Eon et al. *American Journal of Evaluation, 30*(1), 93-105. https://doi.org/10.1177/1098214008327931

Lam. T. C., & Bengo, P. (2003). A comparison of three retrospective self-reporting methods of measuring change in instruction practice. *American Journal of Evaluation*, *24*(1), 65-80. https://doi.org/10.1177/109821400302400106

Mayne, J. (2001). Addressing attribution through contribution analysis: Using performance measures sensibly. *The Canadian Journal of Program Evaluation*, *16*, 1-24.

Mezzof, B. (1981). How to get accurate self-reports of training outcomes. *Training and Development Journal*, *35*(9), 56-61.

Pohl, N. F. (1982). Using retrospective pre-ratings to counteract response-shift confounding. *Journal of Experimental Education*, *50*, 211-214. https://doi.org/10.1080/00220973.1982.11011826

Pratt, C. C., McGuigan, W. M., & Katzev, A. R. (2000). Measuring program outcomes: Using retrospective pretest methodology. *American Journal of Evaluation, 21*(3), 341–349. https://doi.org/10.1177/109821400002100305

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, *52*, 471-499. https://doi.org/10.1146/annurev.psych.52.1.471

Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, *22*(2), 127–160. https://doi.org/10.1177/109821400102200202

Shadish, W. R., Cook, T, D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. *Social Science Review*, *76*(3), 365-386.

Stufflebeam, D. L., & Wingate, L. A. (2005). A self-assessment procedure for use in evaluation training. *American Journal of Evaluation*, *26*(4), 544-561. https://doi.org/10.1177/1098214005279730

Taylor, P. J., Russ-Eft, D. F., & Taylor, H. (2009). Gilding the outcome by tarnishing the past: Inflationary biases in retrospective pre-tests. *American Journal of Evaluation, 30*(31), 31-43. https://doi.org/10.1177/1098214008328517

The Canadian Society for Training and Development (CSTD). (2009). Invest in people's project. Retrieved from http://www.cstd.ca/investing_in_people/tools.html