This is the peer-reviewed, accepted manuscript of an article published by Sage in the *Evaluation Review* on August 31, 2012. The final version is available online (https://doi.org/10.1177/0193841X12458103): Daigneault, Pierre-Marc (2012). 'Measuring stakeholder participation in evaluation: An empirical validation.' *Evaluation Review*, 36 (4), 243-270.

Measuring Stakeholder Participation in Evaluation: An Empirical Validation of the Participatory Evaluation Measurement Instrument (PEMI)

Pierre-Marc Daigneault

Abstract: *Background.* Stakeholder participation is an important trend in the field of program evaluation. Although a few measurement instruments have been proposed, they either have not been empirically validated or do not cover the full content of the concept. *Objectives.* This study consists of a first empirical validation of a measurement instrument that fully covers the content of participation, namely the Participatory Evaluation Measurement Instrument (PEMI). It specifically examines 1) the intercoder reliability of scores derived by two research assistants on published evaluation cases; 2) the convergence between the scores of coders and those of key respondents (i.e., authors); and 3) the convergence between the authors' scores on the PEMI and the Evaluation Involvement Scale (EIS). *Sample.* A purposive sample of 40 cases drawn from the evaluation literature was used to assess reliability. One author per case in this sample was then invited to participate in a survey; 25 fully usable questionnaires were received. *Measures*. Stakeholder participation was measured on nominal and ordinal scales. Cohen's kappa, the intraclass correlation coefficient and Spearman's rho were used to assess reliability and convergence. *Results.* Reliability results ranged from fair to excellent. Convergence between coders' and authors' scores ranged from poor to good. Scores derived from the PEMI and the EIS were moderately associated. *Conclusions.* Evidence from this study is strong in the case of intercoder reliability and ranges from weak to strong in the case of convergent validation. Globally, this suggests that the PEMI can produce scores that are both reliable and valid.

Keywords: Participatory evaluation; stakeholder involvement; stakeholder participation; collaborative inquiry; measurement instrument; empirical validation.

Background and Research Problem

"One of the larger trends in evaluation theory and practice is an increased focus on stakeholder participation" (Mark, 2001, p. 462). Numerous evaluation approaches such as collaborative, democratic-deliberative, empowerment, fourth-generation, inclusive and utilization-focused to name a few, explicitly endorse the principle of stakeholder participation. The abundance of terms used to designate evaluation theories and models in which stakeholders are significantly involved "is surely an indication that participatory approaches to program evaluation are coming of age" (King, 1998, p. 58). Indeed, the participatory principle is now widely accepted, some would even say hegemonic, within the evaluation community (Fleischer, Christie, & LaVelle, 2011; Biggs, 1995, as cited in Gregory, 2000, p. 180; Mathison, 2005; Shea & Lewko, 1995; Whitmore, 1998).

Stakeholder participation is not only a rhetorical device, but also a phenomenon that has taken root in evaluation practice in various contexts of evaluation practice (e.g., Cousins et al., 2011; Cullen, Coryn, & Rugh, 2011; Thayer & Fine, 2001).

Stakeholder participation is one of the major constructs that has caught the attention of researchers, especially those interested in evaluation use (Cousins, 2003; Cullen et al., 2011; Johnson et al., 2009; Poth & Shulha, 2008). Yet, in order for empirical research to contribute effectively to knowledge development, a sound conceptualization and operationalization of stakeholder participation is needed. To that end, Daigneault and Jacob (2009) have developed — based on the work of Cousins and Whitmore (1998) — what they deemed to be a coherent, parsimonious yet content-valid conceptualization of participatory evaluation (PE). Their framework possesses three constitutive dimensions that are theorized as necessary and sufficient conditions for the concept of PE: *extent of involvement*, *diversity of participants* and *control of the evaluation process*. The dimensions are measured on a five-point ordinal scale ranging from .00 (absence of the dimension) to 1.00 (full presence of the dimension). In between these two extremes, .25, .50 and .75 represent a limited, moderate and substantial level of the dimension, respectively. Because of the necessary and sufficient condition concept structure, the overall level of stakeholder participation they have proposed is logically determined by the *minimum* of the three dimensions (Goertz, 2006). For instance, an evaluation case with scores of .75 on the first two dimensions and .25 on the third one would get an overall score of participation of .25. Four dichotomous indicators respectively representing involvement in various evaluation tasks and types of stakeholders involved serve to operationalize the extent of involvement and diversity of participants dimensions (Daigneault & Jacob, 2009). Control of the evaluation process, by contrast, has not really been operationalized with precision. Rating of this dimension is indeed based on a subjective assessment of the balance of power between the evaluator and participants (from exclusive control by the evaluator to exclusive control by participants).

This framework has given rise to a few applications that seem quite promising (Connors & Magilvy, 2011; Jacob, Ouvrard & Bélanger, 2011; Laudon, 2010). For instance, Connors and Magilvy (2011) have positively assessed their use of the framework:

> Overall, we found the index both easy to implement and to understand and relevant to the work of the CON [College of Nursing] evaluation. The language of the instrument was clear and familiar. In particular, examining the degree of stakeholder participation at the four decision points (design, data collection/analysis, judgment, and dissemination) aligned with the collaborative process used in the CON program evaluation. The rating on the dimension of control was the most subjective, as acknowledged by Daigneault and Jacob, when thinking in general terms about the evaluation. However, when specific evaluation decisions were reviewed retrospectively, we were easily able to assign a rating on this dimension. From our perspective, the scale fulfilled Daigneault and Jabob [sic] goals that the instrument be parsimonious, consistent in structure, and useful for differentiating participatory from nonparticipatory evaluation practices. (pp. 82-83)

Yet, the framework as a measurement instrument has not been empirically validated and could clearly benefit from more specific guidance with respect to how to rate each dimension, especially control. Legitimate doubts have indeed been raised about the reliability and validity of the instrument in its nonvalidated form (Cullen, 2009; Cullen et al., 2011).

Since the instrument — hereafter labeled the *Participatory Evaluation Measurement Instrument* (PEMI) for convenience — was specifically developed for the purpose of conducting sound

empirical research, it appears necessary to proceed to a first empirical examination of its reliability and validity. Assessing the reliability of the scores generated by two different coders is indeed a first, fundamental, step in any validation study (DeVellis, 2005, Fleiss, Levin, & Paik, 2004). Next, it is important to assess whether the scores derived from the PEMI can be interpreted as actually measuring the concept of stakeholder participation (see Carmines, Woods & Kimberly, 2005).

Research Objectives and Hypotheses

This study consists of an empirical validation of the PEMI. Specifically, three objectives are pursued:

1. *Intercoder reliability assessment.* To examine the level of intercoder reliability achieved by two research assistants working independently using the PEMI on published reports of a sample of published evaluation cases.
2. *Convergent validation I*. To examine the extent to which the scores achieved by two independent coders, once discrepancies are resolved through discussion, align with those of a key respondent for each case.
3. *Convergent validation II*. To examine the convergence between the scores obtained by key respondents on the PEMI and on the *Evaluation Involvement Scale* (hereafter EIS, Toal, 2009), a validated — albeit imperfect — measure of stakeholder participation.

[INSERT FIGURE 1 ABOUT HERE]

We hypothesize that the scores of coders will display a "fair" level of agreement or better (i.e., Cohen's kappa must be greater or equal to .40). Following the reliability assessment, it is essential to check whether the coders' scores correspond to the "real" level of stakeholder participation observed in cases. The "catch" here is that reality does not allow for unfiltered access to its secrets (if it were possible to directly observe reality, it would be meaningless to develop and test a new measurement instrument of stakeholder participation). Therefore, the "true" scores of participation are unknown and we must rely on external variables that are supposed to covary (or not) with this concept to assess whether our measurement is valid. In other words, a *convergent/discriminant validation* strategy must be used to assess the validity of the inferences derived from the PEMI. The variety of terms used to describe different facets of the unified concept of validity or, more accurately, validation procedures can be quite confusing. In this study, convergent validation refers to the process of comparing different measures of the same concept to see if they converge (covary), whereas discriminant validation refers to the process of comparing different measures of different concept to see if they diverge (Adcock & Collier, 2001; McDonald, 2005).

Coders' scores on the PEMI were compared to those of the authors of the articles reporting the evaluation cases on the same instrument (Objective 2, see Figure 1). Contrary to the coders, the authors have direct experience with the evaluation cases (although they might also have consulted the article to establish their scores). This approach to convergent validation is a similar but simpler version of the monotrait-heteromethod approach as initially developed by Campbell and Fiske, 1959 (cited in Trochim, 2006). Our second hypothesis is that there will be a "fair" level of agreement between the two sets of scores (i.e., Cohen's kappa greater or equal to .40).

In addition, authors' scores on the PEMI and the EIS will be compared (Objective 3, see Figure 1). A few words on the EIS are first warranted. This quantitative scale has been developed to measure stakeholder *involvement* — which correspond to the *extent of involvement* dimension in the PEMI —, not *participation*. This instrument has been empirically validated using Messick's unitary concept of validity and the evidence suggests that it "produces appropriate and adequate inferences and interpretations of involvement in multisite evaluations" (Toal, 2009, p. 361). The EIS thus seems to be a good candidate for validating an instrument like the PEMI. On the one hand, a strong positive correlation is expected between the scores for the extent of involvement dimension on the PEMI and the EIS since they purport to measure the same construct (monotrait-heteromethod). On the second hand, a moderate correlation is hypothesized between the overall level of participation as measured by the PEMI and the EIS scores (heterotrait-heteromethod). This expectation is based on both convergent and discriminant rationales. On the one hand, convergence is expected since involvement is one of the three constitutive dimensions of stakeholder participation. On the other hand, we expect only a moderate association between the two constructs because they are different even though they are closely related. Indeed, stakeholder participation is not exhausted by involvement: diversity of participants and control of the evaluation process are necessary dimensions of participation.

Methods

Data and Sample

Intercoder reliability assessment. Data for the assessment of intercoder reliability came from a purposive sample of evaluation cases that were reported in articles published in peer-reviewed journals. Though limited in size, the final sample was sufficiently large (i.e., n = 40) to conduct quantitative analysis, once studies used for coder training and pilot testing were excluded. It must be stressed that the unit of analysis was the (evaluation) case, not the article.[1]

Articles that were already familiar to the authors were perused to assess whether the PE cases they reported respected three selection criteria. First, cases had to contain sufficient information about the evaluation process to allow for scoring the three dimensions of the PEMI (i.e., who participated, when and how). Second, evaluation cases had to be collectively diverse in terms of their theoretical approach used and their level of stakeholder participation (assessed informally). Third, the study had to contain the email address of the authors or this information had to easily be obtained through a web search or colleagues. Although informally applied, other considerations for case selection included diversity in terms of policy domains (education, health, human services, etc.), origins of authors (United States, Canada, Europe, etc.) and journals.

The database created for the purpose of this study contained 48 cases published between 1985 and 2010 (M = 2000). Cases were published in various journals devoted to program evaluation and other disciplines (see Appendix A). The sample covered many policy domains, mainly education, health and human services, but also agriculture, local governance, environment and international development. Based on an informal assessment, cases in the sample displayed varying levels of stakeholder participation: non-participatory or barely participatory cases (n = 4), limited participation (n = 12), moderate or moderately-high participation (n = 26), high or very-high participation (n = 6). This rough classification should not obscure the fact that cases from the same category of participation can actually be very different as to who is involved, how and when. In addition, cases were rather diverse from a theoretical perspective with respect to their evaluation approach and stakeholder involvement. Evaluation cases reported in articles were indeed qualified

by their authors as participatory, collaborative, empowerment, stakeholder-based, utilization-focused, democratic-deliberative, community-based, responsive, etc. Contrary to our initial expectations, contact information proved impossible to obtain for four cases. Since these cases did not respect the third selection criteria, they were used exclusively for training purposes (see below).

Convergent validation I and II. A second source of data came from a survey of one key respondent for each case in the final sample for which author contact information was either available or easy to obtain and which was not part of the first pilot test (n = 39).[2] The survey was conducted online, in English and French, from the 6th of December 2011 to the 9th of January 2012. An invitation email was personally addressed to potential respondents and contained a link to an online questionnaire (one link per case). Emails mentioned the complete reference to the case for which respondents were contacted and the questionnaire's instructions explicitly asked respondents to base their answers on this specific case. A follow-up message was sent to non-respondents one-week after the initial invitation and a second one was sent a week later. A third and last follow-up message was sent three days before the survey closed. More frequent correspondence has occurred with a few respondents who have shown interest in the study or for whom problems were experienced. The timing and titles of follow-up messages capitalized on behavioral theory to increase the response rate, emphasizing the need for help and study salience by highlighting the specific evaluation case for which contacted persons were involved (see Ritter & Sue, 2007).

It was assumed that the first author of each study was most likely to be knowledgeable about the case and willing to participate in the survey. Second authors were contacted only if the first author's email address was unavailable or was inaccurate, or if the first author explicitly refused to participate in the study (n = 6).[3] In the end, 44 invitations to participate in the survey were sent, including non-contacts. A total of 25 fully completed surveys were received.[4] Another completed survey was received but a misunderstanding occurred. The respondent's answers were general (i.e., not related to the specific case for which this person was contacted). While this survey could not be used to check whether the coders' scores aligned with those of the respondent (i.e., Objective 2), it could nevertheless be used to examine the relationship between scores for the PEMI and the EIS (i.e., Objective 3). It was thus considered a "partially usable questionnaire". The response rate, which was calculated according to the American Association for Public Opinion Research's *Standard Definitions* RR1 (AAPOR, 2011, p. 44), was 55.6%.[5] This response rate compares well to those generally obtained through electronic surveys (Couper, 2000; Kwak & Radler, 2002; Millar & Dillman, 2011).

Instruments and Procedures

Intercoder reliability assessment. Applying the PEMI with an adequate level of reliability requires a certain level of familiarity with program evaluation. Coders were therefore recruited from a larger pool of potential coders who had followed a masters-level course on program evaluation and had been studying and/or working as research assistants in a research center on evaluation (i.e., Per*f*Eval). Two research assistants were recruited (one had to be replaced because of unsatisfactory scores) and asked to familiarize themselves with the PEMI by reading Daigneault and Jacob (2009) and an application of it (e.g., Connors & Magilvy, 2011). A codebook detailing coding conventions was then developed and was updated during the coding process (see final version in Appendix B). The overall level of stakeholder participation (PART), which was measured on a five-point ordinal scale, was derived from the minimum or lowest score of the three dimensions. In

turn, the scores of extent of involvement and diversity of participants depended respectively on four dichotomous indicators (four indicators measuring the steps of the evaluation process in which stakeholders were involved and four indicators measuring the types of stakeholders involved).

The research assistants were then instructed to independently code nine "vignettes" (i.e., short hypothetical cases about a paragraph in length) developed for training purposes. The use of vignettes was justified by the limited size of the sample. Scores were compared and reliability between coders was assessed informally as recommended by Lombard, Snyder-Duch and Campanella-Bracken (2002) when conducting training. Coders' scores were also compared to the scores of the first author (i.e., Daigneault) to ensure a fair understanding of the instrument logic. Clarifications and revisions to the codebook were made when necessary. Coders then continued their training on four real, precoded cases in order to fully integrate the operationalization of the concepts (these cases were those for which we were finally unable to obtain author contact information).

Intercoder reliability was formally assessed in a pilot test based on four evaluation cases (n = 4) of varying levels of stakeholder participation. Using a random number generator and alphabetical ordering by author's name, one case was selected in each category of participation (i.e., one case for nonparticipatory or barely participatory cases, one case for limited participation, etc.). The following standards, which are well-established and widely cited, were used to interpret the values of κ and ICC:

> The guidelines developed by Cicchetti and Sparrow (1981) resemble closely those developed by Fleiss (1981) and also represented a simplified version of those introduced earlier by Landis and Koch (1977). The guidelines state that, when the reliability coefficient is below .40, the level of clinical significance is *poor*; when it is between .40 and .59, the level of clinical significance is *fair*; when it is between .60 and .74, the level of clinical significance is *good*; and when it is between .75 and 1.00, the level of clinical significance is *excellent*. (Cicchetti, 1994, p. 286: italics added)

To go on with the coding of the main sample, intercoder reliability scores had to be equal or greater than .40 for this first round of pilot tests. Unfortunately, results were clearly unsatisfactory for the diversity of participants (κ_{DoP} = .00; ICC_{DoP} = -.19) and slightly unsatisfactory for the overall level of participation (ICC_{PART} = .36). By contrast, reliability scores for extent of involvement and control of the evaluation process were excellent (see Table 1). The codebook was revised and a second pilot was conducted on four new cases (n = 4) selected in the same way as for the first pilot. Since the results of the second pilot displayed fair to excellent levels of reliability, a decision was made to pursue with the coding of the main sample.

Once cases used for training and the two pilots were removed, the main sample contained 36 cases. The cases were double-coded independently at a rate of approximately 6 cases at a time (i.e., which took a few days each time), depending on length of coding and availability of coders. Cases for each coding round were purposively selected by the author to reflect the various levels of stakeholder participation, the evaluation's date of publication and the policy domain. Discrepancies were resolved by discussion between the two coders, with occasional guidance by the authors. Coding conventions were revised and added as needed. Coding took an average of 2.5 hours by case per research assistant, including time to solve discrepancies between the coders. To mitigate the limited size of our sample, a decision was made to add the cases of the second pilot

to those of the main sample. This practice is acceptable when the scores obtained during the pilot are adequate (Lombard et al., 2002). The final sample contained 40 cases.

Convergent validation I and II. The questionnaire sent to key respondents of the evaluation cases (i.e., studies' authors) had two sections. The first section focused on the PEMI. Respondents had to first check boxes about which stakeholder type participated at which step of the evaluation process and then assess their level of control on the evaluation process. A five-point index of participation (PART) was derived from their answers and fed back to respondents for reactions. Respondents' opinions were measured on an ordinal scale ("Do not agree at all", "Agree to some extent", "Totally agree" or "I don't know / I don't want to answer") and an open-ended question asked respondents to justify their choice.

The second section relied on a slightly-modified version of the EIS (Toal, 2007, 2009). In the original scale, "Respondents [are] asked to indicate the response that best reflected the extent to which they were involved in 13 different activities (*No* = 1, *Yes, a little* = 2, *Yes, some* = 3, *Yes, extensively* = 4, or ''I don't think this activity took place'')" (Toal, 2009, p. 354). The results of exploratory factor analysis conducted by Toal (2009) supported the removal of two items with low factor loadings. In addition, instructions to respondents were adapted to provide a better fit to the specific aim of this study. Whereas the original scale asked respondents to rate *their* involvement in the process, the version of the EIS used in this study asked about the involvement of nonevaluative stakeholders: "For each question, please choose the response that best describes the extent to which stakeholders other than the evaluator were involved in this evaluation activity". This modification was especially important since most articles in our sample reported cases written from the perspective of the evaluator and since the PEMI purports to measure stakeholder participation (as opposed to the evaluator's involvement).

The original scale is based on the theoretical work of Cousins and Whitmore (1998) and Burke (1998). Contrary to the PEMI, however, this instrument does not purport to measure the three dimensions of Cousins and Whitmore's framework, but only depth of participation (similar to extent of involvement in the PEMI). The EIS is therefore a closely related but imperfect measure of stakeholder participation. Why use the EIS if it does not perfectly measure stakeholder participation? First of all, it is possible to derive theoretical expectations about the relationship between PEMI and EIS scores that could be empirically assessed. As stated earlier, a moderate correlation is expected between the overall level of participation generated by the PEMI (PART) and the level of stakeholder involvement generated by the EIS. A strong correlation is also expected between the latter and the PEMI's extent of involvement score since both indices purport to measure the same concept. Second, the EIS' usefulness stems from the fact that it has been empirically validated and that the evidence of its validity is convincing: "it appears that the majority of the evidence suggests that the Evaluation Involvement Scale produces appropriate and adequate inferences and interpretations of involvement in multisite evaluations" (Toal, 2009, p. 361).

The questionnaire sent to studies' authors was pilot-tested for clarity and readability by two university professors with significant expertise in program evaluation in general and stakeholder participation in particular. One expert tested the English version of the questionnaire while the other tested the French version. Comments from experts were generally positive but also pointed to a few modifications required in the wording of the questions. In addition, correspondence with an early respondent highlighted an ambiguity with respect to what their answers should refer (i.e.,

general practice vs. the specific case for which they were contacted). Whereas the invitation email was clear on this point (i.e., respondents were instructed to answer the questionnaire with respect to the *specific* case for which they were contacted), the questionnaire was modified early in the process to eliminate this ambiguity.

Data analysis

Intercoder reliability assessment. Two quantitative indices were calculated by SPSS (Version 13) to assess intercoder reliability: Cohen's kappa (κ) and intraclass correlation coefficient (ICC). The Cohen's kappa statistic was used to calculate reliability for the eight dichotomous indicators. The kappa was selected over its main rival for nominal data, namely percentage of agreement, because it is a chance-corrected measure of agreement for which results are easily interpretable (Orwin, 1994). Averaged kappa was calculated for the four indicators of extent of involvement and diversity of participants, respectively. While it is usually recommended to assess reliability scores on an item-by-item basis (e.g., Orwin, 1994), the kappa scores for indicators of a same dimension were averaged. It indeed makes sense theoretically as the indicators are supposed to measure the same construct and it facilitates calculation of κ where the number of cases is small (i.e., only four cases for the pilot tests) and where the distribution of scores for individual indicators is skewed (i.e., some columns of the two-by-two matrices were blank).

The ICC was used to assess intercoder reliability for ordinal-scale variables (PEMI's three dimensions and the overall level of participation — PART). Although the use of weighted kappa is generally advocated for ordinal variables and the ICC for continuous ones, ICC is robust enough to be used with ordinal variables in most situations (Norman, 2010). The two tests have indeed been proven to be equivalent under certain conditions by Fleiss and Cohen (1973, as cited by Cicchetti, 1994; Fleiss et al., 2004; Norman, 2010). Furthermore, ICC's flexibility and relationship with Generalizability or G theory are desirable properties that militate in its favour (Norman, 2010; Orwin, 1994). The ICC model selected was the two-way random effects with measures of absolute agreement (i.e., ICC [2,1], see Shrout & Fleiss, 1979).

Convergent validation I. Cohen's kappa and ICC statistics were also used to examine the extent to which the scores achieved by the two independent coders, once discrepancies resolved, aligned with those of a key respondent for each case (where the key respondent is an author of an article in which the cases were reported).

Convergent validation II. Spearman's rank order correlation coefficient (r_s) was used to examine the convergence between the scores achieved by the authors on the PEMI and on the EIS. Whereas Spearman's test is less statistically powerful than Pearson's correlation coefficient, it is a robust nonparametric test appropriate for ordinal scales that makes few assumptions about the distribution of data. The following standards were used to interpret the results (whether negative or positive): .00 to .20 = very weak correlation; .20 to .40 = weak correlation; .40 to .70 = moderate correlation; .70 to .90 = strong correlation; .90 to 1.00 = very strong correlation (Johnston, 2000).

[INSERT TABLE 1 ABOUT HERE]

Results

Intercoder reliability assessment. Results from the different rounds of coding are presented in Table 1. As it was explained earlier, the final sample (n = 40) was composed of cases from the main sample and the second pilot. As denoted by the kappa statistic and the ICC, intercoder reliability is "good" for diversity of participants and "excellent" for extent of involvement. The ICC score is also "good" for control of the evaluation process. Reliability for the overall level of participation, which is the minimum score of the other dimensions, is "fair". Furthermore, all the results are statistically significant (p = .000) which means that it would be highly improbable that the agreement between coders was due to chance. Thus, the results from this intercoder reliability assessment suggest that the PEMI can be used on evaluation cases reported in the literature to produce reliable scores about a phenomenon that is believed to be stakeholder participation. We now examine whether this belief about the nature of this phenomenon is warranted.

[INSERT TABLE 2 ABOUT HERE]

Convergent validation I. The scores obtained by the authors who fully completed their questionnaires (n = 25) were compared to those of the coders (once discrepancies were resolved). Regarding the dichotomous indicators for diversity of participants and extent of involvement, reliability as measured by Cohen's kappa is "fair" and "poor", respectively (see Table 2). Even though the results for the extent of involvement are not satisfactory (i.e., low kappa and statistically nonsignificant results), it must be noted that they are still better than what would be expected by chance alone (i.e., κ = 0). ICC results for the diversity of participants and extent of involvement dimensions are "poor" and, in the latter case, also fail to attain statistical significance. This result is puzzling since extent of involvement was the dimension for which coders' scores were the most reliable. ICC results for the control of the evaluation process and overall level of stakeholder participation were respectively "good" and "fair" and were both statistically significant. Overall, there is a positive alignment between the coders' and the authors' scores but its magnitude is relatively modest (ranging from poor to good). Overall, these results provide some evidence about the validity of the PEMI, but this evidence is relatively weak.

[INSERT TABLE 3 ABOUT HERE]

Convergent validity II. The authors' scores on the PEMI were compared to their scores on the EIS. As stated earlier, a positive moderate relationship was expected between PART and EIS scores because there are closely related but yet different constructs. In addition, a strong positive relationship was expected between extent of involvement (EoI) and EIS scores since they both purport to measure stakeholder involvement in evaluation. On the one hand, the results support the first hypothesis. The relationship between the overall participation score derived from the PEMI and the involvement score derived from the EIS is one of moderate strength (r_s = .44) and statistically significant (p = .025). On the other hand, the relationship between the two alternative measures of stakeholder involvement is only one of moderate strength (r_s = .52, p = .007). Whereas this result goes in the expected direction, it is clearly below our expectations. An unexpected result is the moderate association (i.e., r_s = .63, p = .001) between the scores for control of the evaluation process and the EIS scores. This will need to be further investigated.

At the aggregate level (i.e., overall level of participation), the results from convergent validation provide strong evidence in favour of the PEMI's validity. The two sets of scores are indeed moderately associated which is congruent with our theoretical expectations. At the dimension level, the validation evidence is weaker since the relationship between extent of involvement and

EIS is not as strong as expected. Yet, the moderate strength of the correlation still constitutes evidence of the validity of the extent of involvement dimension:

> We know for sure that we would hope for a correlation of neither 1.00 nor 0. In the first case, the new test could be considered a veritable clone of the one with which it is being compared. In the second case, the construct validity of the very concept being measured would be called into question. (Cicchetti, 1994, p. 288: see also Adcock and Collier, 2001)

Discussion

This study aimed at examining 1) whether the PEMI could be used by two coders on published evaluation cases to produce reliable scores; 2) whether the coders' conciliated scores aligned with those of a key respondent for each evaluation case, and; 3) whether the scores derived by key respondents on the PEMI and the EIS converged.

[INSERT TABLE 4 ABOUT HERE]

Are inferences derived from the PEMI reliable and valid? It is important to stress that "validity is best thought of as a degree, since no variable completely captures an abstract concept" (McDonald, 2005, p. 939). Similarly, Toal (2009) argued that "validity is not a question of 'yes' or 'no,' but instead a question of 'more' or 'less'" (p. 350). The results from this study were therefore interpreted in terms of the strength of the evidence they bring for (and against) PEMI's validity (see Table 4). On the one hand, the evidence is positive and ranges from moderate to strong in the case of intercoder reliability and convergence between PEMI's and EIS' scores. On the other hand, the strength of the evidence is weaker in the case of the alignment between coders' and authors' scores on the PEMI. A first, natural, explanation for this finding would be that the validity of the PEMI is problematic. We would like to suggest three plausible, alternative explanations to this conclusion. First of all, whereas the research assistants had been trained in the use of the instrument, could rely on numerous coding conventions (see Appendix B) and benefited from useful feedback on their scores, the authors who responded to the survey were "left on their own" when using the PEMI. Second, the different data sources used by the research assistants and the authors (i.e., published articles and direct experience, respectively) might explain the discrepancy between their respective results. Coders had indeed to base their scores on what the articles reported about evaluation cases. Even though sufficient data about each evaluation case was a selection criterion for articles, it cannot be assumed that the articles constitute a perfect representation of cases. Third, memory limitations could have biased the scores of the authors and, as a result, could have brought down the level of agreement between their scores and those of the research assistants. Studies in the sample were published more than 10 years ago on average and some authors expressed concerns about their ability to correctly remember the details of the case. Memory problems were cited as the reason for refusal to participate in the study or abandonment by two authors. While a few respondents initially expressed concerns about their memory as well, they nevertheless seemed to be able to remember the case well enough to fully complete the questionnaires and didn't raise this problem again, whether through the open-ended section of the questionnaire or email.

It is worth stressing that this study, like every study which relies on a purposive sample, incurs risks of selection bias. Indeed, although great care was taken to ensure a certain level of diversity during case selection, the extent to which the findings from this study are generalizable to other evaluation cases and settings remains uncertain. For instance, it is entirely possible that some "mainstream" evaluation model (e.g., participatory evaluation *à la* Cousins; see Cousins & Earl,

1992) are over-represented in the sample simply because they are more likely to report in sufficient details the evaluation process that was followed. Moreover, the fact that only *published* evaluation studies were used may have biased the sample against both nonparticipatory and highly participatory cases. Lastly, while the survey response rate was quite acceptable with respect to public opinion research standards, respondents might nevertheless differ from nonrespondents on significant dimensions, for instance their greater knowledge in or interest for participatory evaluation. However, it is difficult to establish the direction and magnitude of the influence that nonresponse would have in this specific case.

Yet, in the end, the results of this study suggest that the PEMI can produce scores that are both reliable and valid. It must be noted that the statistics used to measure intercoder reliability and convergence between scores, namely Cohen's kappa and ICC, are based on agreement, not consistency. These statistics are thus rather conservative (see e.g., Lombard et al., 2002). Furthermore, a debriefing session with the research assistants revealed that reliability would probably improve if more cases were coded. Some coding conventions were developed relatively late in the coding process and added to the coding book but could unfortunately not be applied to many cases. The debriefing also revealed that testing the PEMI on articles was a tough test for reliability. Indeed — and despite careful selection —, many of the articles reviewed in this study contained incomplete or ambiguous information on participation. The need for interpretation was increased and, in turn, the probability of misunderstandings increased as well. This suggests that reliability would improve if the PEMI was used in a real-world setting. The research assistants finally stated that control of the evaluation process was the most difficult dimension to score as determining a representative score is difficult when there are variations during the process. Moreover, they pointed that the rule used to determine the overall level of participation (PART) can be counterintuitive. They thus suggested that the use of the average score of the dimensions would better reflect the level of stakeholder participation than would the minimum score. This theoretical issue will need to be further investigated (see Daigneault & Jacob, forthcoming).

Notes

1. A total of 36 cases out of 48 (75%) came from single-case articles. Two articles reported 3 cases each (for a total of 6 cases or 12.5%) and three articles reported 2 cases each (for a total of 6 cases or 12.5%). Unless they came from the same article, each of the selected cases was written by different authors to ensure a diverse sample.

2. The first pilot test was based on four cases, not three. However, since the author of one of these cases was also the author of two other cases in the main sample and would therefore be contacted anyway, it was decided to survey this author on all cases (i.e., including the case in the pilot). The rationale for excluding cases from the first pilot was that their unreliable scores could have possibly biased the results. In retrospect, however, these fears were largely unfounded because the scores that were used were those for which discrepancies were resolved.

3. Out of the 39 emails sent, 4 were undeliverable. In two cases, the first author had retired and was not further available for participation in this study. In another case, an alternative valid email address could not be found. The second authors of these three cases were thus contacted. In the case of the last undeliverable email, the correct address of the first author was obtained through a colleague. Two potential respondents also refused to

participate (one explained that the case was too old to be remembered, the other person gave no reason). For one case, a second author was contacted but, for the other, the refusal to participate came too late in the survey process to attempt a meaningful contact with the second author.

4. Two respondents had begun to fill the survey but did not complete it (i.e., break-offs). When offered help to complete the survey, one respondent indicated that he or she could not remember the case well enough while the other cited lack of time as a reason for abandon. No replies were received for a total of 11 invitations to participate. Although it cannot be ascertained that these people indeed received the initial invitation and the reminders (e.g., because of a spam filter), we assume that they had.

5. At 57.8%, the response rate including the partially completed questionnaire (RR2) is slightly higher. For both RR1 and RR2, non-contacts are included in the denominator. If excluded, the response rates would be 61% and 63.4%, respectively.

References

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review, 95*(03), 529-546. doi: doi:10.1017/S0003055401003100

Burke, B. (1998). Evaluating for a change: Reflections on participatory methodology. *New Directions for Evaluation*(80), 43-56.

Carmines, E. G., Woods, J. A., & Kimberly, K.-L. (2005). Validity Assessment *Encyclopedia of Social Measurement* (pp. 933-937). New York: Elsevier.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284-290. doi: 10.1037/1040-3590.6.4.284

Connors, S. C., & Magilvy, J. K. (2011). Assessing vital signs: Applying two participatory evaluation frameworks to the evaluation of a college of nursing. *Evaluation and Program Planning, 34*(2), 79-86.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*(4), 464-494.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan & D. L. Stufflebeam (Eds.), *International Handbook of Educational Evaluation* (Vol. 9, pp. 245-265). Dordrecht: Kluwer Academic.

Cousins, J. B., & Earl, L. M. (1992). The case for participatory evaluation. *Educational Evaluation and Policy Analysis, 14*(4), 397-418.

Cousins, J. B., Elliott, C., Amo, C., Bourgeois, I., Chouinard, J. A., Goh, S. C. (2011). Organizational capacity to do and use evaluation: Results of a pan-Canadian survey of evaluators. *Revue canadienne d'évaluation de programme, 23*(3), 1-35.

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New Directions for Evaluation, Understanding and practicing participatory evaluation.* (80), 5-23.

Cullen, A. (2009). *The politics and consequences of stakeholder participation in international development evaluation.* (Ph.D.), Western Michigan University, United States -- Michigan. Retrieved from http://proquest.umi.com/pqdlink?did=1957706941&Fmt=7&clientId=9268&RQT=309&VName=PQD

Cullen, A., Coryn, C., & Rugh, J. (2011). The politics and consequences of including stakeholders in international development evaluation. *American Journal of Evaluation, 32*(3), 345-361.

Daigneault, P.-M., & Jacob, S. (2009). Toward accurate measurement of participation: Rethinking the conceptualization and operationalization of participatory evaluation. *American Journal of Evaluation, 30*(3), 330-348.

Daigneault, P.-M., & Jacob, S. (forthcoming). Unexpected but Most Welcome: Mixed Methods for the Validation and Revision of the Participatory Evaluation Measurement Instrument. *Journal of Mixed Methods Research.*

DeVellis, R. F. (2005). Inter-rater reliability. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (Vol. 2, pp. 317-322). Amsterdam; London Elsevier Academic Press.

Fleischer, D. N., Christie, C. A., & LaVelle, K. B. (2011). Perceptions of evaluation capacity building in the United States: A descriptive study of American Evaluation Association members. *Revue canadienne d'évaluation de programme, 23*(3), 37-60.

Fleiss, J. L., Levin, B., & Paik, M. C. (2004). The Measurement of Interrater Agreement *Statistical Methods for Rates and Proportions* (pp. 598-626): John Wiley & Sons, Inc.

Goertz, G. (2006). *Social science concepts : A user's guide*. New Jersey: Princeton University Press.

Gregory, A. (2000). Problematizing participation: A critical review of approaches to participation in evaluation theory. *Evaluation, 6*(2), 179-199. doi: 10.1177/13563890022209208

Jacob, S., Ouvrard, L., & Bélanger, J.-F. (2011). Participatory evaluation and process use within a social aid organization for at-risk families and youth. *Evaluation and Program Planning, 34*(2), 113-123.

Johnson, K., Greenseid, L. O., Toal, S. A., King, J. A., Lawrenz, F., & Volkov, B. (2009). Research on evaluation use: A review of the empirical literature from 1986 to 2005. *American Journal of Evaluation, 30*(3), 377-410. doi: 10.1177/1098214009341660

Johnston, I. (2000). I'll give you a definite maybe: An introductory handbook on probability, statistics, and Excel *Section 4: Correlations* Retrieved from http://records.viu.ca/~Johnstoi/maybe/maybe4.htm

King, J. A. (1998). Making sense of participatory evaluation practice. *New Directions for Evaluation*(80), 57-67.

Kwak, N., & Radler, B. (2002). A comparison between mail and web surveys: Response pattern, respondent profile and data quality. *Journal of Official Statistics, 18*(2), 257-273.

Laudon, J. M. D. (2010). *Participatory to the End: Planning and Implementation of a Participatory Evaluation Strategy.* York University, Toronto.

Lombard, M., Snyder-Duch, J., & Campanella-Bracken, C. (2002). Content analysis in mass communication: assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587-604.

Mark, M. M. (2001). Evaluation's Future: Furor, Futile, or Fertile? *American Journal of Evaluation, 22*(3), 457-479.

Mathison, S. (Ed.). (2005). *Encyclopedia of evaluation*. Thousand Oaks, CA: Sage.

McDonald, M. P. (2005). Validity, Data Sources. In K.-L. Kimberly (Ed.), *Encyclopedia of Social Measurement* (pp. 939-948). New York: Elsevier.

Millar, M. M., & Dillman, D. A. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly, 75*(2), 249-269. doi: 10.1093/poq/nfr003

Norman, G. (2010). Likert scales, levels of measurement and the ''laws'' of statistics. *Advance in Health Sciences Education, 15*(5), 625-632.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139-162). New York: Russel Sage Foundation.

Poth, C.-A., & Shulha, L. (2008). Encouraging stakeholder engagement: A case study of evaluator behavior. *Studies in Educational Evaluation, 34*(4), 218-223. doi: 10.1016/j.stueduc.2008.10.006

Ritter, L. A., & Sue, V. M. (2007). Conducting the survey. *New Directions for Evaluation, 2007*(115), 47-50. doi: 10.1002/ev.235

Shea, M. P., & Lewko, J. H. (1995). Use of a stakeholder advisory group to facilitate the utilization of evaluation results. *Revue canadienne d'évaluation de programme, 10*(1), 159-162.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Thayer, C. E., & Fine, A. H. (2001). Evaluation and outcome measurement in the non-profit sector: stakeholder participation. *Evaluation and Program Planning, 24*(1), 103-108.

The American Association for Public Opinion Research. (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys Retrieved from http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156

Toal, S. A. (2007). *The Development and Validation of an Evaluation Involvement Scale for Use in Multi-site Evaluations.* (PhD), University of Minnesota, Minneapolis.

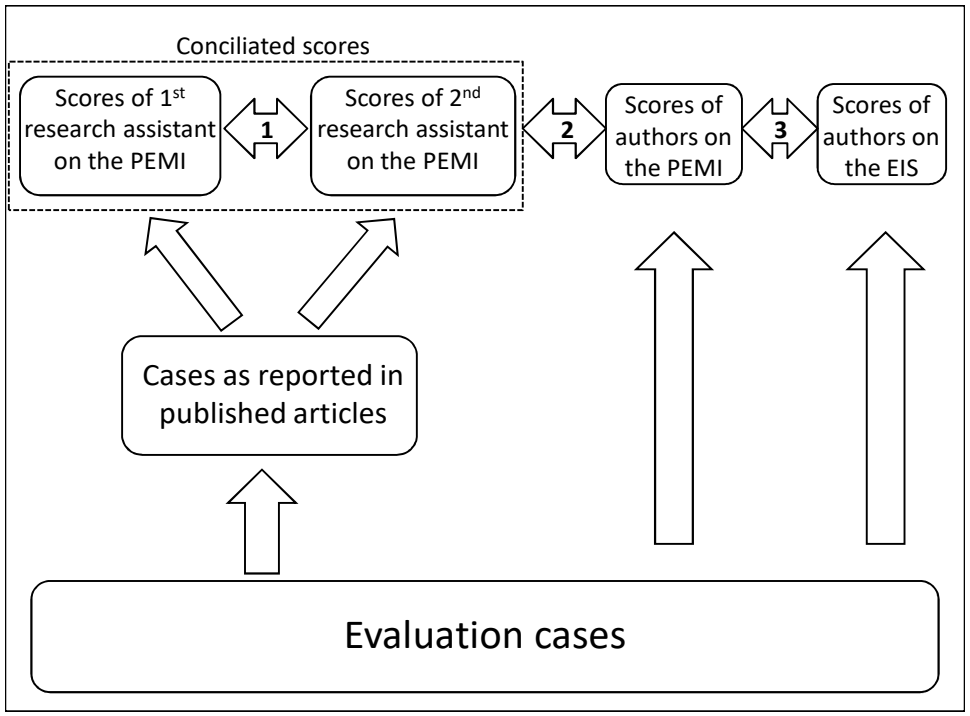
Toal, S. A. (2009). The Validation of the Evaluation Involvement Scale for Use in Multisite Settings. *American Journal of Evaluation, 30*(3), 349-362. doi: 10.1177/1098214009337031

Trochim, W. M. (2006). The multitrait-multimethod matrix. *The research methods knowledge base*. Retrieved from http://www.socialresearchmethods.net/kb/

Whitmore, E. (1998). Editor's notes. *New Directions for Evaluation*(80), 1-3.

Figures

Figure 1: Schematic Representation of Validation Objectives



NOTE: PEMI = Participatory Evaluation Measurement Instrument; EIS = Evaluation Involvement Scale.

Tables

Table 1: Results of the Intercoder Reliability Assessment (Cohen's kappa and intraclass correlation coefficient)

Coding rounds	*n*	κ_{DoP}	κ_{EoI}	ICC_{DoP}	ICC_{EoI}	ICC_{CoEP}	ICC_{PART}
Training round 1 (vignettes)	9	—	—	—	—	—	—
Training round 2	4	—	—	—	—	—	—
Pilot test round 1	4	.00 (1.00)	.82 (.001)***	-.19 (.693)	.96 (.005)**	.80 (.066)	.36 (.260)
Pilot test round 2	4	.88 (.000)***	1.00 (.000)***	.93 (.011)*	1.00 (n/c)	.50 (.236)	.89 (.022)*
Main sample	36	.53 (.000)***	.80 (.000)***	.68 (.000)***	.87 (.000)***	.64 (.000)***	.45 (.003)**
Final sample (i.e., main sample + pilot test round 2)	40	.64 (.000)***	.82 (.000)***	.71 (.000)***	.89 (.000)***	.63(.000)***	.53 (.000)***

NOTE: DoP= Diversity of participants; EoI = Extent of involvement; CoEP = Control of the evaluation process; PART = Level of stakeholder participation; n/c = value could not be calculated.

*p = significant at .05 level; **p = significant at .01 level; ***p = significant at .001 level

Table 2: Alignment between Key Respondents' Scores and Conciliated Scores (Cohen's kappa and intraclass correlation coefficient)

Sample	*n*	κ_{DoP}	κ_{EoI}	ICC_{DoP}	ICC_{EoI}	ICC_{CoEP}	ICC_{PART}
Overlapping sample	25	.47 (.000)***	.14 (153)	.31 (.05)*	.23 (.12)	.62(.000)***	.40 (.024)*

NOTE: DoP= Diversity of participants; EoI = Extent of involvement; CoEP = Control of the evaluation process; PART = Level of stakeholder participation.

*p = significant at .05 level; **p = significant at .01 level; ***p = significant at .001 level

Table 3: Correlation (Spearman's rho) between Scores Derived from the PEMI and the EIS

		PEMI			
	n	*DoP*	*EoI*	*CoEP*	*PART*
EIS	26	.15(.455)	.52(.007)**	.63(.001)***	.44(.025)*

NOTE: EIS = Evaluation Involvement Scale; DoP = Diversity of participants; EoI = Extent of involvement; CoEP = Control of the evaluation process; PART = Level of stakeholder participation.
*p = significant at .05 level; **p = significant at .01 level; ***p = significant at .001 level

Table 4: Strength of Validity Evidence

Validation Procedures	*Findings*	*Statistical Significance*	*Validity Evidence*
Intercoder reliability assessment	Agreement is fair to excellent	***	Strong
Convergent validation I	Agreement is poor to good	N.S. to ***	Weak
Convergent validation II: H1	Moderate correlation (as expected)	*	Strong
Convergent validation II: H2	Moderate correlation (strong expected)	**	Moderate

NOTE: H1 = Hypothesis 1; H2 = Hypothesis 2.
N.S. = non significant; *p = significant at .05 level; **p = significant at .01 level; ***p = significant at .001 level

Appendix A: Reference list of the total sample

Abma, T. A., Nierse, C. J., & Widdershoven, G. A. M. (2009). Patients as Partners in Responsive Research: Methodological Notions for Collaborations in Mixed Research Teams. *Qualitative Health Research, 19*(3), 401-415. doi: 10.1177/1049732309331869 (2 cases).

Alpert, B., & Bechar, S. (2007). Collaborative Evaluation Research: A Case Study of Teachers' and Academic Researchers' Teamwork in a Secondary School. *Studies in Educational Evaluation, 33*(3-4), 229-257.

Andrews, A. B., Motes, P. S., Floyd, A. G., Flerx, V. C., & Lopez-De Fede, A. (2005). Building Evaluation Capacity in Community-Based Organizations: Reflections of an Empowerment Evaluation Team. *Journal of Community Practice, 13*(4), 85-104. doi: 10.1300/J125v13n04_06

Ayers, T. D. (1987). Stakeholders as Partners in Evaluation: A Stakeholder-Collaborative Approach. *Evaluation and Program Planning, 10*(3), 263-271

Berner, M., & Bronson, M. (2005). A case study of program evaluation in local government: Building consensus through collaboration. *Public Performance & Management Review, 28*(3), 309-325.

Brandon, P. R. (1999). Involving program stakeholders in reviews of evaluators' recommendations for program revisions. *Evaluation and Program Planning, 22*(3), 363-372.

Brisson, D. (2007). Collaborative Evaluation in Community Change Initiative: Dilemmas of Control Over Technical Decision Making. *Canadian Journal of Program Evaluation, 22*(2), 21-39.

Campbell, R., Dorey, H., Naegeli, M., Grubstein, L. K., Bennett, K. K., Bonter, F. (2004). An empowerment evaluation model for sexual assault programs: Empirical evidence of effectiveness. *American Journal of Community Psychology, 34*(3-4), 251-262.

Christie, C. A., Ross, R. M., & Klein, B. M. (2004). Moving toward collaboration by creating a participatory internal-external evaluation team: A case study. *Studies In Educational Evaluation, 30*(2), 125-134.

Cooper, D., & Hewitt, W. E. (1989). Working together on an evaluation: a case study. *Canadian Journal of Program Evaluation, 4*(1), 1-10.

Cousins, J. B. (1996). Consequences of Researcher Involvement in Participatory Evaluation. *Studies in Educational Evaluation, 22*(1), 3-27 (3 cases).

Curran, V., Solberg, S., LeFort, S., Fleet, L., & Hollett, A. (2008). A responsive evaluation of an aboriginal nursing education access program. *Nurse Educator, 33*(1), 13-17.

Dawson, J. A., & D'Amico, J. J. (1985). Involving Program Staff in Evaluation Studies: A Strategy for Increasing Information Use and Enriching the Data Base. *Evaluation Review, 9*(2), 173-188.

Dryden, E., Hyde, J., Livny, A., & Tula, M. (2010). Phoenix Rising: Use of a participatory approach to evaluate a federally funded HIV, Hepatitis and substance abuse prevention program. *Evaluation and Program Planning, 33*(4), 386-393.

Dubois-Arber, F., Jeannin, A., & Spencer, B. (1999). Long Term Global Evaluation of a National AIDS Prevention Strategy: The Case of Switzerland. *Aids, 13*(18), 2571-2582.

Greene, J. C. (1987). Stakeholder Participation in Evaluation Design: Is It Worth the Effort? *Evaluation and Program Planning, 10*(4), 379-394 (2 cases).

Hofstetter, C. H. (2004). Unpacking the Evaluation Process: A Study of Transitional Bilingual Education. *Studies in Educational Evaluation, 30*(4), 325-336.

Howe, K. R., & Ashcraft, C. (2005). Deliberative democratic evaluation: Successes and limitations of an evaluation of school choice. *Teachers College Record, 107*(10), 2275-2298.

Keiny, S., & Dreyfus, A. (1993). School self-evaluation as a reflective dialogue between researchers and practitioners. *Studies in Educational Evaluation, 19*(3), 281-295.

King, J. A., & Ehlert, J. C. (2008). What we learned from three evaluations that involved stakeholders. *Studies in Educational Evaluation, 34*(4), 194-200. doi: 10.1016/j.stueduc.2008.10.003 (3 cases).

Llosa, L., & Slayton, J. (2009). Using Program Evaluation to Inform and Improve the Education of Young English Language Learners in US Schools. *Language Teaching Research, 13*(1), 35-54.

Mayo, J. K., Green, C. B., & Vargas, M. E. (1985). Radio Santa Maria: a case study of participatory evaluation. *Development Communication Report* (48), 1.

McAllister, C. L., Green, B. L., Terry, M. A., Herman, V., & Mulvey, L. (2003). Parents, practitioners, and researchers: Community-based participatory research with early head start. *American Journal of Public Health, 93*(10), 1672.

McKenzie, B. (1997). Developing First Nations Child Welfare Standards: Using Evaluation Research within a Participatory Framework. *Canadian Journal of Program Evaluation*, *12*(1), 133-148.

Miller, R. W. (1987). Using Evaluation to Support the Program Advisory Function: A Case Study of Evaluator-Program Advisory Committee Collaboration. *Evaluation and Program Planning, 10*(3), 281-288.

Papineau, D., & Kiely, M. C. (1996). Participatory Evaluation in a Community Organization: Fostering Stakeholder Empowerment and Utilization. *Evaluation and Program Planning, 19*(1), 79-93.

Puma, J., Bennett, L., Cutforth, N., Tombari, C., & Stein, P. (2009). Case Study of a Community-Based Participatory Evaluation Research (CBPER) Project: Reflections on Promising Practices and Shortcomings. *Michigan Journal of Community Service Learning, 15*(2 (Spring)), 34-47.

Quintanilla, G., & Packard, T. (2002). A participatory evaluation of an inner-city science enrichment program. *Evaluation and Program Planning, 25*(1), 15-22.

Rebolloso, E., Fernandez-Ramirez, B., & Canton, P. (2005). The Influence of Evaluation on Changing Management Systems in Educational Institutions. *Evaluation, 11*(4), 463-479 (2 cases).

Ridde, V. (2003). The Experience of a Pluralist Approach in a Country at War: Afghanistan. *Canadian Journal of Program Evaluation, 18*(1), 25-48.

Rockwell, S. K., Dickey, E. C., & Jasa, P. J. (1990). The Personal Factor in Evaluation Use: A Case Study of a Steering Committee's Use of a Conservation Tillage Survey. *Evaluation and Program Planning, 13*(4), 389-394.

Ryan, K. E., & et al. (1996). Progress and Accountability in Family Literacy: Lessons from a Collaborative Approach. *Evaluation and Program Planning, 19*(3), 263-272.

Shea, M. P., & Lewko, J. H. (1995). Use of a stakeholder advisory group to facilitate the utilization of evaluation results. *Canadian Journal of Program Evaluation, 10*(1), 159-162.

Somers, C. (2005). JEvaluation of the Wonders in Nature-Wonders in Neighborhoods Conservation Education Program: Stakeholders Gone Wild! *New Directions for Evaluation*, (108), 29-46. doi: 10.1002/ev.169

Spooner, C., Flaxman, S., & Murray, C. (2008). Participatory research in challenging circumstances: Lessons with a rural Aboriginal program. *Evaluation Journal of Australasia, 8*(2), 28-34.

Suarez-Balcazar, Y., Orellana-Damacela, L., Portillo, N., Sharma, A., & Lanum, M. (2003). Implementing an Outcomes Model in the Participatory Evaluation of Community Initiatives. *Journal of Prevention & Intervention in the Community, 26*(2), 5-20.

Sullins, C. D. (2003). Adapting the empowerment evaluation model: A mental health drop-in center case example. *American Journal of Evaluation, 24*(3), 387-398.

Torres, R. T., Stone, S. P., Butkus, D. L., Hook, B. B., Casey, J., & Arens, S. A. (2000). Dialogue and Reflection in a Collaborative Evaluation: Stakeholder and Evaluator Voices. *New Directions for Evaluation*, (85), 27-38.

Uhl, G., Robinson, B., Westover, B., Bockting, W., & Cherry-Porter, T. (2004). Involving the community in HIV prevention program evaluation. *Health Promotion Practice, 5*(3), 289-296.

Williams, A. M. (2010). Evaluating Canada's Compassionate Care Benefit using a utilization-focused evaluation framework: Successful strategies and prerequisite conditions. *Evaluation and Program Planning, 33*(2), 91-97.

Wye, C. G. (1989). Increasing client involvement in evaluation: A team approach. *New Directions for Program Evaluation, 1989*(41), 35-48.

Appendix B: Coding Conventions (Final Version)

General Instructions

G – 1. Scores for all dimensions rely on the concept of *meaningful* or *significant* involvement. In order to be considered meaningfully or significantly involved, stakeholders must have a significant role in the actual design and/or conduct of the evaluation (including the diffusion of findings). Therefore, being a data source or a passive observer is not sufficient for a stakeholder type to be considered meaningfully or significantly involved (thereby implying a score of .00 on all three dimensions).

G – 2. The score for the overall level of participation is always the *minimum* (i.e., the lowest score) of the three dimensions. For instance, scores of .25, .50 and .75 on extent of involvement, diversity of participants and control for a given case implies that the general level of stakeholder participation is coded at .25.

Extent of Involvement (EoI)

EoI - 1. When certain steps of the evaluation process are not explicitly mentioned (e.g., diffusion of evaluation results), coders must infer from the information available for other steps.

Diversity of Participants (DoP)

DoP – 1. The first and second indicators have been respectively relabelled "Policy-makers, decision-makers and managers" and "People directly responsible for "program delivery". Managerial staff of all hierarchical levels goes into Indicator 1. Indicator 2 is restricted to professionals and front-line civil servants responsible for direct program delivery.

DoP – 2. Indicator 4 is restricted to organized stakeholders and citizens who defend interests that are larger in scope than those of the program and/or organization evaluated. Suppose a gifted education program is implemented by a school district. Teachers and their union's representatives from the district are taken into account by Indicator 2 whereas participants from the American Federation of Teachers would be taken into account by Indicator 4.

DoP – 3. Professional evaluators and their support staff (secretary, research assistant, etc.), whatever their organizational affiliations, are never taken into account for the coding of diversity. Thus, academics hired to act as external evaluators would not affect the coding of diversity whereas academics acting in an expert role on an advisory committee would.

Control of the evaluation process

CoEP – 1. Coding this dimension does not rely on any indicator per se. It is based on a subjective assessment of how control is relatively distributed between the evaluator on the one hand and the non evaluative stakeholders on the other ***only for steps*** *in which stakeholders were involved*. It is thus logically possible to have a high score for control even though stakeholders were only involved in one step.

CoEP – 2. When control varies during the evaluation process, scores should be *representative* of the share of control stakeholders possess during all steps in which they were involved. Scores can be averaged across steps but, contrary to conventional Likert scales, scores should fall on (or be rounded to) established points such as .25 or .50 (not .378 or .516).

CoEP – 3. Scores must be based on what authors affirm about the distribution of control and on how they qualify it. For instances, terms and expressions such as "shared", "equally", "collaboratively", "jointly", "worked together" and "mutual decisions" can be useful indicators of a

shared control between evaluators and stakeholders (.50). Similarly, “consulted”, “inspired from”, “give voice or input”, “facilitated” are often indicative of an unequal distribution of control (.25 or .75). Finally, “retained control”, “totally or completely managed by”, “fully responsible” and “unilaterally decided by” indicate a situation where control should be rated .00 or 1.00. Despite the fact that those terms and expressions are useful to assess control, they should not be applied mechanically; coding is a matter of informed judgement.

CoEP – 4. When there is no indication as to how control is distributed, one must assume that it is shared equally between the evaluator and stakeholders (i.e., a score of .50). Moreover, only the effective distribution of control is relevant, not the number of evaluators and stakeholders involved in an evaluation. The control can totally rest in the hands of the evaluator even if an evaluation team counts one evaluator and ten participants (and vice-versa).

CoEP – 5. A score of 1.00 on this dimension denotes that nonevaluative stakeholders have full control of the evaluation process, meaning that either there is no professional evaluator involved in this evaluation (stakeholders act as full evaluators) or the professional evaluator’s role is confined to that of a technical executant or henchman for stakeholders.