PIERRE-MARC DAIGNEAULT

LA PARTICIPATION À L'ÉVALUATION : DU CONCEPT À LA MESURE

Thèse présentée
à la Faculté des études supérieures et postdoctorales de l'Université Laval
dans le cadre du programme de doctorat en science politique
pour l'obtention du grade de *Philosophiae doctor* (Ph.D.)

DÉPARTEMENT DE SCIENCE POLITIQUE
FACULTÉ DES SCIENCES SOCIALES
UNIVERSITÉ LAVAL
QUÉBEC

2012

Résumé

La popularité croissante des approches participatives représente une tendance lourde dans le champ de l'évaluation des politiques. La prolifération des définitions et des termes utilisés pour désigner la participation génère cependant beaucoup de confusion chez les chercheurs et praticiens du domaine. Il n'existe en outre aucun instrument de mesure adéquat de la participation, ce qui freine l'avancement des connaissances. Trois questions de recherche structurent cette thèse : 1) Qu'est-ce que la participation à l'évaluation?; 2) Comment traduire ce concept en un instrument de mesure opérationnel?; et 3) Est-ce que cet instrument mesure la participation de manière fidèle et valide?

Une conceptualisation cohérente de l'évaluation participative s'inscrivant dans la foulée des travaux de Cousins et Whitmore (1998) est d'abord proposée. Cette conceptualisation, fondée sur la logique des conditions nécessaires et suffisantes, est opérationnalisée en un instrument de mesure de la participation.

L'instrument (*Participatory Evaluation Measurement Instrument* – PEMI) fait ensuite l'objet d'une validation empirique à partir d'un échantillon de 40 cas d'évaluation tirés de la littérature et d'un sondage auprès de leurs auteurs. Trois éléments sont appréciés quantitativement : 1) la fidélité intercodeur; 2) la convergence des scores des codeurs et des auteurs sur le PEMI; et 3) la convergence des scores des auteurs sur le PEMI et un instrument de mesure alternatif. De manière générale, cette étude suggère que le PEMI génère des scores dont la fidélité et la validité sont d'un niveau acceptable.

Troisièmement, une étude de validation du PEMI combinant méthodes qualitatives et quantitatives est présentée. Le recours aux méthodes mixtes a généré un cycle inattendu – mais bénéfique – de révision de l'instrument et de validation quantitative supplémentaire. Les résultats de validation suggèrent que la version révisée du PEMI, désormais fondée sur une structure conceptuelle hybride, est plus en phase avec l'opinion des répondants quant au niveau de participation des cas d'évaluation. La valeur ajoutée des méthodes mixtes à des fins de validation est également discutée.

Une réflexion sur le potentiel scientifique de l'instrument de mesure, en particulier dans le cadre de recherches empiriques sur la relation entre participation et utilisation de l'évaluation, vient conclure cette thèse.

Abstract

The growing popularity of participatory approaches represents an important trend in the field of program evaluation. The proliferation of definitions and terms used to designate stakeholder participation, however, generates a lot of confusion among researchers and practitioners. Moreover, the dearth of adequate instruments to measure participation hinders knowledge accumulation. This dissertation is structured around three research questions: 1) What is stakeholder participation in evaluation? 2) How is this concept translated into an operational measurement instrument? and 3) Does this instrument allow for the reliable and valid measurement of stakeholder participation?

A systematic and coherent conceptualization of participatory evaluation is first proposed based on the work of Cousins and Whitmore (1998). This conceptualization, which is based on the logic of necessary and sufficient conditions, is operationalized in a measurement instrument.

The instrument (*Participatory Evaluation Measurement Instrument* – PEMI) is then empirically validated using a sample of 40 evaluation cases from the literature and a survey of their authors. Three elements are quantitatively assessed: 1) intercoder reliability; 2) convergence between coders' and authors' scores on the PEMI; and 3) convergence between authors' scores on the PEMI and an alternative measurement instrument. Considered globally, this study suggests that the PEMI can generate reliable and valid scores.

Finally, a validation study combining qualitative and quantitative methods is presented. The use of mixed methods has generated an unexpected but most welcome cycle of instrument revision and further quantitative validation. The validation results suggest that the revised version of the PEMI, now based on a hybrid conceptual structure, is more in line with our respondents' opinions with respect to the level of stakeholder participation in their particular evaluation case. The added value of mixed methods for validation purposes is also discussed using counterfactual reasoning.

Reflections on the scientific and practical potential of the measurement instrument, on the relationship between stakeholder participation and evaluation use in particular, conclude this dissertation.

Avant-Propos

La présente thèse adopte le format « avec insertion d'articles ». Trois manuscrits d'articles forment le corps de cette thèse. Il faut préciser que j'ai obtenu l'autorisation de mes coauteurs pour insérer ceux-ci dans ma thèse. Les modifications ayant été apportées aux articles concernent principalement leur mise en page selon le format exigé par les instances universitaires et la cohérence générale de l'ouvrage, soit l'ajout d'un résumé français pour les articles de langue anglaise, l'adaptation de la numérotation des tableaux et figures pour qu'elle suive celle de la thèse et l'uniformisation du style des références.

Le premier article, « Toward Accurate Measurement of Participation: Rethinking the Conceptualization and Operationalization of Participatory Evaluation », a été inséré au deuxième chapitre. La version définitive de l'article a été publiée en 2009 dans l'*American Journal of Evaluation*, vol. 30, n° 3 (pp. 330-348; http://aje.sagepub.com/content/30/3/330), par Sage Publications Ltd/Sage Publications, Inc., tous droits réservés ©.[1] J'en suis l'auteur principal alors que M. Steve Jacob, professeur au Département de science politique de l'Université Laval et superviseur de ma thèse, en est le coauteur. Mon rôle dans la préparation du manuscrit a été significatif et extensif, comprenant la réalisation de la revue de littérature; la définition de la problématique générale; la conception générale du cadre conceptuel, de l'instrument de mesure, des tableaux et des figures; ainsi que la rédaction. M. Jacob a quant à lui complété la revue de littérature par des références pertinentes; a contribué au choix définitif des indicateurs du cadre conceptuel; a lu, commenté et bonifié plusieurs versions du manuscrit.

Le second manuscript, inséré au chapitre trois, est intitulé « Measuring Stakeholder Participation in Evaluation: An Empirical Validation of the Participatory Evaluation Measurement Instrument (PEMI) ». Le texte a été soumis le 2 février 2012 à *Evaluation Review: A Journal of Applied Social Research* et a été accepté pour publication le 20 juillet 2012 (l'article est présentement sous presse). Je suis l'auteur principal de ce manuscrit rédigé en collaboration avec M. Jacob et M. Joël Tremblay, professeur au Département de psychoéducation de l'Université du Québec à Trois-Rivières. J'ai réalisé la revue de

[1] Une version française et mise à jour de cet article sera également publiée dans la seconde édition d'*Approches et pratiques en évaluation de programme*, sous la direction de V. Ridde et C. Dagenais.

littérature, la problématisation, la conception générale du devis et de l'instrument de mesure, la formation et l'entraînement des codeurs, la collecte et l'analyse des données, l'interprétation des résultats ainsi que la rédaction générale. Mes coauteurs ont apporté une contribution concentrée lors des phases de conception du devis et de l'instrument de mesure et de la rédaction. Ils ont également apporté une aide ponctuelle lors du processus de collecte et d'analyse des données.

Le troisième manuscrit d'article, « Unexpected but Most Welcome: Mixed Methods for the Validation and Revision of the Participatory Evaluation Measurement Instrument », a été inséré au chapitre quatre. Il a été soumis le 6 février 2012 au *Journal of Mixed Methods Research* et a été accepté pour publication conditionnellement à des révisions mineures le 25 juillet 2012 (les révisions demandées ont été soumises le 3 août 2012). Je suis le principal auteur de ce manuscrit qui a été rédigé avec M. Jacob. Mon rôle dans la préparation du manuscrit a porté sur les tâches suivantes : problématisation, revue de littérature, collecte et analyse des données, interprétation des résultats, révision de l'instrument et rédaction. La contribution de M. Jacob a consisté en une vérification systématique du codage des données qualitatives et de leur l'interprétation. Celui-ci a également participé à la révision de l'instrument de mesure et à la mise en forme finale du manuscrit (structure, style, etc.).

Remerciements

Si la rédaction de la thèse est d'abord et avant tout une entreprise individuelle, plusieurs personnes y contribuent de manière directe et indirecte. Ainsi, derrière chaque doctorant qui soutient sa thèse se trouvent des mentors, professionnels, collègues, membres de la famille et amis qui l'ont soutenu et encouragé dans cette entreprise de longue haleine. Je souhaite ici les remercier.

Je suis d'abord et avant tout reconnaissant à Steve Jacob, mon directeur de recherche à la maîtrise et au doctorat, pour m'avoir initié à l'évaluation des politiques et ainsi permis de trouver « ma voie » au plan professionnel. Steve m'a prodigué moult conseils, encouragements et ressources (financières, logistiques, etc.) qui m'ont permis de développer mon plein potentiel. Je tiens à souligner plusieurs de ses qualités qui font de lui un excellent directeur de recherche : sa très grande disponibilité, son attitude posée et réfléchie (très efficace pour rassurer l'étudiant anxieux que je suis) et un respect qui l'amène à accompagner, guider les étudiants dans leurs recherches plutôt qu'à leur imposer « son » programme de recherche et « ses » perspectives et méthodes. Steve est à bien des égards un modèle dont je me suis inspiré tout au long de mes études et dont je continuerai à m'inspirer.

Je souhaite ensuite remercier Mathieu Ouimet pour sa contribution scientifique et matérielle à titre de codirecteur de mes recherches. Je lui suis particulièrement reconnaissant de m'avoir initié à certains enjeux novateurs en science politique (politiques fondées sur des données probantes, hiérarchie des devis basée sur la validité de leurs inférences, méthodes de revue systématique et de synthèse, etc.) et de m'avoir fait profiter de son expertise sur le transfert des connaissances. Je remercie également Mathieu de m'avoir poussé hors de ma « zone de confort » en remettant en question certaines de mes positions épistémologiques et méthodologiques. En science comme dans d'autres domaines, le progrès découle souvent de remises en question et de discussions animées.

Outre mon directeur et mon codirecteur de recherche, plusieurs personnes ont apporté leur contribution à cette thèse. Je souhaite d'abord remercier Kristen Leppington, pour la révision linguistique des sections anglaises de la thèse. Concernant l'article inséré au

deuxième chapitre, je souhaite remercier Jean-François Bélanger, Martin Cossette, Brad Cousins, Nouhoun Diallo, Gary Goertz, Jennifer Greene, Jean King, Mbaïrewaye Mbaï-Hadji, Laurence Ouvrard ainsi que trois évaluateurs anonymes de l'*American Journal of Evaluation* pour leurs commentaires pertinents.

Je remercie Joël Tremblay, coauteur de l'article inséré au chapitre trois, pour sa contribution. Cet article, ainsi que celui inséré au quatrième chapitre, a bénéficié de l'apport de Marvin C. Alkin et de Marie Gervais pour le prétest des versions anglaise et française du questionnaire. Merci à Stacie Toal pour ses précisions sur l'instrument de mesure qu'elle a développé ainsi qu'à David Collier et Nathalie Loye pour leur aide concernant les différents types de validation. Je tiens également à remercier Geoffroy Desautels et Marylie Roger pour leur codage consciencieux des cas d'évaluation, ainsi que Jean-Simon Couture, du Centre de services APTI de la Faculté des sciences sociales, pour le développement du questionnaire électronique. Enfin, ces articles n'auraient probablement pas vu le jour sans la générosité et l'intérêt des répondants.

Au-delà de la thèse proprement dite, plusieurs personnes ont marqué mon évolution aux plans scientifique et professionnel. Je tiens d'abord à remercier Louis Bélanger, François Blais, Jean Crête et Louis Imbeau, professeurs au Département de science politique, pour avoir su me communiquer leur sens de la rigueur et leur intérêt pour les questions épistémologiques et méthodologiques.

Je dois également remercier mes collègues avec qui j'ai pu échanger de manière constructive sur certains aspects de mes recherches mais aussi de manière plus générale sur la science politique, le cheminement doctoral et le métier de politologue. Je remercie tout spécialement Pierre-Olivier Bédard, Jean-François Bélanger, Maude Benoît, Marie-Hélène Cantin, François Chouinard, Benoît Collette, Martin Cossette, Marylise Cournoyer, Jérôme Couture, Nouhoun Diallo, Isabelle Dufour, Kim Fontaine-Skronski, Tania Gosselin, David Houle, Félix Kuntzsch, Vincent Laborderie, Mbaïrewaye Mbaï-Hadji, Alexandre Morin-Chassé, Laurence Ouvrard, Marie-Ève Proulx, Rima Slaibi, Ronan Teyssier et Stéphanie Yates.

Je suis ensuite reconnaissant au Conseil de recherches en sciences humaines du Canada (CRSH) pour le soutien financier attribué par l'intermédiaire de son programme de Bourses d'études supérieures du Canada. Je remercie également l'École de la fonction publique du Canada et le Fonds québécois de la recherche sur la société et la culture (FQRSC) pour leur soutien financier dans la réalisation de recherches liées à cette thèse.

Enfin, sur le plan personnel, mes parents, les membres de ma famille, mes amis et ma conjointe, Mélanie Turmel-Huot, obtiennent toute ma reconnaissance pour leur soutien indéfectible. Merci!

Je dédie cette thèse à la nature collective de l'entreprise scientifique.

Un concept net doit porter la trace de tout ce qu'on a refusé d'y incorporer. D'une manière générale, à l'origine d'une conceptualisation, il faut effacer les teintes vagues et flottantes d'un phénomène pour en dessiner les traits constants.

Gaston Bachelard ([1951], 2001). *La dialectique de la durée.* 3e éd., Paris : PUF, p. 15.

*Comme le coureur de fond, le doctorant doit tenir la distance. Mais à la différence du marathonien, personne n'a tracé pour lui de ligne d'arrivée. Le plus dur dans la thèse, c'est de finir***.**

Éloïse Lhérété (2011). La solitude du thésard de fond. *Sciences humaines, 10*(230), p. 10.

Table des matières

Liste des tableaux

Liste des figures

Liste des encadrés

Liste des abréviations et des sigles

AAPOR	The American Association for Public Opinion Research
CoEP	Contrôle du processus évaluatif (*Control of the evaluation process*)
DoP	Diversité des participants (*Diversity of participants*)
EIS	*Evaluation Involvement Scale*
EoI	Étendue de l'implication (*Extent of involvement*)
ÉP	Évaluation participative
H	Hypothèse
ICC	Coefficient de corrélation intraclasse (*Intraclass correlation coefficient*)
IDCV	*Instrument Development and Construct Validation*
PART	Niveau global de participation
PE	*Participatory evaluation*
PEMI	*Participatory Evaluation Measurement Instrument*
P-PE	*Practical participatory evaluation*
RR	Taux de réponse (*Response rate*)
s.a.	Sans auteur
s.p.	Sans page
T-PE	*Transformative participatory evaluation*

Liste des symboles

©	Tous droits réservés (*Copyright*)
κ	Kappa de Cohen
M	Moyenne
Mdn	Médiane
Max	Maximum
Min	Minimum
n	Nombre de cas de l'échantillon
p	Niveau de signification statistique
r_s	Rho de Spearman
U	Statistique U de Mann-Whitney
z	Valeur z (approximation de la distribution normale)

Introduction

Au sens général, évaluer consiste à porter un jugement de valeur sur un objet : « Evaluation is the process of determining the merit, worth, and value of things » (Scriven, 1991, p. 1: cité dans Vedung, 1997, p. 2). On peut ainsi évaluer une œuvre d'art, un plat gastronomique, les aptitudes professionnelles d'un employé ou encore la qualité des derniers modèles automobiles.

L'évaluation peut également être appliquée à ce que les autorités publiques « font », c'est-à-dire aux politiques, programmes, projets et mesures qu'elles adoptent et mettent en œuvre pour pallier certains problèmes publics (p. ex., lutter contre l'itinérance, favoriser le développement économique des régions ressources, etc.). Cette pratique, que l'on désigne par les termes *évaluation de politique* ou *évaluation de programme*, vise à déterminer si les actions entreprises par les autorités publiques sont pertinentes, plausibles, efficaces, efficientes, économes, équitables et durables, notamment. Au contraire de la pratique de l'évaluation entendue au sens commun, l'évaluation des politiques et programmes publics repose cependant sur une ambition scientifique (Rossi, Lipsey, & Freeman, 2004). Il s'agit certes de juger, mais également de mesurer (Jacob, 2004). En effet, le jugement porté sur l'action publique ne doit pas être arbitraire ou encore motivé par des préférences personnelles ou partisanes. L'évaluation doit au contraire être fondée sur le recours à des données de qualité et à une méthodologie transparente, systématique et reproductible.

Cette forme de rétroaction sur l'action publique est indispensable dans un contexte où les autorités publiques ne disposent que de moyens limités pour apprécier la valeur de leurs interventions. Tout d'abord, contrairement à la plupart des entreprises privées, les autorités publiques, qui sont souvent en situation de monopole, ne sont en règle générale que peu ou pas soumises aux lois du marché. Les concepts de profits et de parts de marché qui servent d'étalon de mesure dans le secteur privé sont donc la plupart du temps inapplicables dans le secteur public. Ensuite, si les élections permettent certes aux citoyens d'exprimer leurs préférences sur les politiques à travers le choix de gouvernants, ce choix est indirect, agrégé (c.-à-d., que le vote porte sur l'ensemble de la plateforme d'un parti ou d'un candidat), motivé par plusieurs considérations (p. ex., les compétences des candidats, leur idéologie, etc.) et n'est pas tenu à une fréquence suffisante pour constituer une rétroaction efficace.

Enfin, les sondages d'opinion et les consultations publiques sont des outils de pilotage de l'action publique utiles mais d'une portée limitée. D'une part, les citoyens sondés ne possèdent pas toujours les connaissances nécessaires à la formation et à l'expression d'un jugement éclairé. D'autre part, la satisfaction n'est que l'un des critères (voir ci-dessus) mobilisés pour évaluer des politiques publiques et certainement pas toujours le plus important. Il est en effet possible que les personnes sondées ou consultées soient satisfaites d'une politique inefficace et vice-versa.

L'évaluation constitue ainsi un outil indispensable aux élus, concepteurs de politiques et gestionnaires de programmes notamment, pour prendre des décisions éclairées sur les programmes publics, qu'il s'agisse de rendre des comptes sur ceux-ci, de décider de leur sort (maintien, abandon ou réforme) ou encore de générer des connaissances généralisables. Entre deux politiques, laquelle faut-il adopter? Faut-il maintenir ou abolir tel programme? Comment améliorer l'efficacité d'une intervention donnée? Les intentions du gouvernement en termes de politiques ont-elles été mises en œuvre comme prévu dans la société? Les ressources prévues pour ce programme ont-elles été utilisées de manière appropriée? L'évaluation de politique, en offrant des réponses crédibles à ces questions, peut contribuer à rehausser la qualité de la prise de décision publique.

Survol de la problématique : la participation

Cette thèse s'inscrit dans un courant qui structure de manière importante la théorie, la recherche et la pratique de l'évaluation, soit la *participation des parties prenantes* à l'évaluation. L'implication des parties prenantes (*stakeholders*) – gestionnaires, professionnels de première ligne, bénéficiaires des programmes, etc. – dans le processus d'évaluation est un principe qui est maintenant bien accepté par la communauté des évaluateurs (Mathison, 2005a; Whitmore, 1998). Certains ont souligné que la participation constitue l'une des tendances les plus importantes observées dans le domaine (Mark, 2001). Les approches évaluatives qui font participer les parties prenantes au processus d'évaluation telles que l'évaluation centrée sur l'utilisation (Patton, 2008), l'évaluation habilitative (Fetterman, 2000) et l'évaluation démocratique-délibérative (House & Howe, 2000) sont d'ailleurs fort populaires. Que le recours aux approches participatives soit motivé par des considérations *pragmatiques* (c.-à-.d., accroître l'utilisation de l'évaluation),

politiques (c.-à-.d., assurer l'équité et la justice sociale) ou *épistémologiques* (c.-à-.d., accroître la « validité » du savoir généré par l'évaluation) (Weaver & Cousins, 2004), il est de plus en plus valorisé. Certains auteurs ont d'ailleurs dénoncé ce qu'ils considèrent être une « orthodoxie participative », un consensus problématique sur les approches participatives (Biggs, 1995, cité dans Gregory, 2000, p. 180). De là à soutenir que l'évaluation participative ne comporte que des avantages ou est appropriée en toutes circonstances, il n'y a qu'un pas que tous ne sont pas prêt à franchir (Gregory, 2000; Jacob et Daigneault, 2011; Jacob et Ouvrard, 2009). Quelques auteurs adoptent même un ton plus critique sur les démarches participatives lorsqu'ils dénoncent l'amateurisme et les menaces que font peser ces approches sur la validité, la crédibilité et l'éthique de l'évaluation (Scriven, 1997, 2005). Cela étant dit, qu'on défende l'idéal participatif ou qu'on le remette en question, l'influence qu'il exerce dans le champ de l'évaluation est indéniable.

La popularité croissante de la participation en évaluation cause cependant certains problèmes conceptuels. D'une part, la multiplication des termes et des acceptions font de la participation un concept ambigu et vague dont l'utilité est limitée dans le cadre de recherches empiriques rigoureuses. Ces problèmes conceptuels ont d'ailleurs été soulignés à maintes reprises par le passé (Huberman, 1995; Murray, 2002; Rebien, 1996; Ridde, 2006) mais les doléances de ces auteurs sont pour la plupart restées lettre morte. D'autre part, la charge normative (positive) du concept de participation fait en sorte qu'il est parfois difficile de distinguer ce qui relève, d'un côté, de la réflexion fondée sur des préférences normatives soutenues par des anecdotes, et, de l'autre, d'inférences fondées sur la mobilisation d'un protocole de recherche scientifique théoriquement reproductible.

Or, seul un concept construit et opérationnalisé de manière adéquate peut contribuer à une production scientifique cumulative de qualité (p. ex., sur la relation entre participation et utilisation de l'évaluation). Il semble en effet que ce thème a pris beaucoup d'expansion depuis quelques années, comme en témoignent les nombreuses recherches récentes sur le sujet (p. ex., Cousins & Chouinard, à paraître; Cullen, Coryn, & Rugh, 2011; King et collab., 2011; Laudon, 2010; Lawrenz, King, & Ooms, 2011; Poth & Shulha, 2008; Rodrígues-Campos, 2012; Toal & Gullickson, 2011). Certains vont même jusqu'à affirmer que la participation constitue la principale variable indépendante étudiée en relation avec

l'utilisation: « L'influence d'une implication accrue des parties prenantes dans tous les aspects du processus évaluatif a dominé la recherche sur l'utilisation de l'évaluation dans les années 90 » (Poth, 2008, p. 36 : traduction libre).

Visée générale et objectifs de recherche

Cette thèse vise à circonscrire la nature de l'évaluation participative et à en permettre la mesure. Elle est structurée autour de trois grandes questions de recherche :

1) *Qu'est-ce que la participation à l'évaluation?*
2) *Comment traduire ce concept en un instrument de mesure opérationnel?*
3) *Est-ce que cet instrument mesure la participation de manière fidèle et valide?*

1) Qu'est-ce que la participation à l'évaluation? Le concept de participation à l'évaluation est vague et ambigu. En effet, le sens attribué au concept varie d'un chercheur à l'autre et différents termes sont utilisés pour désigner ce phénomène. Il s'agit donc de procéder à une analyse systématique du concept afin de mettre au jour son « ontologie », sa nature réelle (voir Goertz, 2006). Analyser systématiquement l'ontologie de la participation, c'est d'abord en circonscrire les dimensions constitutives. Il s'agit ensuite de traduire ces dimensions sur le plan conceptuel et d'identifier la logique d'agrégation qui unit ces dimensions entre elles, soit la structure du concept (Goertz, 2006). C'est enfin déterminer comment le concept de participation se distingue de concepts apparentés tels que l'évaluation « conventionnelle » ou « traditionnelle ». L'objectif est de former un concept clair, univoque, cohérent, parcimonieux mais couvrant adéquatement le domaine de la participation, précis et utile dans le cadre de recherches empiriques (Gerring, 1999).

2) Comment traduire ce concept en un instrument de mesure opérationnel? Un concept est une construction mentale qu'on utilise pour faire sens du monde qui nous entoure (Sartori, 1984], 2009; Schedler, 2011).[2] Or, il n'y a pas de correspondance parfaite entre le monde conceptuel et la réalité empirique. En d'autres termes, les concepts véhiculés par le langage

[2] Tout au long de cette thèse, nous entendrons *concept* au sens d'un *concept empirique*, soit « Any concept that is amenable, no matter how indirectly, to observations. Thus empirical concepts involve observational terms and *referents*. Contrasted with theoretical terms, logical concepts (e.g., "analytics") and metaphysical concepts (e.g., "absolute beings") » (Sartori, [1984], 2009, p. 137). Un concept empirique peut ainsi être très abstrait, comme le conflit ou la participation, mais néanmoins observable (Sartori, [1975], 2009).

(démocratie, révolution, amitié, etc.) ne peuvent être observés directement dans la réalité. Il est donc nécessaire d'opérer une traduction du concept, de l'opérationnaliser en indicateurs et en règles qui en permettent l'application dans l'empirie. Il s'agit de développer un instrument de mesure prêt à utiliser qui semble traduire de manière adéquate le concept de participation.

3) Est-ce que cet instrument mesure la participation de manière fidèle et valide? Un instrument permet une mesure fidèle dans la mesure où, lorsqu'appliqué au même objet, il produit une mesure semblable à travers le temps et l'espace (Carmines, Woods, & Kimberly, 2005; Durand & Blais, 2006). Il s'agit d'abord de déterminer si deux personnes utilisant sur un même échantillon l'instrument de mesure de la participation qui a été développé dans le cadre de cette thèse arrivent à un résultat comparable. Une fois la fidélité de la mesure établie, il s'agit d'évaluer si ce qui est mesuré est « vraiment » la participation.

Contribution

En offrant des réponses crédibles aux questions précédentes, cette thèse apporte une contribution à la fois théorique, méthodologique et empirique au domaine de l'évaluation des politiques. Tout d'abord, sur le plan *théorique*, une conceptualisation de l'évaluation participative cohérente et ancrée dans les écrits du domaine sera proposée. Les approches participatives jouissent d'un pouvoir d'attraction fort et croissant dans le champ de l'évaluation, ce qui ouvre malheureusement la porte à toutes sortes d'abus dans le discours et dans la pratique (Cousins & Chouinard, à paraître). Cette conceptualisation, nécessaire dans l'état actuel d'ambiguïté qui règne autour de la participation, permettra de réduire la confusion en ce qui a trait aux frontières du concept et de faciliter la réflexion théorique à son sujet. Ensuite, sur le plan *méthodologique*, un instrument de mesure ancré dans la conceptualisation précédente de la participation sera développé. Cet instrument, baptisé *Participatory Evaluation Measurement Instrument* (PEMI), permettra aux commanditaires, aux évaluateurs et aux chercheurs de mesurer avec précision le niveau de participation de cas réels d'évaluation. Enfin, sur le plan *empirique*, des résultats viendront confirmer la fidélité et la validité des scores générés par l'instrument.

Organisation

Cette thèse comporte quatre chapitres. Le premier chapitre présente une revue de la littérature générale sur la conceptualisation et la mesure, l'évaluation de programme et la participation à l'évaluation. Le second chapitre contient une analyse systématique et une opérationnalisation du concept de participation dans un contexte évaluatif. Le troisième chapitre présente une étude de validation empirique quantitative de l'instrument de mesure de la participation proposé au second chapitre. Le quatrième chapitre présente également une validation empirique de l'instrument, mais à l'aide de méthodes mixtes. Il propose par ailleurs une version révisée de la conceptualisation et de l'instrument de mesure de la participation à l'évaluation. Tous les chapitres de la thèse à l'exception du premier couvrent, à des degrés divers, les questions de recherche présentées à la section précédente (voir Tableau 1).

Tableau 1 : Traitement des questions de recherche par chapitre

	1) Qu'est-ce que la participation à l'évaluation?	2) Comment traduire ce concept en un instrument de mesure opérationnel?	3) Est-ce que cet instrument mesure la participation de manière fidèle et valide?
Chapitre 2	✓	✓	✓
Chapitre 3	✓	✓	✓
Chapitre 4	✓	✓	✓

Note : un *grand* crochet indique que le chapitre est substantiellement centré sur la question de recherche tandis qu'un *petit* crochet indique que le chapitre aborde la question de recherche de manière secondaire.

> Les chercheurs débutants pensent que le but de la recension des écrits est de trouver des *réponses* relativement au sujet de recherche; au contraire, les chercheurs expérimentés étudient les recherches antérieures pour développer des *questions* plus intelligentes et plus pénétrantes à propos du sujet.
> (Yin, 1994, cité et traduit dans Chevrier, 2006, p. 53)[3]

1 Recension des écrits

Ce premier chapitre présente une recension *sélective* et *ciblée* des écrits traitant de trois thèmes centraux à cette thèse, soit la conceptualisation et la mesure, l'évaluation des politiques, ainsi que la participation à l'évaluation/évaluation participative. Le ciblage des sources permet d'éviter ou, à tout le moins, de minimiser, les redondances contreproductives qui ne manqueraient pas de survenir entre les différents chapitres d'une thèse avec insertion d'articles.

Nous examinerons dans un premier temps le rôle joué par le concept dans l'entreprise scientifique, sa nature, son analyse et sa mesure. Cet exposé méthodologique s'avèrera essentiel à la compréhension du propos des sections qui suivent. Nous circonscrirons dans un second temps l'évaluation des politiques sur les plans théorique, historique et pratique. Bien que l'évaluation existe depuis fort longtemps, sa nature transdisciplinaire commande en effet de la situer par rapport aux pratiques apparentées avec lesquelles on la confond parfois. Nous nous tournerons ensuite vers la participation des parties prenantes (*stakeholders*) à l'évaluation qui constitue une tendance lourde du domaine de l'évaluation. Dans tous les cas, il s'agira d'insister sur les enjeux, préoccupations et débats qui caractérisent ces objets afin de situer notre problématique de recherche.

[3] Les caractères italiques et standards ont été interposés par rapport à la citation originale.

1.1 Conceptualisation et mesure[4]

1.1.1 Le concept, ce grand négligé

Thinking without the positing of categories and concepts in general would be as impossible as breathing in a vacuum.
(1949, cité dans Shadish, Cook, & Campbell, 2002, p. 65)

Peu importe leur discipline ou leur tradition épistémologique, les scientifiques qui étudient l'être humain et la société ne peuvent conduire leur entreprise de connaissance en l'absence de concepts. Des concepts tels que nationalisme, démocratie, genre, anomie et développement économique jouent en effet un rôle capital dans plusieurs propositions et modèles théoriques mais ne peuvent être directement observés dans la réalité. Les concepts et les mots qui leur servent de support constituent les unités de base de la pensée (Sartori, [1975] 2009, [1984], 2009); ils sont des « contenants » sémantiques. Ces unités de base peuvent être organisées et combinées de manière à former des propositions, des hypothèses, des schèmes d'analyse et des théories. Goertz (2006) rappelle d'ailleurs que les théories scientifiques sont construites à partir des « briques » que sont les concepts. Le rôle des concepts ne se limite toutefois pas au niveau théorique, loin s'en faut. L'étude empirique implique la réduction du réel en observations et en faits qui doivent d'abord être circonscrits pour ensuite être organisés, classés et mis en relation. Le concept présente une fonction de découpage du réel : il sert donc tout autant de « contenant à données » (*data container*) que de contenant sémantique (Sartori, [1975] 2009).

Affirmer que la pratique scientifique ne saurait se passer des concepts à proprement parler ainsi que des règles pour les analyser et les apprécier est une évidence qu'il est néanmoins utile de rappeler. Puisque les écrits théoriques et empiriques foisonnent dans les sciences sociales, l'enjeu n'est dès lors pas que les scientifiques n'ont pas recours aux concepts mais bien que, collectivement, ils les utilisent de manière non systématique et négligent les principes qui doivent guider leur analyse. En effet, bien des chercheurs qui manient les concepts le font de manière informelle et intuitive, sans porter une attention particulière à

[4] Jusqu'à la sous-section « La mesure » (p. 15), cette section reprend presque textuellement une portion de l'article suivant : « Les concepts souffrent-ils de négligence bénigne en sciences sociales? Éléments d'analyse conceptuelle et examen exploratoire de la littérature francophone à caractère méthodologique » (Daigneault & Jacob, à paraître, 2012).

cet aspect de la recherche. Ces scientifiques ne sont d'ailleurs pas sans rappeler Monsieur Jourdain qui parlait en prose sans le savoir dans le *Bourgeois gentilhomme* de Molière; ils emploient, construisent et reconstruisent des concepts mais sans être conscients que c'est effectivement ce qu'ils font. En matière de conceptualisation, bien des scientifiques du social se basent sur un « savoir populaire », sur des intuitions plus ou moins implicites (Gerring, 1999). De cela découle des conséquences néfastes sur la communication entre scientifiques et sur l'accumulation des connaissances. Certains reprochent aux scientifiques de déformer le sens des concepts en faisant violence à leur étymologie et d'inventer de nouveaux termes à un rythme effréné (Sartori, [1975] 2009). Cette situation peut être mieux comprise à partir de l'éclairante métaphore du jeu de cartes proposée par Sartori ([1975] 2009) :

> Le rapport de l'instrument linguistique à la connaissance scientifique ressemble – avec quelques différences, certes – au rapport des cartes à un jeu de cartes. Le jeu (avec ses possibilités presque infinies) peut être joué seulement parce que les cartes et les règles de leur combinaison sont données – elles sont en effet statiques. D'une manière pas trop dissemblable, seule une utilisation disciplinée des termes et des procédures de leur composition (et décomposition) permet aux scientifiques de jouer ce jeu. Par contraste, nous les scientifiques du social investissons de plus en plus de nos énergies à simplement *altérer les cartes*. S'il en est ainsi, nous ne faisons pas progresser la science mais bien la confusion la plus absolue. Nous démantelons plutôt que reconstruisons la connaissance cumulative que nous avons atteinte, quel que soit son niveau. (p. 64)

Si la communauté savante agit ainsi, nous émettons l'hypothèse que c'est d'abord en raison du fait que l'analyse conceptuelle souffre de négligence « bénigne » ou « bienveillante » (*benign neglect*)[5] en sciences sociales. Cette situation ne doit pas surprendre étant donné que ni les concepts comme objet de recherche, ni l'analyse conceptuelle comme méthode n'ont reçu beaucoup d'attention dans les écrits à caractère méthodologique (Gerring, 1999; Goertz, 2006). Schedler (2011) soutient à cet égard que l'analyse des concepts n'est souvent considérée que comme une simple « distraction », comme un « prélude à la recherche sérieuse » (traduction).

[5] Dans le domaine des politiques publiques, la « négligence bénigne » est un concept désignant une politique (à caractère ethnique en particulier) de laissez-faire consistant à ne pas se préoccuper d'un problème afin de ne pas l'aggraver. Par extension, on parle de négligence bénigne lorsque des individus se désintéressent d'une situation jugée indésirable et ne font pas d'effort pour y remédier.

Si ce constat vise la science politique contemporaine, il s'applique de manière plus large aux sciences sociales. Duval (2004) affirme ainsi que « …"l'analyse conceptuelle" n'est pas une méthode d'analyse, ni une opération de recherche, véritablement identifiée comme telle, et encore moins codifiée, en sociologie » (p. 131) et que les concepts occupent sans doute une place plus importante dans les analyses des philosophes des sciences que dans la production des scientifiques. D'autres soulignent que l'attention particulière accordée aux référents empiriques dans l'entreprise scientifique se fait au détriment des concepts (Robert-Demontrond, 2004). La faible visibilité de l'analyse conceptuelle dans les cours et ouvrages méthodologiques des différentes disciplines scientifiques laisse croire – à tort – qu'elle ne représente aucun enjeu, qu'elle ne pose aucun défi particulier. Or, il existe des principes, des règles et des normes pour analyser adéquatement les concepts, au même titre que pour réussir son terrain en ethnographie ou analyser des données statistiques ordinales.

1.1.2 Éléments d'analyse conceptuelle[6]

Il faut préciser d'emblée que le traitement des concepts et de leur analyse présenté dans cette section ne vise ni à être exhaustif ni à contribuer à l'avancement des connaissances. L'objectif visé est plus modeste : il s'agit plutôt d'exposer quelques éléments choisis d'analyse conceptuelle à l'intention des scientifiques du social. Le lecteur intéressé par les derniers développements sur la nature des concepts et leur analyse est ainsi invité à consulter les sources portant sur la logique, l'épistémologie ou la philosophie du langage.

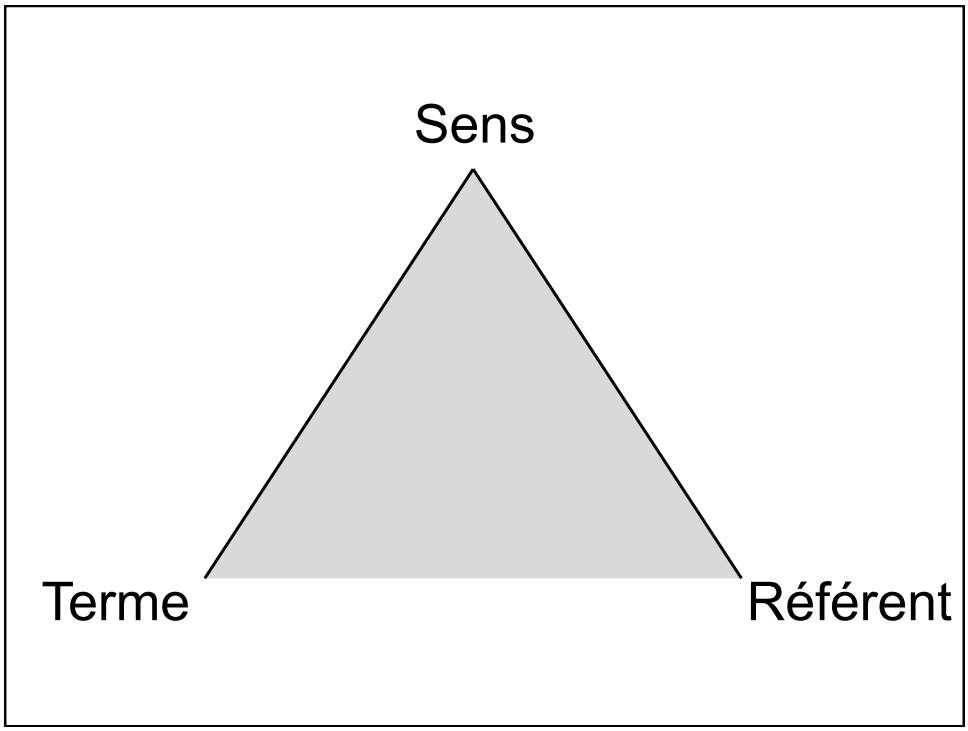
Qu'est-ce qu'un concept?

Pour reprendre une formule bien connue, le scientifique sait probablement reconnaître un concept quand il en voit un. Une conception implicite des concepts est toutefois loin d'être suffisante afin d'assurer une utilisation appropriée de ceux-ci. En effet, seule une conception explicite des concepts permet de bien appréhender les enjeux ayant trait à leur

[6] D'un point de vue philosophique, la position défendue ici s'inscrit en faux contre la thèse *opérationaliste* (voir Chang, 2009). Selon cette thèse, du moins sa version « réductrice », le sens d'un concept se limite aux opérations et instruments utilisés pour le mesurer. Si on utilise deux instruments pour mesurer l'intelligence, deux tests différents par exemple, on mesure en fait deux concepts différents. Or, pour nous, il est clair que la conceptualisation peut – et dans une certaine mesure doit – être appréhendée de manière indépendante de la mesure, que la première précède la seconde (approche déductive) ou vice-versa (approche inductive). Cela étant dit, même dans le cas d'une approche inductive, le sens d'un concept n'est pas épuisé par son opérationnalisation. Dans le cas contraire, il serait en effet futile de procéder à la validation d'un instrument de mesure, comme nous le faisons dans cette thèse.

analyse. Dans la tradition classique, le concept est considéré comme une entité symbolique faisant le pont entre le monde réel et la pensée (Schedler, 2011). Selon une conception éclairante qui puise ses origines dans les travaux menés par Charles Kay Ogden et Ivor Armstrong Richards dans les années 1920 et qui a été largement reprise depuis (Gerring, 1999; Sartori, [1984] 2009; Schedler, 2011), le concept est représenté sous la forme d'un triangle (voir Figure 1). Le concept est ainsi composé de trois dimensions : le sens, le terme et le référent.

Figure 1 : La conception tripartite du concept



Note : Adapté de Sartori (2009 [1984]).

Le concept possède d'abord une dimension sémantique. Il est porteur de sens, c'est-à-dire qu'il exprime une idée. C'est l'unité de base de la pensée. On désigne également le sens comme la *compréhension* d'un concept (synonymes : *connotation* ou *intension*). La compréhension « consiste en l'ensemble des caractéristiques et attributs associés à, ou inclus dans, un mot, terme ou concept donné » (Sartori, [1984] 2009, p. 103; traduction).

Le concept est ensuite caractérisé par un terme, un mot, c'est-à-dire un symbole servant à désigner et véhiculer l'unité de sens. Plusieurs termes peuvent désigner une même idée,

peuvent avoir le même sens. Ce sont alors des (quasi) synonymes, comme *automobile* et *voiture*. Un seul terme peut également posséder plusieurs sens : il s'agit alors d'homonymes. Par exemple, le terme *gouvernance* peut être compris soit comme un terme générique référant au mode de gouverne d'une organisation ou d'une société politique, soit comme un nouveau mode de gouverne caractérisé par l'inclusion d'acteurs de la société civile dans une relation de coordination avec l'État (Rouillard & Burlone, 2011).

Le troisième élément constitutif du concept est sa dimension référentielle, c'est-à-dire qu'il réfère à des objets du monde réel (voir n. 2, supra). Le référent, aussi appelé *extension* (synonyme : *dénotation*), désigne ainsi la classe des objets à laquelle le concept réfère (Sartori, [1984] 2009). La distinction entre la compréhension et l'extension d'un concept peut être clarifiée par un exemple. La *démocratie* peut ainsi être sommairement définie comme un régime politique au sein duquel les gouvernants sont désignés par la majorité des citoyens au moyen d'élections. Il s'agit de la compréhension du concept, soit l'attribut qui le caractérise. D'autres attributs peuvent évidemment être mobilisés pour définir la démocratie, notamment la liberté de devenir candidat aux élections et le respect des droits civils.[7] Quant à l'extension du concept de démocratie, elle réfère à la classe des objets du monde réel qui présentent cet attribut. On retrouve notamment dans cette classe les régimes politiques de la Suède, de la France, du Canada et du Japon.[8]

Quelles sont les qualités d'un « bon » concept?

Les concepts peuvent être évalués à l'aune de plusieurs critères ou dimensions. Gerring (1999) propose à titre d'exemple un cadre d'évaluation qui comporte huit critères. D'autres auteurs ont proposé un traitement exhaustif des désordres ou problèmes pouvant affecter les concepts (Sartori, [1984] 2009; Schedler, 2011). S'il ne fait aucun doute que les concepts utilisés en sciences sociales doivent satisfaire à plusieurs éléments d'évaluation, deux critères apparaissent fondamentaux : la *clarté* et la *précision*. Sur ces critères reposent

[7] À titre d'exemple, Robert Dahl (2000) propose six conditions nécessaires (mais non suffisantes) au concept de démocratie dans les sociétés politiques à grande échelle : 1) des titulaires de charge publique élus; 2) des élections libres, justes et fréquentes; 3) la liberté d'expression; 4) des sources d'information alternatives; 5) l'autonomie d'association, et; 6) une citoyenneté inclusive.

[8] La démocratie n'est évidemment pas l'apanage de l'État-nation. Les régimes politiques de plusieurs villes, provinces, communes, régions et organisations appartiennent ainsi à l'extension du concept.

en effet la communication efficace de l'activité scientifique et la réalisation de recherches empiriques valides et cumulatives. En premier lieu, un concept doit être clair et univoque; il ne doit pas être ambigu. Cela signifie que dans un contexte donné, un terme ne doit avoir qu'un seul sens (Sartori, [1984] 2009). Il faut ainsi éviter autant que possible les synonymes (plusieurs termes pour un même sens) et les homonymes (plusieurs sens pour un même terme). Cette qualité terminologique des concepts concerne la relation entre le terme et le sens, soit le côté gauche du triangle (voir Figure 1).

En second lieu, un concept doit être précis. Cela signifie que le concept ne doit pas être vague, qu'il doit permettre de distinguer efficacement les phénomènes empiriques auxquels il s'applique de ceux auxquels il ne s'applique pas (Sartori, [1984] 2009). Le pouvoir discriminant des concepts concerne la relation entre le sens et le référent, soit le côté droit du triangle (Figure 1). La définition et l'opérationnalisation du concept doivent être fondées sur des éléments au pouvoir discriminant élevé.

Comment analyser les concepts? Une introduction à l'approche « classique »

S'il n'existe pas de « livre de recettes » ou de règles immuables et valables en toutes circonstances pour analyser les concepts (Gerring, 1999; Schedler, 2011), il est néanmoins possible de tracer les grandes lignes d'une approche que l'on pourrait qualifier de « classique ». Cette approche, qui a été popularisée par Sartori ([1984] 2009) mais qui remonte à Aristote (voir Goertz, 2009), mobilise les outils logiques afin d'analyser les concepts en fonction de leurs conditions *nécessaires* et *suffisantes*. Nous nous concentrons sur cette approche parce qu'elle offre des outils qui nous semblent simples et potentiellement utiles aux scientifiques du social.

Le principe fondamental de cette approche consiste à offrir une définition *minimale* d'un concept. On entend par « minimale » une définition qui mobilise uniquement les caractéristiques (attributs, propriétés ou dimensions) essentielles d'un concept, c'est-à-dire celles qui sont *nécessaires* à sa définition. Il faut donc distinguer les caractéristiques nécessaires de celles qui sont contingentes ou accidentelles, c'est-à-dire qu'on peut retrouver ou non dans la définition de ce même concept sans en affecter significativement l'extension. Selon cette conception, un concept est défini adéquatement lorsqu'il comporte assez de caractéristiques nécessaires permettant d'identifier avec précision ses référents

empiriques et de les distinguer des objets auxquels le concept ne s'applique pas. En d'autres termes, chacun des attributs faisant partie du concept doit être nécessaire. Les attributs pris conjointement doivent en outre être *suffisants* pour définir le concept, c'est-à-dire qu'aucun attribut supplémentaire n'est requis pour le définir. Un concept est ensuite défini de manière parcimonieuse lorsque ses attributs n'incluent pas de caractéristiques accidentelles (Sartori, [1984] 2009, pp. 126-127). La distinction entre une caractéristique essentielle et accidentelle d'un concept peut être illustrée à l'aide de l'exemple du régime démocratique. La démocratie *requiert* une forme ou l'autre de suffrage permettant aux citoyens de choisir leurs gouvernants : c'est une condition nécessaire. En revanche, un régime démocratique *peut* posséder une chambre haute ou sénat. Puisque certaines démocraties en possèdent et d'autres non, il s'agit d'une caractéristique contingente ou accidentelle.

Comment faire pour mettre au jour les attributs nécessaires d'un concept? Sauf dans les cas où l'on crée de nouveaux concepts, la première étape consiste à passer en revue la documentation pertinente à ce concept (Sartori, [1984] 2009). Les attributs d'un concept sont généralement présentés de manière explicite ou implicite dans la section théorique des études, notamment au sein des définitions. Il suffit d'extraire de ces dernières les attributs et de les organiser. Un tableau croisé au sein duquel chaque ligne réfère à une étude et chaque colonne à un attribut constitue un outil efficace d'analyse. Les attributs ainsi recensés peuvent former plusieurs configurations dont certaines se prêtent mieux à l'analyse que d'autres (Sartori, [1984] 2009). À titre d'exemple, les sources consultées pourraient avoir permis de cibler quinze caractéristiques pour un concept dont trois sont communes à toutes les études. Un consensus partiel chez les auteurs (quant aux attributs d'un concept) et la prévalence élevée de certains attributs dans les écrits relatifs à ce concept sont généralement des indicateurs de leur importance. Toutefois, la sélection des attributs doit ultimement se fonder sur leur caractère nécessaire dans la définition d'un concept. Mais comment établir cette relation de nécessité? En fait, les attributs nécessaires d'un concept sont ceux qui le circonscrivent dans son extension, c'est-à-dire dans son application empirique. Si on cible la capacité de voler comme un attribut essentiel du concept d'oiseau, on se retrouve avec certains cas problématiques telle l'autruche qui n'a pas la capacité de voler mais qui est un oiseau (Sartori, [1984] 2009). On doit donc en

conclure de deux choses l'une : soit la capacité de voler est un attribut contingent (mais fréquent) des oiseaux, soit l'autruche n'est pas un oiseau.

Le va-et-vient entre les attributs et leurs implications pour circonscrire l'extension d'un concept n'est d'ailleurs pas sans rappeler la méthode de l'*équilibre réfléchi* décrite par le philosophe politique John Rawls (1951, 1971, cité dans Arnsperger & van Parijs, 2003). Cette démarche largement intuitive consiste à confronter les implications de nos principes éthiques à nos jugements moraux bien réfléchis dans des circonstances diverses. En cas de conflit, nous sommes forcés d'abandonner le principe en question ou de le réviser de manière à résoudre le conflit moral qu'il cause. Nous devons également pouvoir étayer nos jugements moraux par des raisons. Le processus d'analyse conceptuelle est très similaire à la recherche d'un équilibre réfléchi sur le plan éthique. En effet, lorsqu'un conflit[9] apparaît entre l'attribut d'un concept et son extension, on doit examiner les raisons qui nous poussent à cibler tel attribut ou à considérer tel objet comme étant un référent de ce concept. L'analyse conceptuelle est ainsi un processus intuitif qui vise la cohérence.

1.1.3 La mesure

Aussi bien conçu qu'il puisse l'être, le concept demeure une construction théorique et abstraite. Si le concept est fondamental dans le cadre de développements à caractère théorique, il ne peut cependant pas être appliqué *directement* au réel (Carmines et collab., 2005; McDonald, 2005). À titre d'exemple, soumettre à un test empirique l'hypothèse selon laquelle la participation à l'évaluation accroît l'utilisation de cette dernière requiert la mesure des concepts de participation et d'utilisation et donc leur traduction en termes opérationnels :

> Ce qu'il importe de retenir, c'est que l'hypothèse renvoie à un ou plusieurs concepts, que ces concepts sont abstraits et qu'on a besoin de signes concrets de ces concepts pour être en mesure de confirmer ou d'invalider l'hypothèse. [alinéa] Le processus par lequel on passe des concepts abstraits à des indicateurs concrets, c'est la ***mesure***. La mesure est définie comme l'***ensemble des opérations empiriques, effectuées à l'aide d'un ou de plusieurs instruments de mise en forme de l'information, qui permet de classer un objet dans une catégorie pour une caractéristique donnée***. (Durand & Blais, 2006, p. 188)

[9] Dans le cas de l'analyse conceptuelle, le conflit n'est pas moral mais cognitif. La conceptualisation entre en conflit avec les idées reçues ou nos intuitions à propos de la nature du concept.

Cette opération de traduction du concept vers la mesure, appelée opérationnalisation, consiste en l'identification d'indicateurs du concept et de ses dimensions, d'une part, et des règles précises d'assignation aux catégories, d'autre part.

Toutes les conceptualisations ne se valent pas – nous l'avons vu à la section précédente – et il en va de même pour les différentes mesures d'un concept. Il existe en effet plusieurs critères d'appréciation du processus de mesure.

Une mesure se doit d'abord d'être *fidèle* ou fiable, c'est-à-dire qu'elle doit donner des résultats constants à travers le temps (stabilité) et l'espace (équivalence), du moment que l'objet auquel est appliqué l'instrument demeure le même (Carmines et collab., 2005; Durand & Blais, 2006). La fidélité intercodeur ou interjuge est un cas spécifique d'équivalence où l'on cherche à déterminer si les valeurs que prend la mesure découlent du phénomène observé ou des observateurs (DeVellis, 2005). La fidélité est indissociable du concept d'erreur de mesure. L'erreur de mesure survient lorsque la mesure effectuée avec un instrument diffère du « vrai » score (hypothétique car impossible à appréhender de manière directe). L'erreur de mesure peut être systématique (c.-à-d., l'écart entre la vraie mesure et les mesures effectuées va toujours dans la même direction) ou aléatoire (c.-à-d., l'écart ne présente pas de patron particulier). On considère que la fidélité entretient une relation négative avec l'erreur de mesure aléatoire : « For example, if a bathroom scale shows a weight as 5 pounds greater than the true weight on the first reading, 8 pounds greater on the second reading, and 10 pounds less on the third reading, the readings are being affected by random error and the reliability of the scale is low » (Carmines et collab., 2005, p. 934).

La fidélité est un critère d'appréciation qui porte exclusivement sur la qualité de la mesure « en soi » (Durand & Blais, 2006, p. 196). La fidélité peut être parfaite, constante dans le temps et l'espace, sans que la mesure ne soit pour autant *valide*. Un instrument produit des inférences descriptives valides si ce qu'il mesure correspond à ce qu'il est censé mesurer. Pour poursuivre avec l'exemple précédent, un pèse-personne peut donner une lecture parfaitement fidèle lorsqu'il est utilisé par différentes personnes à différentes occasions tout en surévaluant le poids réel de chaque individu de 50 kg. Dans ce cas, on ne peut considérer que c'est un instrument de mesure valide du poids (selon ce pèse-personne, un bambin dont

le poids réel est de 11 kg pourrait se voir faussement attribuer un « poids » dépassant les 60 kg!). En ce sens, la fidélité est une condition nécessaire mais non suffisante de la validité, comme le soulignent Fleiss, Levin et Paik (2004) à propos d'un contexte où plusieurs codeurs sont employés pour effectuer la mesure :

> If agreement among the raters is good, then there is a possibility, but by no means a guarantee, that the ratings do in fact reflect the dimension they are purported to reflect. If their agreement is poor, on the other hand, then the usefulness of the ratings is severely limited, for it is meaningless to ask what is associated with the variable being rated when one cannot even trust those ratings to begin with. (p. 598)

Puisqu'il est impossible d'observer « sans filtre » un concept, apprécier la validité d'une mesure est une tâche beaucoup plus complexe qu'apprécier sa fidélité. L'examen de la validité d'un instrument de mesure est traditionnellement fondé sur trois « types » de validité – validité de contenu, critérielle et de construit – qui sont métaphoriquement représentés sous les traits de la « Sainte Trinité conduisant au salut psychométrique » (Guion, 1980, cité dans Adcock & Collier, 2001, p. 537: adaptation). Même si cette conception demeure influente, la validité est désormais conçue comme un concept unitaire par un nombre croissant de chercheurs, et est tour à tour désignée par les termes *validité de mesure*, *validité de construit* ou tout simplement *validité* (voir Adcock & Collier, 2001; Toal, 2009).

Selon la conception unitaire, la validité est fondée sur un argument incorporant des preuves empiriques générées par différents types de *validation* (par opposition à validité). Bien qu'il existe une multitude de termes pour désigner les types de validation – apparente (*face validation*), d'échantillonnage, statistique, etc. (Carmines et collab., 2005; Durand & Blais, 2006) –, Adcock et Collier (2001) ont proposé de regrouper ceux-ci en trois catégories. D'abord, la *validation de contenu* consiste à évaluer dans quelle mesure l'opérationnalisation capture l'ensemble des dimensions importantes d'un concept et seulement celles-ci. Si on cherche à mesurer l'intelligence d'adolescents à l'aide d'un instrument dont les questions portent exclusivement sur la culture générale, des dimensions importantes de l'intelligence sont négligées, notamment la capacité de raisonner de manière logique et d'effectuer des calculs mathématiques. Par conséquent, la validation du contenu de cet instrument indiquerait certainement des lacunes à cet égard. Ensuite, la *validation*

convergente/discriminante vise à déterminer si les scores générés par des indicateurs alternatifs d'un même concept sont positivement associés (convergence), d'une part, et si les scores générés par des indicateurs de deux concepts différents sont négativement ou faiblement associés, d'autre part (Adcock & Collier, 2001). À titre d'exemple, les scores sur un instrument *x* visant à mesurer le niveau de démocratie d'un État devraient être positivement associés à ceux d'un instrument *y* mesurant le même concept et négativement associés à un instrument mesurant le niveau d'autoritarisme. Enfin, la *validation nomologique* (ou de construit) cherche à déterminer si, pour une relation causale établie ou plausible, les scores mesurés pour deux concepts corroborent la relation anticipée. S'il ne s'agit pas d'un principe ou d'une loi invariable, une hypothèse centrale dans le champ de l'évaluation des politiques est à l'effet que l'évaluation participative accroît l'utilisation de l'évaluation (p. ex., Cousins, 2003). Ainsi, une corrélation positive entre les scores obtenus avec un instrument de mesure de la participation et avec un instrument de mesure de l'utilisation pour un échantillon donné serait une indication de la validité de ceux-ci.

Dans tous les cas, il faut insister sur le fait qu'appréhender la validité en termes dichotomiques (c.-à-d., valide ou invalide) est contreproductif. En effet, la validité des mesures générées par un instrument est toujours une question de degré et est déterminé par la qualité des preuves découlant du processus de validation (McDonald, 2005; Toal, 2009). Cette « qualité » est déterminée par les *risques de biais* occasionné par la démarche méthodologique utilisée. Une démarche rigoureuse permet ainsi de minimiser les menaces à la validité des inférences de la mesure (Shadish, Cook & Campbell, 2002).

La recension des écrits sur les concepts (nature, importance, règles d'analyse, etc.) et la mesure (critères d'évaluation, procédures de validation, etc.) présentée précédemment a contribué à poser des jalons essentiels à la compréhension de l'objet d'étude de cette thèse, soit l'évaluation participative. Puisque cette dernière est un type d'évaluation, nous nous tournons maintenant vers une présentation de l'évaluation de politique qui vise à la situer sur les plans théorique, historique et pratique.

1.2 Évaluation de politique

Cette thèse porte sur l'évaluation participative et s'inscrit de ce fait dans le domaine de l'évaluation de politique. Bien que la pratique de l'évaluation ne soit pas nouvelle, ce domaine ne s'est établi que récemment au plan professionnel. Les politologues ne sont pas toujours familiers avec le champ couvert par l'évaluation, au contraire des études électorales ou des relations internationales. Il convient par conséquent de présenter les grandes lignes de son histoire et d'en définir la nature.

1.2.1 Une brève histoire de l'évaluation

Cette section offre un bref aperçu du développement historique de l'évaluation des politiques en tant que pratique et profession. Le portrait tracé ici n'est pas exhaustif mais schématique; il se contente de brosser à grands traits l'évolution de cet outil de pilotage de l'action publique.

Au-delà des comptes-rendus humoristiques à saveur créationniste sur ses origines divines ou plutôt diaboliques (voir Encadré 1), force est de constater que l'évaluation possède une généalogie riche et ancienne. Certains soutiennent que la naissance de l'évaluation remonte à plus de 4000 ans, en Chine, alors qu'était utilisé un système relativement sophistiqué pour évaluer et sélectionner les fonctionnaires de l'Empire (Guba & Lincoln, 1989; Patton, 2008; Shadish, Cook, & Leviton, 1991). D'autres estiment plutôt que l'histoire de l'évaluation est indissociable d'une tradition scientifique « appliquée » remontant à l'Antiquité qui cherche à comprendre les problèmes publics et à agir sur ceux-ci (Rossi et collab., 2004). Pour d'autres encore, les origines de l'évaluation sont plus récentes et remonteraient aux grandes transformations économiques et sociales engendrées par la Révolution industrielle (Goyette, 2009).

Une histoire relativement récente

Les racines de l'évaluation de programme sont anciennes, certes, mais l'évaluation telle qu'elle se pratique aujourd'hui possède des origines beaucoup plus récentes. Selon un récit dominant du champ, l'évaluation « moderne » se serait principalement développée aux États-Unis entre 1930 et la fin de la Seconde Guerre mondiale (Alkin & Christie, 2004; Goyette, 2009; Guba & Lincoln, 1989; Mathison, 2008; Rossi et collab., 2004; Shadish et

collab., 1991; Shadish & Luellen, 2004) et ce, même si on a noté la réalisation d'évaluations avant cette période dans d'autres pays, notamment en Suède (OCDE, 1999). Sans aller jusqu'à soutenir que l'évaluation est une « invention » exclusivement états-unienne, il faut néanmoins reconnaître le rôle central qu'a joué ce pays dans le développement de cet outil (Shadish & Luellen, 2004).

Encadré 1 : La Genèse de l'évaluation

Au commencement Dieu créa le Paradis et la Terre.

Dieu se retourna alors et, voyant ce qu'Il avait créé, proclama « Cela est très bon ». Nous étions au soir du sixième jour.

Et au septième jour, Dieu se reposa du travail accompli.

Un de Ses archanges vint alors Lui demander, « Dieu, comment savez-Vous que ce que Vous avez créé est "très bon"? Sur quelles données fondez-Vous Votre jugement? Quels résultats espériez-Vous atteindre exactement? Et n'êtes-Vous pas un peu trop près de la situation pour porter un jugement juste et impartial? »

Dieu réfléchit à ces questions tout le septième jour et cela perturba grandement Son repos. Au huitième jour, Dieu déclara « Lucifer, va en Enfer ».

Ainsi naquit l'évaluation dans la gloire du Feu divin.

—L'histoire véritable du Paradis perdu *selon Halcom*

Source : Patton, M. Q. (2008). *Utilization-focused evaluation* (4e éd.), p. 1 : traduction.

Durant la période d'après-guerre, l'évaluation s'est ensuite développée à un rythme accéléré. Cette période a été qualifiée de « florissante » (Shadish & Luellen, 2004), voire de « boom » (Rossi et collab., 2004) en matière évaluative. On assiste alors à l'institutionnalisation progressive de l'évaluation en tant que pratique et à sa diffusion dans plusieurs pays du monde :

> By the end of the 1950s, program evaluation was commonplace. Social scientists engaged in assessments of delinquency prevention programs, psychotherapeutic and psychopharmacological treatments, public housing programs, educational activities, community organization initiatives, and numerous other initiatives. Studies were undertaken not only in the United States, Europe, and other industrialized countries but also in less developed nations. Increasingly, evaluation components were included in programs for family planning in Asia, nutrition and health care in Latin America, and agricultural and community development in Africa (Freeman, Rossi, and Wright, 1980; Levine, Solomon, and Hellstern, 1981). (Rossi et collab., 2004, p. 9)

Deux facteurs en particulier ont contribué à l'essor de l'évaluation.[10] Premièrement, la multiplication des programmes publics et la hausse importante des dépenses à caractère social qui ont significativement accru les besoins informationnels de l'État quant à sa performance ont fourni un terreau fertile à l'émergence de l'outil de l'évaluation (Goyette, 2009; OCDE, 1999; Rossi et collab., 2004; Shadish et collab., 1991). Aux États-Unis, des initiatives telles que la *Great Society* et la *War on Poverty* lancées par les présidents Kennedy et Johnson sont d'ailleurs fréquemment mentionnées à ce titre (Rossi et collab., 2004; Shadish et collab., 1991).

Le développement important des sciences sociales est un second facteur ayant alimenté l'essor de l'évaluation (OCDE, 1999; Rossi et collab., 2004; Shadish & Luellen, 2004). Il s'agit en effet d'une phase de consolidation pour l'évaluation sur le plan technique (Goyette, 2009). Le progrès dans les techniques de sondage et les procédures statistiques avancées ont permis la réalisation d'évaluations plus rigoureuses et plus sophistiquées (Rossi et collab., 2004). Sur le plan sociologique, on constate un bond important du nombre de diplômés des cycles supérieurs de diverses disciplines pouvant répondre aux besoins croissants de l'État en matière d'évaluation :

> Universities were a more fertile source of evaluators and, eventually, of the training programs to produce evaluators with the skills that were relevant to social program evaluation. Academicians clearly had pertinent expertise in social science methods, so many graduates of professional schools and social science departments were drawn to work in the field of evaluation. With the increase in employment opportunity provided by federal, state, and local evaluation funds, graduate schools experienced an influx of students seeking professional training in social science disciplines, including economics, education, political science, psychology, and sociology. U.S. Census data indicated an 895% increase in doctoral production, from 1469 in 1950 to 13,153 in 1986. Many of these social scientists went into evaluation work either part or full-time. (Shadish & Luellen, 2004, s.p.)

Alors que la croissance dans les dépenses de programmes a alimenté le développement de l'évaluation durant les Trente Glorieuses, c'est paradoxalement la lutte aux déficits publics qui a joué ce rôle dans les années 1980 (OCDE, 1999). L'évaluation était alors utilisée comme outil pour réaliser des économies et améliorer l'efficience de l'action publique (Bemelmans-Videc, 1989). Sur ce point, Varone et Jacob (2004) soutiennent :

---

[10] Parmi les autres facteurs mentionnés dans la littérature, notons également le mouvement de professionnalisation de l'administration publique (Rossi et collab., 2004).

> … [L]'évaluation peut être instrumentalisée pour opérer et légitimer des réductions budgétaires dans certains secteurs, en fonction d'analyses coûts-bénéfices des politiques publiques concernées. En situation de récession économique et de déficits publics récurrents, les informations livrées par l'évaluation se voient parfois valorisées lors du processus budgétaire ou, à tout le moins, lors de l'engagement de diverses dépenses publiques. Cette stratégie de rationalisation financière, pilotée dans les années 1980 par les ministères des finances et d'autres gardiens du budget, caractérise notamment les situations de l'Australie, du Canada, du Royaume-Uni, de la Norvège et des Pays-Bas. (p. 286)

Plus récemment, dans les années 1990 et 2000, les mouvements de la Nouvelle Gestion publique (dans sa variante de gestion axée sur les résultats) et des politiques fondées sur les données probantes (*evidence-based policy*) ont également contribué à asseoir la fonction d'évaluation dans les administrations publiques des pays développés (Hansen & Rieper, 2009; OCDE, 1999). Après tout, l'évaluation est un produit de connaissance qui fournit des données probantes sur les résultats attendus d'une politique :

> La recherche des résultats est, dans les pays de l'OCDE, un élément principal des récentes réformes du secteur public. L'évaluation joue un rôle dans ce contexte parce qu'elle apporte des informations en retour sur l'efficacité, l'efficience et la performance des politiques qui se rapportent au secteur public, et elle peut être la clef de l'amélioration de ces politiques et d'innovations. En un mot, elle contribue à mettre en place une gestion publique fiable. (OCDE, 1999, p. 4)

L'évaluation constitue ainsi une pratique en plein essor depuis la Deuxième Guerre mondiale. Force est de constater que cet essor, loin de s'essouffler, se consolide et s'affirme de nos jours (Leeuw, 2009; OCDE, 1999; Patton, 2008).

Les générations d'évaluation

La section précédente a retracé dans ses grandes lignes l'évolution de l'évaluation en tant que pratique. Une perspective alternative consiste à examiner le développement de l'évaluation sur le plan des idées. Guba et Lincoln (1989) ont développé un schème de classification qui décrit l'évolution récente de l'évaluation en termes paradigmatiques. Ce schème a été repris et adapté par plusieurs auteurs (p. ex., Baron & Monnier, 2003; Goyette, 2009). Dans sa version originale, le schème comprend quatre générations (Goyette, 2009; Guba & Lincoln, 1989; Lincoln, 2004).[11]

[11] Certains auteurs ont en effet plaidé pour l'ajout d'une 5e génération (p. ex., Baron & Monnier, 2003) mais notre propos se concentre sur le schème original.

La première génération, qui remonte au début du 20e siècle, est celle de la *mesure* des apprentissages à l'aide de tests standardisés. Le rôle de l'évaluateur est celui d'un technicien consistant à observer les résultats (des élèves, des militaires, etc.) dans certains domaines (réussite scolaire, intelligence, etc.). L'accent est mis sur la mesure de ce que l'on sait vrai, sur les « faits » (Guba & Lincoln, 1989, p. 23).

La seconde génération, axée sur la *description*, prend forme dans les années 30 et 40 sous l'influence de Ralph Tyler et de sa fameuse Eight Year Study (Marvin C. Alkin, 2004; Guba & Lincoln, 1989). Le coffre à outils de l'évaluateur n'est alors plus limité aux tests standardisés. Parallèlement, le rôle de l'évaluateur s'élargit pour inclure une description des forces et faiblesses du programme par rapport aux objectifs visés :

> These desired learning outcomes were labeled *objectives*. Tyler was engaged to carry out the same kind of work with the Eight Year Study secondary schools, but with one important variation from conventional evaluation (measurement): the purpose of the study would be to refine the *developing curricula and make sure they were working*. Program evaluation was born. (Guba & Lincoln, 1989, p. 28)

La troisième génération se développe dans la période s'étirant de la Deuxième Guerre mondiale aux années 70. Le paradigme évaluatif qui s'y développe est celui du *jugement*. L'évaluateur demeure un technicien et un descripteur des forces et faiblesses du programme mais il assume en plus un rôle normatif explicite (Goyette, 2009; Guba & Lincoln, 1989). Il s'agit en effet de juger de la valeur d'un programme à partir de critères explicites qui ne se résument pas nécessairement aux objectifs des concepteurs. Au niveau des devis et des méthodes, le coffre à outils de l'évaluateur se diversifie mais on constate toutefois que les devis expérimentaux et quasi expérimentaux associés à une épistémologie néopositiviste y occupent une large place. C'est en outre au cours de cette période que l'évaluation se professionnalise et s'institutionnalise (Goyette, 2009).

La quatrième génération, axée sur la *négociation*, est le fruit d'une remise en cause de l'épistémologie néopositiviste de la génération précédente (Goyette, 2009). Si le niveau de sophistication des méthodes et outils scientifiques développés précédemment est indéniable, Guba et Lincoln (1989) critiquent notamment la proximité malsaine qui s'est établie entre décideurs et évaluateurs, l'incapacité des évaluateurs à prendre en compte le pluralisme des valeurs et un engagement contreproductif envers le paradigme scientifique.

Le rôle de l'évaluateur change radicalement. L'évaluation doit être réalisée dans un contexte naturel et non contrôlé[12] et doit répondre aux intérêts et préoccupations des parties prenantes. Le rôle de l'évaluateur devient celui d'un facilitateur du dialogue entre les diverses parties prenantes du programme, un dialogue ayant pour objectif de permettre aux participants de développer une signification commune, une « construction » plus éclairée sur la valeur du programme.

Le compte-rendu schématique de l'évolution de l'évaluation que nous avons présenté jusqu'ici a contribué à esquisser un portrait, fragmentaire il est vrai, de la nature de cette pratique. Au-delà des balises très larges posées jusqu'ici, est-il possible de définir avec plus de précision la nature de l'évaluation?

1.2.2 Définir la nature de l'évaluation : mission impossible?

Définir avec précision la nature de l'évaluation n'est pas chose aisée. Le concept d'évaluation est en effet polysémique (Jacob, 2004). De même, les auteurs d'un document de référence de l'Organisation de coopération et de développement économiques (OCDE, 1999) sur l'évaluation affirment :

> Il n'existe pas de consensus général sur ce qu'est l'évaluation. Il existe de multiples définitions de ce concept, souvent contradictoires. Ce flou se reflète dans la diversité des disciplines (économie, études politiques et administratives, statistiques, sociologie, psychologie, etc.), des institutions et des praticiens opérant dans ce domaine et dans le large éventail de questions, de besoins et de clients auxquels répond l'évaluation. (p. 13)

L'absence de consensus sur la signification de l'évaluation implique-t-elle qu'il faille renoncer à la conceptualiser? C'est l'opinion de certains constructivistes pour qui définir la nature de l'évaluation constitue carrément une entreprise futile :

> … [W]e will argue that there is *no* "right" way to define *evaluation*, a way that, if it could be found, would forever put an end to argumentation about how evaluation is to proceed and what its purposes are. We take definitions of evaluation to be mental constructions, whose correspondence to some "reality" *is not* and *cannot be*

[12] L'évaluateur est alors amené à utiliser la méthode ethnographique : « Fieldwork is the hallmark of research for ethnographers. The method essentially involves working with people for long periods of time in their natural setting to see people and their behavior with all the real-world incentives and constraints. This **naturalistic** approach avoids the artificial response typical of controlled or laboratory conditions » (Fetterman, 2000, s.d.).

> an issue. There is no answer to the question, "But what is evaluation really?" and there is no point in asking it. (Guba & Lincoln, 1989, p. 21)

Cette position constructiviste et relativiste ne résiste cependant pas à une analyse rigoureuse. Personne ne conteste sérieusement l'idée que les concepts sont des constructions mentales, que ces constructions évoluent dans le temps et qu'elles ne font pas consensus. À titre d'exemple, le concept de démocratie est polysémique et n'a certainement pas le même sens aujourd'hui qu'il avait lors de sa création en Grèce antique. Il en va de même pour l'évaluation mais cela ne signifie pas qu'il faille renoncer à définir le concept, bien au contraire. Toutes les conceptualisations ne se valent pas sur le plan de la clarté, de la précision, de la capacité de différentiation, de la parcimonie, de la cohérence, de l'utilité, etc. (Gerring, 1999; Sartori, [1984], 2009). En refusant de fixer le sens de leurs concepts, ces constructivistes se protègent contre la critique. Il s'agit d'une stratégie d'immunisation qui légitime l'utilisation de la rhétorique et du verbiage. Dans un contexte certes un peu différent, Jacques Bouveresse (1998) dénonce d'ailleurs avec verve cette posture de l'imprécision adoptée par plusieurs intellectuels :

> La question cruciale que l'on est obligé de se poser ici est évidemment de savoir comment l'exigence de précision a pu devenir à ce point, dans l'esprit de la plupart de nos intellectuels, l'ennemie numéro un de la pensée authentique. C'est une banalité de dire qu'un souci exagéré de la précision peut constituer un obstacle à la découverte et à la création intellectuelle. Mais cela n'autorise aucunement à transformer une condition nécessaire en une condition suffisante et à croire qu'il suffit de penser de façon vague, approximative et rhétorique, pour être certain de le faire de façon créatrice et profonde.

Puisque les concepts sont essentiels à la théorisation, à la recherche empirique et à la communication efficace entre scientifiques, une mauvaise conceptualisation nuit très certainement à l'accumulation des connaissances (Daigneault & Jacob, à paraître, 2012; Goertz, 2006; Sartori, [1984], 2009; Schedler, 2011). Il est par conséquent essentiel de justifier les choix conceptuels, contrairement à ce que les relativistes soutiennent.

Définir un concept implique de mettre au jour sa structure interne, soit les éléments qui le composent (dimensions, attributs, etc.) et la logique qui unit ces éléments entre eux. C'est vers cette tâche que nous nous tournons maintenant, tout de suite après avoir identifié le principal défi auquel est confronté celui ou celle voulant conceptualiser l'évaluation, soit sa nature transdisciplinaire.

1.2.3 Premiers repérages sémantiques dans un contexte transdisciplinaire

L'absence d'ancrage disciplinaire exclusif contribue à la fragmentation des connaissances, compliquant ainsi la délimitation du concept d'évaluation et de son champ d'application. Goyette (2009) soutient à cet égard : « … il y a quasiment autant d'histoires de l'évaluation que de gens qui la racontent. Ainsi, les représentations de l'évaluation varient en fonction des avancées des différentes disciplines (économie, science politique, sociologie, administration de la santé, etc.) ou des impératifs de certains secteurs » (p. 30).

L'évaluation a été qualifiée de *transdiscipline* par Scriven, au même titre que la statistique, la logique ou la psychométrie. Le terme transdiscipline désigne « …une discipline qui se focalise sur les enjeux d'une autre discipline mais qui a elle-même les attributs d'une discipline » (s.a., 2005). Les théoriciens, chercheurs et praticiens de l'évaluation proviennent en effet de disciplines aussi diverses que les sciences de l'éducation, la psychologie, l'épidémiologie, la science politique, le travail social, l'administration publique et l'économie (Goyette, 2009; Jacob, 2010; Mathison, 2008; Patton, 2008; Rossi et collab., 2004; Shadish et collab., 1991). Il n'y a pas ou très peu de départements universitaires offrant des programmes de formation entièrement consacrés à l'évaluation.

Ceci étant dit, l'évaluation est une profession qui s'institutionnalise progressivement. Le programme de titres professionnels (évaluateur accrédité) récemment lancé par la Société canadienne d'évaluation témoigne de ce mouvement, tout comme la multiplication des associations professionnelles consacrées à l'évaluation à travers le monde au cours des dernières années (Jacob & Boisvert, 2010; Patton, 2008). Les évaluateurs peuvent en outre compter sur des normes éthiques et de pratique professionnelle bien établies et sur des publications diversifiées (Bickman & Reich, 2004; Dubois & Marceau, 2005). L'institutionnalisation de l'évaluation s'effectue également au plan théorique où on note le développement d'un corpus commun de concepts et de cadres théoriques (Dubois & Marceau, 2005; Mathison, 2008; Shadish, 1998; Shadish et collab., 1991).

1.2.4 Évaluer, c'est porter un jugement sur la valeur

Au-delà des frontières disciplinaires, c'est d'abord la dimension normative qui constitue le fondement théorique du champ de l'évaluation, comme l'affirment Shadish, Cook et Leviton (1991) :

> Without its unique theories, program evaluation would be just a set of loosely conglomerated researchers with practical allegiances to diverse disciplines, seeking to apply social science methods to studying social programs. Program evaluation is more than this, more than applied methodology. Program evaluators are slowly developing a unique body of knowledge that differentiates evaluation from other specialties while corroborating its standing among them. Evaluation is diverse in many ways, but its potential for intellectual unity is in what Scriven calls "the logic of evaluation" […] which might bridge disciplinary boundaries separating evaluators. (p. 31)

Au sens général et peut-être aussi le plus fondamental, évaluer consiste en effet à porter un jugement sur la valeur d'un objet (Jacob, 2010; Québec (Province). Secrétariat du Conseil du trésor. Sous-secrétariat aux politiques budgétaires et aux programmes, 2002; Rossi et collab., 2004). Étymologiquement, le terme *évaluation* dérive d'ailleurs de l'ancien français *esvaluer* qui signifie estimer la valeur. Il n'est donc pas surprenant que la dimension normative soit centrale dans la plupart des définitions de l'évaluation, par exemple : « [l]es évaluations de programme sont des *appréciations* analytiques systématiques concernant les principaux aspects d'un programme et sa *valeur* et qui s'attachent à fournir des conclusions fiables et utilisables » (OCDE, 1999, p. 15: italiques ajoutés); et « [t]he common definition of evaluation is 'a systematic inquiry into the *worth* or *merit* of an object' (Sanders, 1994) » (cité dans Forss, Rebien, & Carlsson, 2002, p. 32 : italiques ajoutés). De son côté, Weiss (1972) souligne que « l'évaluation est un terme élastique qui s'étire pour couvrir des *jugements* de différentes sortes » (dans Vedung, 1997, p. 3: traduction et italiques ajoutés). La définition générale contenue dans l'*Encyclopedia of Evaluation* insiste elle aussi sur la dimension normative de l'évaluation :

> Evaluation is an applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs, value, merit, worth, significance, or quality of a program, product, person, policy, proposal, or plan. Conclusions made in evaluations encompass both an empirical aspect (that something is the case) and a normative aspect (judgment about the value of something). **It is the value feature that distinguishes evaluation from other types of inquiry**, such as basic science research, clinical epidemiology, investigative journalism, or public polling. (Fournier, 2004: caractères gras ajoutés)

Michael Scriven (1974), l'un des « pères fondateurs » du champ, soutient lui aussi que le jugement est ce qui définit l'évaluation: « Evaluation research must produce as a conclusion exactly the kind of statement that social scientists have for years been taught is illegitimate: a judgement of value, worth, or merit » (dans Shadish et collab., 1991, p. 75).

Un bref aparté est ici nécessaire sur la distinction entre la valeur intrinsèque (*merit*) d'un objet, c'est-à-dire ses qualités propres, indépendantes du contexte, et sa valeur extrinsèque (*worth*), soit sa valeur, son utilité dans un contexte donné. Pour illustrer cette distinction, imaginons un instant que quelques individus se retrouvent isolés sur une île déserte suite à un naufrage, sans moyen de communiquer avec l'extérieur. Imaginons également que l'un des naufragés est le plus grand artiste peintre de son époque, ce qui dénote une valeur intrinsèque exceptionnelle du point de vue artistique. Dans ce contexte particulier, en revanche, la valeur extrinsèque des compétences professionnelles de l'artiste est certainement moins intéressante que celles d'un pêcheur ou d'un menuisier. Il en va de même pour tout objet d'évaluation, que l'on parle d'une personne, d'un produit ou d'une politique.

Qu'il s'agisse de porter un jugement sur la valeur intrinsèque ou extrinsèque d'un objet, l'évaluateur suit toujours – de manière plus ou moins explicite – une logique similaire. Cette logique, qui a été mise au jour par Scriven, se décline en quatre étapes : « …(a) selecting criteria of merit that something must do to be good, (b) setting standards of performance about how well it must do on the criteria, (c) measuring performance on each criterion and comparing it to standards, and (d) synthesizing results into a value statement » (Shadish et collab., 1991, p. 48).

La citation précédente sur le processus évaluatif, ainsi que les définitions proposées par les auteurs du champ, témoignent de la nature fondamentalement normative de l'évaluation. La conception qu'ont les évaluateurs du rôle de l'évaluation corrobore par ailleurs ce constat. Un sondage réalisé auprès d'un échantillon de membres d'un groupe thématique de l'American Evaluation Association sur l'utilisation de l'évaluation a en effet révélé que près de 80 % des répondants sont en accord ou fortement en accord avec la fonction normative de l'évaluation (Preskill & Caracelli, 1997, p. 215).

Si, en définitive, l'évaluation consiste à juger de la valeur d'un objet, cela n'en épuise pas le sens. Il s'agit maintenant de préciser la nature de l'évaluation en examinant les objets sur lesquels est appliqué le jugement.

1.2.5 La politique publique comme objet

Puisqu'un jugement peut être porté sur divers objets – une personne, un produit, une nouvelle technologie, une œuvre d'art – une première façon de circonscrire la nature de l'évaluation consiste à définir l'objet auquel elle s'applique, soit, dans le cas présent, la politique publique. Les écrits sur les politiques publiques en proposent plusieurs définitions (voir p. ex., Howlett & Ramesh, 2003; Lemieux, 2002; Pal, 2001; Thoenig, 2004). La définition générale proposée par Lemieux (2002) est complète, c'est-à-dire qu'elle contient tous les éléments constitutifs d'une politique – les acteurs et leurs activités, les problèmes et les solutions – tout en ayant le mérite d'être simple : « Les politiques publiques consisteraient donc en un ensemble d'activités (ou de non-activités) par des acteurs politiques, visant à apporter des solutions à des problèmes » (Lemieux, 2002, p. 6).

Deux éléments de cette définition appellent toutefois à discussion. D'une part, il est utile de préciser que les politiques ne visent que les problèmes *publics*, c'est-à-dire ceux qui effectuent un passage réussi de la sphère privée à la sphère publique (Sheppard, 2004). Pour qu'ils soient considérés publics, les problèmes doivent être perçus comme tels « …par les acteurs qui sont les maîtres de l'ordre du jour gouvernemental » (Lemieux, 2002, p. 5). D'autre part, le terme *acteurs politiques* est imprécis et inapproprié. Dans une acception courante mais restrictive, ce terme désigne les élus et leur personnel politique. La définition est alors trop discriminante car elle ne permet pas de rendre compte des politiques qui sont le fait des acteurs administratifs. Dans une acception large puisant au paradigme de la gouvernance, les partis politiques, groupes d'intérêt, experts et citoyens sont considérés comme des acteurs politiques à part entière (Jacob & Daigneault, 2011). Comme le rappelle Lemieux (2002), ils sont parties prenantes aux décisions et non-décisions à la base des politiques publiques. Si la première acception manque de pouvoir discriminant, la seconde manque en revanche de mordant analytique. Une définition plus large de l'acteur politique rend en effet poreuse la frontière entre politique publique et action collective, lorsqu'elle ne la fait pas complètement disparaître. À titre d'exemple, une corvée de nettoyage d'un parc

organisée par des résidents d'un quartier est-elle une politique publique? Il importe de souligner que l'ouverture du processus de politiques publiques à de « nouveaux » acteurs dépend toujours d'une décision plus ou moins explicite des gouvernants qui sont détenteurs de l'autorité publique. Pour des raisons qui doivent maintenant être évidentes, nous proposons de substituer l'expression *autorités publiques* à celle d'*acteurs politiques*. Si cette décision terminologique et sémantique n'est pas neutre sur le plan de l'angle d'analyse[13], elle permet cependant de circonscrire de manière appropriée l'objet que constitue la politique publique.

Par ailleurs, certains établissent une distinction entre politiques, stratégies, programmes, projets, mesures, interventions, initiatives, etc. Il est vrai que ces objets présentent des différences en termes de portée (p. ex., la portée de la politique est plus large que celle du programme). Au-delà de cette différence de portée, cependant, tous ces objets se conforment à la définition de politique publique présentée précédemment. Un programme ou une mesure peuvent prétendre au statut de « politique publique » parce qu'ils sont des activités initiées par des autorités publiques et visant à apporter des solutions à des problèmes publics. Nous utiliserons donc ces termes de manière interchangeable.

1.2.6 Le recours à la méthode scientifique

En définissant l'évaluation comme une pratique consistant à porter un jugement sur la valeur d'une politique publique, certaines balises contribuant à en délimiter le sens ont déjà été établies. Ceci dit, les politiques peuvent faire l'objet de jugements plus ou moins sophistiqués de la part d'une variété d'acteurs incluant les tribunaux, les élus et le personnel politique, les groupes d'intérêt et les citoyens (Howlett & Ramesh, 2003). Pour prendre un exemple d'actualité, deux clients discutant dans un café de la pertinence d'une enquête publique sur la collusion et la corruption dans le domaine de la construction posent un jugement sur la politique du gouvernement québécois dans ce domaine. Si les observations de Monsieur et de Madame Tout-le-monde sur les politiques publiques sont parfois assimilées à de l'évaluation, c'est que le terme est utilisé de manière inappropriée. En effet,

[13] « En limitant les politiques publiques aux actions des autorités gouvernementales, on adopte volontairement ou involontairement le point de vue de ceux qui ont à gérer les politiques publiques, ce qui a été reproché avec raison aux spécialistes des organisations qui ont, de la même façon, adopté de façon trop exclusive le point de vue de ceux qui occupent des postes de direction et ont à gérer des organisations » (Lemieux, 2002, p. 4).

« … les sociétés contemporaines mobilisent abondamment le concept d'évaluation au risque, parfois, de le dénaturer. Le spectre qui s'étend de l'évaluation spontanée à l'évaluation scientifique est très large » (Jacob, 2010, p. 262).

Or, c'est la base informationnelle objective, valide et fiable du jugement porté qui fonde la spécificité de l'évaluation en tant que champ d'étude et de pratique professionnelle. Le jugement porté par l'évaluateur n'est en effet ni arbitraire, ni fondé sur l'opinion ou la croyance mais sur les méthodes, outils et techniques des sciences sociales. À ce sujet, les termes employés dans les définitions précédentes insistent sur la nature scientifique de l'évaluation : « … appréciations analytiques systématiques… », « … conclusions fiables… » (OCDE, 1999, p. 15); « … applied inquiry process for collecting and synthesizing evidence that culminates in conclusions about the state of affairs… », « [c]onclusions made in evaluations encompass […] an empirical aspect… » (Fournier, 2004, s.p.). D'autres auteurs associent de manière encore plus directe l'évaluation à la méthode scientifique : « **Program evaluation** is the use of social research methods to systematically investigate the effectiveness of social intervention programs… » (Rossi et collab., 2004, p. 16). De la même manière, Jacob (2010) affirme que l'évaluation « … implique une démarche méthodologique, transparente et reproductible » (p. 258).

La dimension normative n'est donc pas suffisante pour que l'on soit en présence d'évaluation. Comme le souligne Scriven (1991, cité dans Rossi et collab., 2004, p. 17), l'évaluation possède deux « bras » : l'un consistant à juger, l'autre à colliger et à analyser des données. Les deux composantes sont nécessaires pour définir la nature de l'évaluation. Stake soutient la même position : « Both description and judgment are essential—in fact, they are the two basic acts of evaluation » (cité dans Guba & Lincoln, 1989, p. 30).

Sans tomber dans une régression à l'infinie où chaque terme qui entre dans une définition doit être défini à son tour, il convient néanmoins de circonscrire ce qu'on entend par la *méthode scientifique*. Nous adoptons ici la conception postpositiviste[14] de la recherche

[14] Le postpositivisme partage avec le positivisme une ontologie réaliste selon laquelle il existe une réalité indépendante de l'observateur et à laquelle il est possible d'accéder à travers l'observation et l'expérimentation. Le premier se distingue toutefois du second par ses positions épistémologiques moins naïves sur la possibilité de découvrir des lois vraies et universelles (c.-à-d., on peut démontrer qu'une théorie est fausse mais pas qu'elle est vraie), sur la neutralité de la mesure (c.-à-d., toute observation est chargée de

scientifique en sciences sociales proposée par King, Keohane et Verba (1994, pp. 7-9). Cette conception, qui s'applique autant à la recherche quantitative que qualitative, est fondée sur quatre caractéristiques : (1) le but de la science est l'inférence; (2) ses procédures sont publiques; (3) ses conclusions sont incertaines, et; (4) son contenu est sa méthode. De ces caractéristiques découle l'implication selon laquelle la science est une entreprise collective. Chaque scientifique a en effet besoin de ses pairs pour évaluer, reproduire, réviser et étendre la portée de ses recherches. La production scientifique doit par ailleurs être jugée sur les critères de pertinence scientifique et sociale, de fidélité et de validité (King, Keohane & Verba, 1994; Shadish, Cook & Campbell, 2002). Si la définition de la science offerte ici n'est en aucun cas exhaustive, elle offre toutefois des balises suffisantes à la compréhension de l'argumentation qui est présentée ici.

1.2.7 Informer la prise de décision

Si on récapitule, l'évaluation cherche à porter un jugement sur les politiques publiques en se fondant sur les méthodes et les standards des sciences sociales. La question qui nous reste maintenant à examiner pour bien circonscrire le concept d'évaluation est celle de ses finalités et de son contexte d'application.

Il faut préciser d'emblée que l'évaluation de politique est, à l'instar de la médecine et de l'ingénierie, un champ orienté vers la pratique (Shadish et collab., 1991). Le rôle de l'évaluateur consiste à fournir des réponses à des questions pratiques telles que celles-ci :

- Quelle est la nature du problème public? Qui affecte-t-il et de quelle manière? Est-ce la responsabilité de l'État d'agir pour le solutionner? Quelles sont les ressources nécessaires pour le résoudre?
- Le programme public est-il conçu adéquatement pour agir sur les causes du problème? Sa théorie est-elle cohérente et plausible?
- Le programme est-il mis en œuvre tel qu'attendu? La prestation des services est-elle fidèle au plan original? Pourquoi ou pourquoi pas? Est-il nécessaire d'adapter le programme aux circonstances locales?
- Le programme est-il efficace? En d'autres termes, atteint-il ses objectifs? Les changements observés sont-ils bel et bien attribuables au programme? Quels sont les effets non prévus, désirables et indésirables, qu'il induit?

---

théorie) et sur le caractère unitaire et indépendant de la réalité (c.-à-d., plusieurs perspectives peuvent coexister à propos de cette réalité) (Rubin et Rubin, 2011, p. 19).

- Le programme est-il rentable? Y a-t-il un programme alternatif qui permette d'obtenir les mêmes résultats à moindre coûts?

L'évaluation est une activité résolument orientée vers l'action et la résolution de problèmes : « Evaluation is just one part of a complex, interdependent, nonlinear set of problem-solving activities » (Shadish et collab., 1991, p. 21). Rossi, Lipsey et Freeman (2004) ajoutent : « …the role of evaluation is to provide answers to questions about a program that will be used and will be actually used. This point is fundamental to evaluation—its purpose is to inform social action » (p. 20).

Plus précisément, quatre fonctions principales sont généralement attribuées à l'évaluation (Rossi et collab., 2004; Vedung, 1997). La première fonction, *sommative*, est orientée vers la reddition de comptes et l'imputabilité. Cette fonction de contrôle consiste à s'assurer que la performance du programme est conforme aux attentes. On cherche ainsi à s'assurer que le programme est mis en œuvre de manière appropriée (p. ex., que les fonds sont dépensés et que les services sont fournis tel que prévu) et qu'il atteint ses objectifs à l'intérieur des paramètres prescrits. L'information générée par l'évaluation est alors généralement destinée aux décideurs et leur permet de prendre une décision éclairée quant à l'avenir du programme : adoption, maintien, réforme ou abandon. Une seconde fonction, *formative*, est orientée vers l'amélioration des programmes publics. L'objectif de l'évaluation est alors de cerner les forces et faiblesses du programme afin de renforcer les premières et de corriger les secondes. L'évaluation formative est d'abord destinée au personnel du programme, soit les gestionnaires du programme et le personnel responsable de la mise en œuvre et de la prestation des services. Troisièmement, l'évaluation peut contribuer au *développement des connaissances*. Une évaluation de la théorie d'action ou de l'impact d'un programme innovateur de prévention du suicide peut faire progresser les connaissances scientifiques sur le sujet. Un autre exemple réside dans l'évaluation d'un examen de la qualité des évaluations fédérales (Daigneault, 2010). L'une des visées de cette évaluation consistait à tirer des leçons de cet exercice et à formuler des recommandations générales sur la pratique de la méta-évaluation qui pourraient être utiles aux évaluateurs et responsables des services d'évaluation du Canada et d'ailleurs. Même lorsque l'évaluation vise la production de connaissances, cependant, ces dernières conservent une orientation tournée vers la résolution de problèmes et l'action.

La quatrième et dernière fonction de l'évaluation constitue en fait une catégorie composite rassemblant plusieurs fonctions *stratégiques* : légitimer une décision déjà prise sur la base d'autres considérations, servir de munition politique pour persuader autrui, temporiser, démontrer la rationalité et la compétence de gestionnaires d'une organisation, etc. Rossi, Lipsey et Freeman (2004) désignent cet ensemble de fonctions par l'expression « agendas cachés » pour bien mettre en exergue la nature politique de ces fonctions de l'évaluation et les comparer avec les trois fonctions précédentes, substantives.

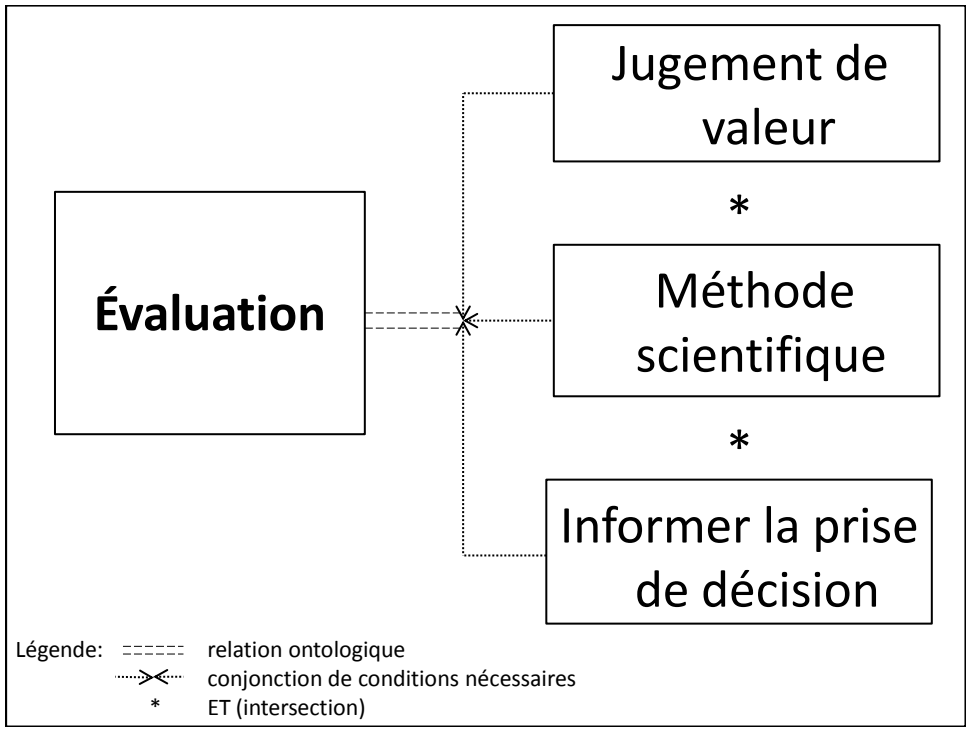
1.2.8 L'évaluation : concept et définition

La discussion précédente visait à clarifier la nature de l'évaluation. Il doit maintenant être clair que l'évaluation de politique comporte trois dimensions constitutives : (1) elle porte un jugement de valeur sur une politique; (2) elle prend appui sur la méthode scientifique; (3) elle possède une orientation pratique qui vise généralement à informer la prise de décision relative à une politique ou à un programme. Ces trois dimensions sont mobilisées de manière plus ou moins explicite par plusieurs auteurs du domaine pour définir la nature de l'évaluation. Garon et Roy (2001) soutiennent par exemple que « [q]uels que soient les modèles utilisés, toute démarche d'évaluation présuppose une procédure en trois étapes, qui sont les fondements sur lesquels elle s'appuie : la recherche, le jugement et la prise de décision (Allard, 1996) » (p. 101). On retrouve les mêmes dimensions dans la définition de l'évaluation proposée par Rossi (2004).

Il est cependant peu courant de retrouver une conceptualisation qui spécifie à la fois les composantes du concept d'évaluation *et* la logique qui unit ces composantes au concept général. Il faut donc remédier à cette lacune. Nous soutenons que chacune des trois dimensions doit être considérée comme une condition *nécessaire* au concept. Cela signifie qu'en l'absence de l'une ou l'autre de ces dimensions, on ne peut véritablement parler d'évaluation. Ces dimensions sont en outre conjointement *suffisantes* au concept, c'est-à-dire qu'aucune autre dimension n'est requise au concept que l'on désigne par l'expression « évaluation des politiques ». Cette conceptualisation à trois dimensions est illustrée à la Figure 2 en utilisant la notation graphique de Gary Goertz (2006). La définition suivante de l'évaluation est d'ailleurs compatible avec la conceptualisation proposée :

> L'évaluation se définit comme une démarche scientifique qui examine de façon systématique et objective les processus, les produits ou les effets d'une politique ou d'un programme public, et qui, en fonction de critères bien définis, porte un jugement sur sa valeur et sa contribution. Elle rassemble de l'information sur les interventions publiques afin de produire des connaissances crédibles, pertinentes et utiles à la conduite de l'action publique. (Jacob, 2010, p. 258)

Figure 2 : Ontologie du concept d'évaluation de politique



1.2.9 Distinguer l'évaluation de pratiques apparentées

Définir un concept ne se résume pas à l'identification de ses caractéristiques fondamentales, il faut également déterminer les relations qu'il entretient avec les concepts voisins (Gerring, 1999; Sartori, 2009 [1984]). Comme le souligne Gerring (1999), « [i]t is impossible to redefine one term without redefining others, for the task of definition consists of establishing relationships with neighboring terms » (p. 382). Il est donc impératif de montrer comment l'évaluation se distingue d'autres pratiques de recherche et de pilotage de l'action publique avec lesquelles elle est souvent confondue. Pour ce faire, les dimensions constitutives identifiées à la section précédente sont très utiles (voir Tableau 2).

Recherche scientifique

Le questionnement à propos de la distinction entre évaluation et recherche scientifique est presque un passage obligé pour les nouveaux évaluateurs qui cherchent à définir leur identité professionnelle. Il s'agit notamment d'un sujet très populaire sur EVALTALK, un forum électronique de discussion consacré à l'évaluation et géré par l'American Evaluation Association (Mathison, 2008).

Tableau 2 : Comparaison entre l'évaluation et diverses pratiques apparentées

Outil	Jugement de valeur	Méthode scientifique	Informer la prise de décision
Évaluation de politique	Oui	Oui	Oui
Recherche de base	Non	Oui	Non
Recherche appliquée	Non	Oui	Oui
Analyse de politique (*knowledge of*)	Non	Oui	Parfois
Analyse de politique (*knowledge for*)	Oui	Parfois	Oui
Mesure et gestion de la performance	Oui (portée limitée)	Non	Oui
Vérification traditionnelle	Oui (conformité)	En partie	Oui
Vérification de performance (*value for money*)	Oui (les 3 "E")	En partie	Oui

Mathison (2008) affirme que l'évaluation se distingue de la recherche, d'une part, par le recours à certaines méthodes spécifiques au champ et, d'autre part, par des normes d'évaluation de la qualité de la production évaluative qui diffèrent de celles utilisées en recherche. Cette position semble toutefois reposer sur une base argumentaire pour le moins discutable. En premier lieu, les méthodes « spécifiques » au champ de l'évaluation telles que *photovoice* et la « technique du changement le plus significatif » (*the most significant change technique*) peuvent être et sont employées dans le cadre d'études scientifiques plus conventionnelles. En outre, le nombre d'évaluateurs qui connaissent et utilisent ces méthodes est probablement très restreint. On voit donc difficilement comment ces

méthodes pourraient permettre de distinguer efficacement l'évaluation de la recherche. En second lieu, Mathison souligne avec justesse que les critères utilisés pour juger de la qualité des évaluations sont différents de ceux de la recherche. Aux impératifs habituels de validité et de fidélité s'ajoutent en effet des considérations d'utilité et de faisabilité, notamment (Joint Committee on Standards for Educational Evaluation, 1994). Ceci dit, ces standards sont le symptôme et non la cause d'une différence « ontologique » entre l'évaluation et la recherche scientifique. Si on utilise des standards différents, c'est d'abord parce que l'évaluation est orientée vers la pratique, qu'elle cherche à éclairer la prise de décision. Dans ce contexte, où les relations avec les parties prenantes sont inévitables, les critères d'utilité et de faisabilité acquièrent sans contredit une importance fondamentale.

La conceptualisation tridimensionnelle de l'évaluation présentée précédemment est un outil qui permet de distinguer efficacement évaluation et recherche. Si l'évaluation partage avec la recherche le recours à une méthode systématique et transparente pour produire des connaissances valides, elle s'en distingue à un ou deux égards, selon qu'on parle de recherche de base (c.-à-d., fondamentale) ou appliquée (voir Tableau 2). L'évaluation se distingue tout d'abord de la recherche de base par son orientation appliquée. L'évaluation vise en effet à informer la prise de décision relative à un programme ou à une politique. Elle partage ainsi une dimension constitutive avec la recherche appliquée.

Ensuite, l'évaluation se distingue à la fois de la recherche de base et appliquée par sa dimension normative. Alors que les scientifiques évitent les jugements *de* valeur dans le cadre de leurs recherches, le rôle de l'évaluateur consiste au contraire à porter un jugement *sur* la valeur du programme à l'aide de critères normatifs tels que la pertinence, la plausibilité de la théorie, la fidélité de la mise en œuvre, l'efficacité, l'efficience, l'équité et la pérennité (Scriven, 1974, cité dans Shadish et collab., 1991, p. 75)[15].

[15] Selon le contexte et son orientation théorique, l'évaluateur peut s'en tenir aux valeurs des commanditaires et autres parties prenantes de l'évaluation (théorie *descriptive* de la valeur) ou au contraire se servir de valeurs différentes comme la participation démocratique ou l'habilitation (théorie *prescriptive* de la valeur) (Shadish, Cook et Leviton, 1991). Dans tous les cas, il importe d'insister sur le fait que l'évaluateur porte un jugement normatif sur le programme.

Analyse de politique

L'évaluation et l'analyse de politique sont des pratiques relativement similaires amenant plusieurs auteurs à les comparer (Geva-May & Pal, 1999; Lemieux, 2006). La multiplication des termes anglais – *policy analysis*, *policy science*, *policy studies* et *policy evaluation* – et leur traduction vers le français ne fait qu'accroître la confusion. Il convient d'abord de distinguer deux acceptions du concept d'analyse de politique, l'une descriptive et explicative, l'autre prescriptive (Pal, 2001). Comme le souligne Lemieux (2002) : « On peut étudier les politiques publiques pour faire avancer les connaissances ou on peut les étudier principalement pour améliorer l'action. En anglais, on exprime parfois cette distinction en parlant de 'knowledge of' et de 'knowledge for' » (p. 1).

La première acception réfère à l'étude *scientifique* des politiques publiques et ne constitue en ce sens qu'une variante plus ou moins appliquée de la recherche scientifique (Tableau 2). Ceux qui étudient les politiques publiques visent à produire des recherches en matière de politique qui sont pertinentes (*policy relevant*), espérant ainsi influencer de manière plus ou moins directe la prise de décision. Cette orientation de l'analyse des politiques vers la résolution de problème demeure et ce, même si elle s'est étiolée avec le temps (Howlett & Ramesh, 2003, pp. 3-4). Il faut ainsi retenir de cette première acception que l' « analyse » des politiques correspond à leur étude scientifique.

La seconde acception réfère quant à elle à une pratique résolument orientée vers l'action. Entendue en ce sens, l'analyse de politique consiste à comparer différentes solutions de politiques à un problème public (choix A, B et C), les apprécier et à recommander la meilleure d'entre elles, comme l'indique ce passage passant en revue les écrits portant sur cet outil :

> … [P]olicy analysis is regarded as a politically oriented discipline; it provides power by producing knowledge of and in the policy process (Dunn, 1981: 2); it determines which of the various alternative public or government policies will most achieve a given set of goals in light of the relations between the policies and the goals (Nagel, 1990) and in light of politically feasible courses of action; it generates information and evidence in order to help the policymaker choose the most advantageous action (Quade, 1975: 5); and it identifies and evaluates alternative policies and programs intended to 'lessen or resolve social, economic, or political problems' (Patton and Sawicki, 1986: 21). (Geva-May & Pal, 1999, p. 263)

L'analyse de politique prescriptive partage ainsi avec l'évaluation une orientation normative (juger les alternatives de politiques) et, tout comme elle, est axée vers la prise de décision. Mais qu'est-ce qui différencie ces pratiques? Certains soutiennent que la principale différence entre analyse de politique et évaluation réside dans le moment de leur réalisation : la première serait prospective alors que la seconde serait rétrospective (Geva-May & Pal, 1999, p. 263; OCDE, 1999, p. 13). Cette distinction est cependant loin d'être absolue. D'une part, l'évaluation peut être réalisée à n'importe quel moment du cycle des politiques, comme le démontre l'existence des termes *évaluation prospective* (*ex ante*) et *évaluation concomitante* (*in itinere*) (Jacob, 2010; Mayne, 2006; OCDE, 1999). D'autre part, l'analyse de politique peut parfois être rétrospective. En effet, chercher à influencer les décisions futures en matière de politiques publiques n'est pas incompatible avec l'analyse rétrospective. À titre d'exemple, une analyse des différents scénarios de politiques en matière de frais de scolarité universitaires peut être fondée sur les initiatives passées des gouvernements d'ici et d'ailleurs.

La différence fondamentale entre analyse de politique et évaluation se situe plutôt au niveau de la méthode. L'évaluation s'appuie sur les méthodes, outils, techniques et standards des sciences sociales[16] pour porter son jugement tandis que l'analyse de politique prescriptive adhère à des standards qui ne sont pas toujours aussi stricts :

> We noted with Wildavsky that policy analysis is a craft that depends on method and imagination. It is not that an evaluator should not be expected to be creative, use previous knowledge or intuition in coaching the client and formulating evaluation questions and methodologies – but that their data interpretation should stick to very strict research codes in order to be able to present a reliable information base. While the policy analyst should make use of journalistic modes of inquiry, the evaluator should make sure that objective research methods are being used – whether comparative, experimental, case studies, etc. – along with valid and reliable research tools (Weimer and Vining, 1992; Bardach, 1974). (Geva-May & Pal, 1999, p. 264)

L'analyste de politique (au sens prescriptif du terme) laisse en effet une large place à l'intuition, à l'anecdote, aux considérations politiques dans son travail et est souvent guidé par un ordre du jour politique qui colore son traitement de l'information (Geva-May & Pal,

[16] Il importe de souligner l'existence d'approches évaluatives ayant recours à des standards d'une orientation épistémologique et méthodologique différente, notamment constructiviste (voir p. ex., Guba & Lincoln, 1989).

1999). En outre, l'argumentation et l'art de la rhétorique y sont souvent mobilisés[17] afin de persuader décideurs et citoyens du bien-fondé des solutions proposées. Ceci dit, affirmer que l'analyse de politique n'a *jamais* recours à une méthode explicite et systématique serait inexact. La différence avec l'évaluation tient plutôt au fait que l'analyse de politique n'est qu'occasionnellement fondée sur une telle méthode et, surtout, qu'elle ne suit pas toujours les normes de validité associées à la recherche scientifique postpositiviste (Tableau 2).

Mesure et gestion de la performance

L'évaluation de politique se distingue également des indicateurs et tableaux de bord utilisés pour mesurer et gérer la performance des organisations publiques. Tout d'abord, l'évaluation est un exercice réalisé de manière ponctuelle et donc délimité dans le temps, alors que les outils de gestion de la performance fournissent des informations sur une base continue :

> Performance monitoring can be viewed as *periodically* monitoring progress toward explicit short-, intermediate, and long-term results. It also can provide feedback on the progress made (or not made) to decision makers, who can use the information in various ways to improve performance. (H. J. Smith, Kuseck, & Rist, 2004, s.p.: italiques ajoutés)

Comme le souligne cette citation, les deux outils de pilotage ont cependant en commun le fait de chercher à éclairer la prise de décision publique en générant des données évaluatives (voir Tableau 2). Les indicateurs du tableau de bord sont de nature évaluative car ils produisent un jugement sur la performance d'une organisation ou d'un programme (Schwartz & Mayne, 2005a, 2005b). Ici s'arrêtent toutefois les ressemblances entre les deux outils. Contrairement à l'évaluation, les outils de gestion de la performance tel le tableau de bord ne permettent pas de porter un jugement aussi explicite et approfondi sur l'objet évalué que l'évaluation[18]. En outre, si les indicateurs mesurent et permettent d'apprécier un phénomène, on ne peut toutefois pas considérer qu'ils produisent des informations à partir de la méthode scientifique : « L'évaluation est plus exigeante [que la mesure des performances] : elle s'efforce de trouver des explications aux résultats constatés et de comprendre la logique des interventions publiques… » (OCDE, 1999, p. 13). Cela

[17] Il en va de même dans plusieurs travaux en sciences sociales. Il existe en effet des normes et des « technologies littéraires » propres à chaque discipline pour présenter ses travaux de recherche et convaincre ses pairs (Sandelowski & Barroso, 2006).

[18] Puisque la portée de l'évaluation peut varier selon le contexte, il s'agit ici d'une observation générale.

étant dit, les évaluateurs sont heureux lorsqu'ils peuvent compter sur des données provenant de systèmes de gestion de la performance pour réaliser leurs évaluations.

Vérification traditionnelle et de performance

À l'origine, la vérification (ou audit) était une pratique consistant à rendre une opinion d'expert sur des états financiers (Leeuw, 1992). Cette pratique a évolué avec le temps et n'est désormais plus limitée aux questions comptables : « Broadly defined, *auditing* is a procedure in which an independent third party systematically examines the evidence of adherence of some practice to a set of norms or standards for that practice and issues a professional opinion » (Schwandt, 2004, s.p.). On constate un fort engouement pour la vérification et l'agrément au point où certains soutiennent que l'on vit désormais dans la « société de l'audit » (Power, [1997] 2005).

Il importe de distinguer deux variantes à la vérification. La première est la *vérification traditionnelle* qui s'intéresse au respect des lois, règlements et procédures, comptables notamment, dans une perspective de conformité, de régularité et de contrôle (Leeuw, 1992; Mayne, 2006; OCDE, 1999). Comme l'affirme Mayne (2006) : « In general, auditing is *checking*… » (p. 12). La vérification met l'accent sur les ressources financières et leur relation avec les instruments de politique (Leeuw, 1992). La seconde variante est la *vérification de performance* (ou vérification d'optimisation des ressources) dans laquelle le jugement sur la valeur du programme déborde la simple conformité aux règles et procédures. Il s'agit dans ce cas de déterminer si un programme « nous en donne pour notre argent », c'est-à-dire est efficace, économe et efficient (les trois « E »).

Peu importe sa variante, la vérification est normative, orientée vers la prise de décision – la reddition de comptes en particulier – et est fondée sur une méthode systématique (Schwandt, 2004; Schwartz & Mayne, 2005a). Elle partage ainsi de nombreux points communs avec l'évaluation de politique. On note cependant quelques différences.

En ce qui concerne les méthodes et les données, l'évaluateur dispose d'un arsenal méthodologique varié et sophistiqué, comprenant devis expérimentaux et quasi expérimentaux de type quantitatif mais également des méthodes qualitatives telles que l'ethnographie et la théorisation ancrée, qui lui permettent de formuler des inférences

descriptives et causales (Leeuw, 1992; Mayne, 2006; Schwandt, 2004). La vérification a une ambition plus limitée et se concentre principalement sur une preuve documentaire qui est parfois complétée par des entrevues, des sondages et de l'observation. Bien que systématique, la méthode utilisée en vérification semble différer de la méthode scientifique postpositiviste (Tableau 2). Tandis que l'identité et la pratique professionnelle de l'évaluateur restent ancrées dans les sciences sociales, celles du vérificateur demeurent davantage ancrées dans les disciplines de la comptabilité et de la gestion.

Une seconde différence réside dans la portée et l'objet du jugement de valeur porté. Premièrement, la vérification s'appuie sur des critères qui sont généralement préétablis alors que l'évaluateur peut utiliser différents critères (pertinence, fidélité, efficacité, équité, etc.) selon les visées de l'évaluation et les besoins du commanditaire ou client de l'évaluation (Jacob, 2010; Leeuw, 1992). Deuxièmement, la vérification adopte principalement une perspective légale et réglementaire axée sur le contrôle et ce, même lorsqu'elle cherche à apprécier la performance (Leeuw, 1992; OCDE, 1999). Troisièmement, la vérification s'intéresse principalement aux systèmes, procédures et pratiques de gestion et aux extrants des politiques (Mayne, 2006). L'évaluateur peut aller plus loin en amont (besoins et théorie d'une politique) et en aval (effets voulus et non voulus, à court, moyen et long termes).

En somme, la perspective de l'évaluateur est beaucoup plus large que celle du vérificateur. Malgré des différences réelles, il n'en demeure pas moins qu'en pratique l'évaluation et la vérification ne sont pas des pratiques professionnelles totalement imperméables :

> Malgré les différences évidentes entre l'audit traditionnel et l'évaluation, on a posé la question de savoir s'il convenait de rapprocher ces fonctions ou si cette convergence est déjà en train de se produire. Les délimitations commencent à se brouiller : l'audit a étendu son champ aux questions de performances [sic] et, à côté de l'audit financier traditionnel, il existe maintenant un audit des performances. Ce dernier, effectué par exemple en Suède, au Royaume-Uni (audits "value for money", rentabilité) et aux Pays-Bas, est un exemple d'appréciation similaire à l'évaluation de programme. Les méthodes et critères appliqués sont très semblables. On observe aussi des signes de convergence entre les cultures professionnelles, par exemple avec des réseaux communs comme l'*European Evaluation Society* (EES). (OCDE, 1999, p. 19)

Certains ont d'ailleurs noté qu'évaluateurs et vérificateurs font partie du même « réservoir » de compétences que les décideurs peuvent mobiliser selon leurs besoins (voir Schwandt, 2004).

1.3 L'évaluation participative

Maintenant que la nature de l'évaluation a été clarifiée, il convient de s'interroger sur les spécificités d'un courant participatif. Selon le cadre d'analyse de Guba et Lincoln (1989) discuté à la section précédente, les trois premières générations d'évaluation ont été pratiquement muettes sur le rôle des différentes parties prenantes (*stakeholders*) à l'évaluation. Tout au plus a-t-il été fait mention des rôles de commanditaire de l'évaluation et celui de source de données des sujets humains. Au-delà des spécifications de la commande d'évaluation, la responsabilité de la conduite même de l'évaluation était exclusivement dans les mains de l'évaluateur. La quatrième génération est la première où l'on a reconnu le rôle significatif que les décideurs, professionnels de première ligne, bénéficiaires et autres acteurs ont à jouer dans la conduite même de l'évaluation. L'évaluation participative était née.

1.3.1 Une popularité indéniable

L'intérêt pour les approches participatives en évaluation n'est pas nouveau. Si la pratique de l'évaluation participative (ÉP) comme telle remonte aux années 70 (Cousins & Chouinard, à paraître; Cousins & Whitmore, 1998), on lui note des ancêtres sociohistoriques et philosophiques beaucoup plus anciens (Brisolara, 1998). Toutefois, ce n'est que dans les années 80 et 90 que les approches participatives sont réellement devenues populaires (Cousins, Donohue, & Bloom, 1996; Díaz-Puente, Montero, & de los Ríos Carmenado, 2009; Lennie, 2005; Poth, 2008; Preskill, Zuckerman, & Matthews, 2003; Shulha & Cousins, 1997). Le mouvement récent en faveur de la démocratie participative et de la nouvelle gouvernance observé en dehors du champ de l'évaluation a d'ailleurs certainement contribué à la popularité de l'idéal participatif (DeLeon, 1997; Jacob & Daigneault, 2011). Aujourd'hui, la multiplication des termes utilisés pour décrire les approches centrées sur les participants témoigne de la maturité de leur développement (Daigneault & Jacob, 2009; King, 1998; Toal, 2009). La participation est ainsi devenue

l'une des tendances les plus importantes du champ de l'évaluation (Mark, 2001). Cette tendance est d'ailleurs visible autant sur les plans de la théorie que de la pratique.

Sur le plan théorique, le principe participatif est aujourd'hui largement admis dans le monde de l'évaluation (Whitmore, 1998). Comme le souligne Wye (1989), « [e]valuators are increasingly aware of the desirability of client involvement in the evaluation process » (p. 35). La participation est devenue un concept central en évaluation (Carlsson, Ericksson-Baaz, Fallenius, & Lövgren, 1999; Fleischer & Christie, 2009; Poth & Shulha, 2008), concept avec lequel « … chaque évaluateur et théoricien de l'évaluation est d'accord » (Mathison, 2005b, p. xxxiii : traduction). Le consensus autour des approches et techniques participatives est si fort que certains l'ont assimilé à un « article de foi » (Shea & Lewko, 1995, p. 159) ou l'ont qualifié d' « orthodoxie participative » afin d'en dénoncer les effets pervers (Biggs, 1995, cité dans Gregory, 2000, p. 180).

Loin d'être un phénomène passager, l'attrait normatif de la participation semble une tendance lourde dans le domaine de l'évaluation. Morgan (1996) soutient à ce propos : « The issue of stakeholder involvement, responsibility, and collaboration in the evaluation process has long held a central position in evaluation debates (Patton, 1988a; Weiss, 1988a, 1988b; Guba & Lincoln, 1989; Alkin, 1985, 1990) » (p. 35). L'Organisation mondiale de la santé a, dès la fin des années 1990, recommandé que les évaluations d'interventions de promotion de la santé soient participatives (cité dans Springett & Wallerstein, 2008, p. 201).

Cousins, Donohue et Bloom (1996) ont effectué un sondage auprès de 564 répondants actifs en évaluation de diverses associations professionnelles nord-américaines sur le thème de l'évaluation participative. Leurs résultats démontrent la force normative exercée par ce concept. Ainsi, une forte majorité de répondants étaient en accord ou fortement en accord avec les énoncés suivants :

- les personnes responsables de la mise en œuvre des programmes devraient participer à la conduite des évaluations (88 % de répondants, $n = 544$);
- les personnes ayant un intérêt vital pour les programmes (p. ex., concepteurs des programmes, commanditaires, directeurs) devraient participer à la conduite des évaluations (81 %, $n = 542$);

- plus il y a de groupes de parties prenantes impliqués dans l'évaluation, mieux c'est (75 %, *n* = 547). (Cousins et collab., 1996, p. 216: traduction)

Plus récemment, Fleischer et Christie (2009) ont mené un sondage auprès d'un échantillon de membres de l'American Evaluation Association. Les répondants devaient indiquer leur niveau d'accord avec dix énoncés, mesurés sur une échelle allant de « 1 » (fortement en désaccord) à « 5 » (fortement en accord), concernant le rôle de l'évaluateur dans certaines activités évaluatives. Avec 98 % des répondants en accord ou fortement en accord (*n* = 1047), « impliquer les parties prenantes dans le processus évaluatif » est l'activité pour laquelle l'approbation était la plus forte (pp. 164-165 : traduction). Ce chiffre constitue une illustration sans équivoque de l'attrait normatif exercé par la participation sur l'évaluateur contemporain.

La participation exerce par ailleurs une influence prépondérante sur les chercheurs du domaine à titre d'objet d'étude. Cette influence s'est particulièrement manifestée au cours des dernières années (Cousins & Chouinard, à paraître; Cullen et collab., 2011; Jacob & Ouvrard, 2009; King et collab., 2011; Rodrígues-Campos, 2012; Smits, Champagne, & Brodeur, 2011; Toal, 2009), ce qui témoigne de l'intérêt toujours renouvelé de la part des chercheurs du domaine pour la participation.

1.3.2 Une approche évaluative?

Qu'on les désigne par les termes de théories, modèles ou approches, les différentes approches évaluatives (scientifique, axée sur l'utilisation, fondée sur la théorie du programme, etc.) offrent toutes des lignes directrices sur la manière de réaliser une évaluation de qualité :

> [Elles] offrent […] des prescriptions d'orientation pratique qui ont trait à des questions qui concernent la pertinence et l'organisation de l'évaluation, notamment sa visée, ses questions, ses méthodes, le rôle de l'évaluateur et des parties prenantes ainsi que la manière de gérer les contraintes de l'évaluation. (Daigneault, 2011, p. 3)

Si, à plusieurs égards, l'ÉP peut être considérée comme une approche particulière d'évaluation, nous considérons qu'il s'agit plus exactement d'une *classe* ou d'un type d'approches évaluatives. Certes, l'ÉP réfère dans certains cas à une approche évaluative relativement bien définie (voir p. ex., Cousins & Earl, 1992). Cependant, l'ÉP est d'abord

et avant tout un terme général utilisé pour désigner toute évaluation ou approche d'évaluation au sein de laquelle les parties prenantes participent de manière significative (Greene, 1988; King, 2005; Papineau & Kiely, 1996; Rossi et collab., 2004).

1.3.3 Quelques éléments de définition

Une conceptualisation systématique de l'ÉP sera proposée au prochain chapitre, mais il n'est pas inutile de tenter de circonscrire quelque peu le concept dès maintenant à l'aide d'une définition influente : « *Participatory evaluation* is an overarching term for any evaluation approach that involves program staff or participants actively in decision making and other activities related to the planning and implementation of evaluation studies » (King, 2005, s.p.). De même, Cousins et Chouinard (à paraître) soutiennent que ce qui distingue l'évaluation participative est l'engagement des parties prenantes non évaluatives dans l'acte même de la recherche. Il importe par conséquent de déboulonner deux mythes courants à propos de la nature de l'ÉP (King, 2005). Le premier est que toute évaluation dans laquelle les parties prenantes sont impliquées sous une forme ou l'autre – à titre de source de données par exemple – est participative. Le second est que le recours aux méthodes qualitatives dans le cadre d'une évaluation en fait automatiquement une ÉP. Or, comme la définition ci-dessus l'indique, c'est plutôt le type de relation qu'entretient l'évaluateur avec les participants qui détermine si une évaluation est participative ou non.

1.3.4 Deux courants et un cadre d'analyse

Au-delà de la définition précédente, très générale, plusieurs raisons motivent ou justifient le recours à une approche participative et permettent de classer les types d'ÉP en courants (Cousins & Whitmore, 1998; Weaver & Cousins, 2004). Un premier courant, la *practical participatory evaluation* (P-PE), est fondé sur une justification pratique/pragmatique. L'évaluation axée sur l'utilisation (*utilization-focused evaluation*) constitue l'exemple phare de ce courant (Patton, 2008). La participation est alors envisagée dans une perspective de résolution de problèmes, comme pouvant contribuer à l'utilité des connaissances générées par l'évaluation :

> The core premise of P-PE is that stakeholder participation in evaluation will enhance evaluation relevance, ownership, and thus utilization. The utilization construct has been traditionally conceptualized in terms of three types of effects or uses of evaluation findings: (1) instrumental, the provision of support for discrete

> decisions; (2) conceptual, as in educative or learning function; and (3) symbolic, the persuasive or political use of evaluation to reaffirm decisions already made to further a particular agenda (Leviton and Hugues, 1981; King 1988; Weiss, 1972, 1979). Typically, impact is conceptualized in terms of effects on an undifferentiated group of "users" or "decision makers." (Cousins & Whitmore, 1998, p. 6)

Un second courant, la *transformative participatory evaluation* (T-PE), est fondé sur des motivations politiques et normatives. La participation à l'évaluation est en effet conçue comme un moyen de conscientisation, d'autodétermination et d'autonomisation des parties prenantes :

> Through direct involvement and participation in the research process, persons from oppressed groups or marginalized sectors that do not normally have a voice in policy or programme decision making are now provided with such opportunities. The focus for politically-oriented collaborative inquiry is very much emancipatory or concerned with the amelioration of social inequities inherent in the societal structures of the status quo. (Weaver & Cousins, 2004, p. 21)

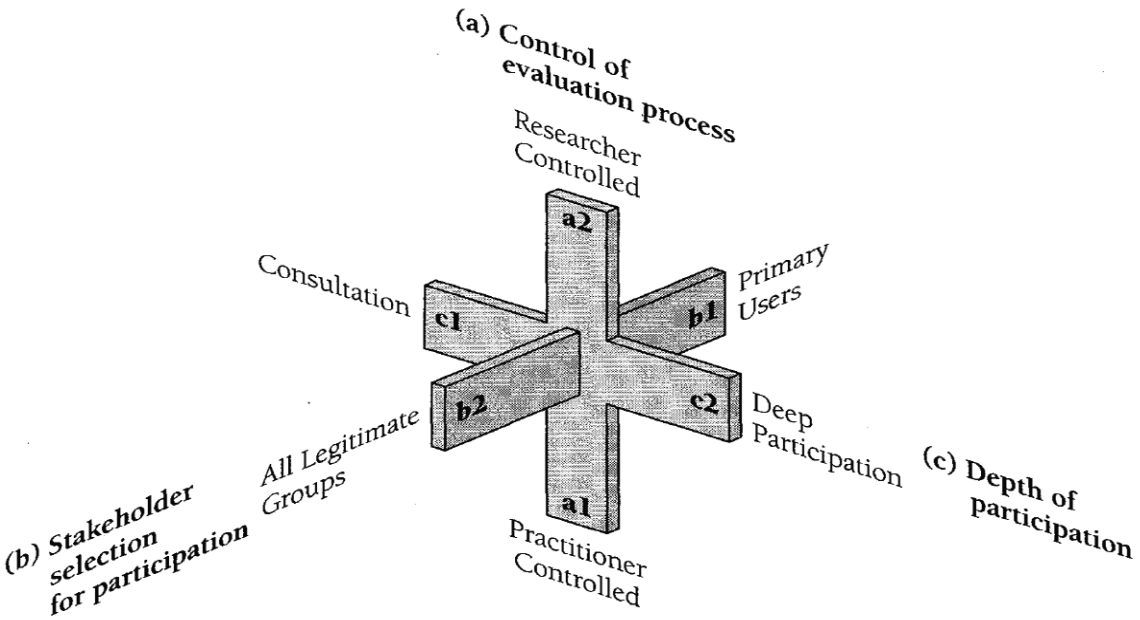
Les tenants du courant transformatif ne considèrent pas la participation exclusivement de manière instrumentale. En effet, la participation est également vue comme une question d'équité, de démocratie et de justice sociale et ce, peu importe ses conséquences. L'évaluation de type habilitative (*empowerment evaluation*) de Fetterman (2000) et l'évaluation démocratique-délibérative de House et Howe (2000) sont deux exemples d'approches s'inscrivant dans la T-PE.

Une autre justification, qui au contraire des deux précédentes n'a pas été formalisée au sein d'un courant à part entière, est épistémologique (Weaver & Cousins, 2004). La participation est alors un moyen d'arriver à la production d'un savoir informé par plusieurs perspectives, notamment celles des professionnels responsables de la prestation des services, et donc plus « valide ». La *responsive evaluation* (Stake & Abma, 2005) est un exemple d'approche évaluative dont les motivations participatives sont d'abord et avant tout épistémologiques.

Au-delà des courants et des justifications, Cousins et Whitmore (1998) ont développé, dans la foulée de travaux antérieurs sur le sujet (voir Cousins, Donohue et Bloom, 1996; Cousins et Earl, 1992, 1995), un cadre d'analyse tridimensionnel de la recherche collaborative (Figure 3). Ces trois dimensions de processus sont théoriquement indépendantes et peuvent

être utilisées pour comparer les processus collaboratifs que l'on retrouve non seulement dans les approches d'évaluation participatives mais dans la recherche participative généralement. La première dimension est le *contrôle du processus évaluatif* qui renvoie au contrôle des décisions techniques liées à la réalisation de l'évaluation (par opposition aux décisions concernant la pertinence d'initier une évaluation). Elle s'étend du contrôle total exercé par l'évaluateur au contrôle total exercé par les autres participants à l'évaluation. La deuxième dimension, *sélection des parties prenantes* qui participent à l'évaluation, renvoie aux types de participants impliqués dans l'évaluation. Elle s'étend de l'inclusion des principaux utilisateurs (*primary users*) à celle de tous les groupes légitimes. La troisième dimension est la *profondeur de la participation*. Elle couvre un spectre allant de la consultation (aucun pouvoir décisionnel) à l'implication dans tous les aspects de l'évaluation (conception, collecte et analyse de données, dissémination et utilisation des résultats).

Figure 3 : Les trois dimensions processuelles de la recherche collaborative



Les différents courants et approches d'évaluation peuvent être classés à l'aide de ces dimensions (Cousins & Whitmore, 1998, p. 12). Pour le courant pratique (P-PE), on dénote

une participation des utilisateurs principaux (b1) dans toutes les phases de l'évaluation (c2) et un contrôle partagé entre l'évaluateur et les participants (a1-a2). Le courant transformatif (T-PE) est similaire au courant pratique sur le plan de la profondeur de la participation (c2) mais s'en distingue toutefois par le fait que, d'une part, tous les utilisateurs légitimes sont impliqués (b2) et que, d'autre part, le contrôle est partagé mais repose ultimement dans les mains des participants (a1).

1.3.5 Des problèmes conceptuels persistants

Les premiers éléments de définition et le cadre conceptuel présentés précédemment pourraient laisser croire que l'ÉP est clairement conceptualisée. Ils visaient précisément à offrir quelques jalons sémantiques pour circonscrire le concept de participation. Or, la prolifération des approches participatives (démocratique, démocratique-délibérative, habilitative, « illuminative », fondée sur la parties prenantes, « développementale », etc.) témoigne d'un besoin réel de clarté conceptuelle (Cousins & Chouinard, à paraître). Si ces approches ne se résument pas à leur nature plus ou moins participative, leur coexistence dans un domaine aussi encombré contribue à mélanger les cartes quant au sens à attribuer à la participation. Autrement dit, la relation entre les termes employés pour désigner l'ÉP et le sens qu'on leur donne est problématique (le côté gauche du triangle dans le Figure 1). Plusieurs auteurs ont ainsi dénoncé la nature insaisissable et toujours changeante du concept (Hayward, Simpson, & Wood, 2004; Huberman, 1995). La métaphore du kaléidoscope illustre bien les lacunes conceptuelles de la participation: « Le mot 'participation' est kaléidoscopique; il change de couleur et de forme selon la volonté des mains qui le tiennent et, tout comme les images momentanées du kaléidoscope, il peut être très fragile et illusoire, changeant de forme d'un moment à l'autre » (Whyte, Nair et Ascroft, 1994, p. 16 cité dans Hayward et collab., 2004, p. 104 : traduction). En outre, certains éléments dans la conceptualisation de la participation posent problème, notamment la structure du concept. Enfin, la participation à l'évaluation doit encore être opérationnalisée de manière cohérente. Dans leur forme actuelle, les différentes conceptualisations de l'ÉP n'établissent pas de frontière claire et précise entre ce qui est participatif et ce qui ne l'est pas.

Les problèmes conceptuels qui caractérisent l'ÉP sont sérieux et nuisent à l'avancement des connaissances théoriques et empiriques dans le champ de l'évaluation. Face à ces problèmes, une réponse en trois volets semble nécessaire. Il s'agit dans un premier temps de procéder à une analyse systématique du concept de participation afin de le refonder sur des bases plus solides. Il s'agit dans un second temps de traduire cette conceptualisation refondée en termes opérationnels ou, autrement dit, de concevoir un instrument de mesure complet et cohérent. Il s'agit enfin de vérifier empiriquement si cet instrument mesure de manière fidèle et valide la participation à l'évaluation. Les chapitres qui suivent s'attèlent à ces tâches.

2 Vers une juste mesure de la participation : repenser la conceptualisation et la mesure de l'évaluation participative[19]

Résumé : Alors que l'évaluation participative (ÉP) constitue une tendance importante dans le champ de l'évaluation, son ontologie n'a pas fait l'objet d'une analyse systématique. En conséquence, le concept d'ÉP souffre d'ambiguïté et d'une théorisation inadéquate. En outre, il n'existe aucun instrument mesurant avec justesse la participation des parties prenantes à l'évaluation. Dans un premier temps, cet article tente de surmonter ces lacunes par une appréciation des conceptualisations actuelles de l'ÉP qui prend appui sur les travaux de G. Goertz (2006) et de J. Gerring (1999). Dans un deuxième temps, une version revue du cadre développé par J. B. Cousins et E. Whitmore (1998) est proposée comme alternative aux conceptualisations actuelles. Ce cadre révisé est ensuite opérationnalisé et adapté en un instrument de mesure de la participation. La conceptualisation et l'instrument proposés pourront potentiellement contribuer à la production de connaissances empiriques solides sur l'évaluation et à la réflexion sur la pratique de l'ÉP.

Mots-clés : *évaluation participative; évaluation collaborative; ontologie; analyse conceptuelle; opérationnalisation et mesure*

Abstract: While participatory evaluation (PE) constitutes an important trend in the field of evaluation, its ontology has not been systematically analyzed. As a result, the concept of PE is ambiguous and inadequately theorized. Furthermore, no existing instrument accurately measures stakeholder participation. First, this article attempts to overcome these problems by using the works of G. Goertz (2006) and J. Gerring (1999) on concept formation and evaluation to assess current conceptualizations of PE. Second, an amended version of the framework developed by J. B. Cousins and E. Whitmore (1998) is proposed as an alternative to current conceptualizations. This amended framework is then operationalized and adapted in a participation measurement instrument. The proposed conceptualization and instrument have the potential to contribute to the

---

[19] La version définitive de cet article, intitulée « Toward accurate measurement of participation: Rethinking the conceptualization and measurement of participatory evaluation », a été publiée en 2009 dans l'*American Journal of Evaluation*, vol. 30, n° 3 (pp. 330-348; http://aje.sagepub.com/content/30/3/330), par Sage Publications Ltd/Sage Publications, Inc., tous droits réservés ©. Il constitue une version bonifiée d'une communication présentée dans le cadre du colloque annuel de l'American Evaluation Association, Baltimore, Maryland, le 10 novembre, 2007. Il a été produit grâce au soutien financier du Conseil de recherches en sciences humaines du canada (CRSH), de l'École de la fonction publique du Canada et du Fonds québécois de la recherche sur la société et la culture (FQRSC). Les opinions exprimées dans cet article n'engagent toutefois que les auteurs. Nous désirons remercier J. Bradley Cousins, Jennifer Greene et Jean King pour leur critique approfondie d'une version antérieure du texte. Un remerciement spécial va à Laurence Ouvrard pour ses commentaires sur l'opérationnalisation du cadre. Nous désirons également remercier Jean-François Bélanger, Martin Cossette, Nouhoun Diallo et Mbaïrewaye Mbaï-Hadji pour leurs commentaires. Enfin, nous sommes reconnaissants à Kristen Leppington de sa révision du manuscrit final.

production of sound empirical knowledge about evaluation and to reflections on PE practice.

Keywords: *participatory evaluation; collaborative evaluation; ontology; conceptual analysis; operationalization and measurement*

Involving stakeholders in the evaluation process is a principle that is now generally accepted within the evaluation community (see Mathison, 2005a; Whitmore, 1998). Some authors even refer to this trend as the ''participatory orthodoxy,'' underlining the wide consensus on participatory methods (i.e., Biggs, 1995 as cited in Gregory, 2000, p. 180). Even for those who do not endorse participation as an ideal, the saliency in the evaluation field of such themes as stakeholder participation, inclusion, and empowerment can hardly be disputed.

While the popularity of *participatory evaluation* (PE) is good news for the proponents of stakeholder involvement, it also raises serious concerns in terms of conceptual development and production of empirical knowledge. As Huberman (1995) wrote more than 10 years ago: ''Participatory evaluation is a noble but elusive construct. It seems to recur every thirty years in a new rhetorical guise, but it presents the same tough conceptual and practical problems'' (p. 104). Many have concurred with this diagnosis (Murray, 2002; Rebien, 1996; Ridde, 2006).

PE is plagued by insufficient and/or inadequate conceptualization. First of all, multiple labels (e.g., collaborative evaluation, stakeholder evaluation, empowerment evaluation, interactive evaluation, democratic evaluation, fourth-generation evaluation, etc.) are used to characterize the same phenomenon (i.e., PE) which leads to misunderstandings among scholars and practitioners. PE is also polysemic, that is, it stretches to cover very different realities (Cousins, 2003; Garaway, 1995; Jackson & Kassam, 1998; Murray, 2002; Whitmore, 1998). For instance, this term ranges from a type of evaluation that ''seeks to include program personnel in the evaluation process'' (Torres et al., 2000, p. 27) to an approach which aims to be ''an educational process through which social groups produce action-oriented knowledge about their reality, clarify and articulate their norms and values, and reach a consensus about further action'' (Brunner & Guzman, 1989, p. 11). The polysemic nature of PE is not surprising, given the various rationales for undertaking it.

Weaver and Cousins (2004) have identified three main rationales or justifications for stakeholder participation: *pragmatic* (problem-solving orientation), *political* (social justice orientation), and *epistemological* (validity of knowledge orientation). In addition, PE is inadequately theorized from an ontological perspective. Many references to participatory approaches to evaluation are found in the literature, but most of these discussions take a normative or prescriptive perspective (i.e., advocating for stakeholder involvement) or limit themselves to a vague and informal definition of the meaning of the term. Systematic analyses of the constitutive dimensions or fundamental attributes of PE are indeed relatively rare. A last problem, which in part derives from the other shortcomings mentioned, is that few satisfactory operationalizations of PE exist in evaluation literature, thereby hindering adequate measurement. This should not be surprising as ''operationalizability'' rests on consistent conceptualization.

This situation of conceptual ambiguity and ''unoperationalization'' is especially problematic in the light of repeated calls urging evaluators to undertake further empirical research about evaluation in general (e.g., Christie, 2003; Mark, 2001; Smith, 1993) and on PE in particular (e.g., Cousins, 2001; Cousins & Earl, 1999; Mark, 2001). Case reports of practice are useful in that respect, but knowledge claims that are based on such reports are not as strong as those based on systematic and rigorous empirical research.

The purpose of this article, which is divided in three parts, is to overcome these problems. The first part presents our theoretical approach to conceptualization, which is essential to the comprehension of the rest of the article. The second part proposes an amended version of the framework developed by Cousins and Whitmore (Cousins & Whitmore, 1998) and an argument defending the usefulness of this conceptualization. In the third part, a theoretically guided operationalization of the amended framework and its adaptation in a measurement instrument are presented.

2.1 Theoretical Framework to Conceptualization

Concepts play a central role in the social sciences (Gerring, 1999; Goertz, 2006). They act as building blocks for hypotheses and theories, among other functions. Various approaches to conceptualization can be mobilized for that purpose. We rely here on the frameworks

developed by Gerring (1999) and especially Goertz (2006) to guide our conceptualization of PE. Using a consistent and systematic framework in the conceptualization process can substantially enhance the results of this endeavor.

2.1.1 Eight Criteria for Evaluating the Concept of PE

What is a valuable concept? Gerring (1999) identified eight evaluation criteria related to the functions fulfilled by concepts, namely familiarity, resonance, parsimony, coherence, differentiation, depth, theoretical utility, and field utility. According to Gerring, conceptualization is a matter of prioritization and tradeoffs between the different functions fulfilled by concepts, not of applying a ‘‘cookbook.’’ According to this ‘‘criterial’’ framework, some steps can be taken to ensure that conceptualization is *pareto-optimal* which means that, beyond a certain point, improving the performance of a concept on one dimension will imply losses on other dimensions. For instance, more parsimony (the shortness of a term and of the number of its attributes) could mean less differentiation as fewer attributes are mobilized to distinguish this concept from others. Concepts should therefore be formed in relation to their purpose in a specific research endeavor.

We have already argued that PE is plagued by many problems. Translating these problems in the language of this ‘‘criterial’’ framework allows us to better grasp the areas needing improvement. First of all, greater *internal coherence* and *parsimony* are needed as PE is often defined with respect to an unduly long list of attributes. For instance, Jackson and Kassam (1998) have listed as many as nine defining characteristics whereas Burke (1998) has identified seven principles of PE and as many key elements of its process. Second, given the current state of the literature on PE, effectively distinguishing between participatory and non-PE (*differentiation*) is difficult. Where exactly does the border lie between PE and ‘‘conventional’’[20] evaluation? Unfortunately, most authors have given only a vague definition of what they mean by PE and then do not bother to distinguish it explicitly from nonparticipatory approaches. Moreover, few of the current

[20] Various terms have been used by different authors to refer to nonparticipatory evaluation, for instance *conventional evaluation* (Rebien, 1996), *distanced evaluation* (O'Sullivan & D'Agostino, 2002), *independent evaluation* (Rossi et collab., 2004), *old-style evaluation* (Weiss, 1986), *technocratic evaluation* (Murray, 2002), and *traditional evaluation* (VanderPlaat, Samson, & Raven, 2001). Even though these terms are ambiguous and need further specification, we will use the terms *conventional* or *traditional* throughout the text to qualify evaluations that are not participatory.

conceptualizations allow for comparing and ranking different evaluations or theoretical approaches according to their level of participation. When they do, as in the case of the framework developed by J. Bradley Cousins and his colleagues (Cousins, 2005; Bradley Cousins, Donohue, & Bloom, 1996; Bradley Cousins & Whitmore, 1998; Weaver & Cousins, 2004), the ranking of evaluation approaches is made with respect to certain process dimensions of collaboration (e.g., the diversity of participants in an evaluation) but not according to the overall degree of participation.

2.1.2 Emphasizing Ontology and Concept Structure

Gerring's (1999) framework allows us to make a diagnosis about PE's conceptualization, but it is less useful in pointing to possible remedies. This is why we supplement it with Goertz's (2006) framework which offers a consistent and practical guide to conceptualization. Goertz has adopted an ontological, causal, and realist perspective on conceptualization. It is ontological as it focuses on a phenomenon's essential attributes rather than on its secondary, accompanying or superficial characteristics: ''Concepts are theories about ontology: they are theories about the fundamental constitutive elements of a phenomenon'' (Goertz, 2006, p. 5). These essential characteristics of a concept play an explanatory role in theories and hypotheses as causal mechanisms. This approach stresses that to conceptualize is to reflect upon and analyze what a concept really is, that is, the phenomenon it refers to and its fundamental attributes—the disease—in contrast to the statistical and factor analytic approaches that emphasize its consequences—the symptoms. For instance, PE should be defined independently of its plausible consequences (e.g., evaluation use or empowerment) to avoid circularity.

This approach insists on concept structure and *concept-measure consistency*, which refers to ''the degree to which the numeric measure reflects well the basic structure of the concept''(Goertz, 2006, p. 95). Important concepts have three levels, namely the basic, secondary, and indicator/data levels. The *basic* level is the most general and characterizes concepts as they are used in theoretical propositions (e.g., PE increases evaluation use). The *secondary* level is made up of the constitutive dimensions or fundamental attributes of a concept. For instance, we argue that control of the evaluation process is one of PE's fundamental attributes. The most concrete level is the *indicator/data level* or

operationalization level that guides the collection of empirical data about a phenomenon or, in other words, its measurement.

A sound concept has a consistent structure from its indicators to its basic level. Apart from the statistical approach, Goertz (2006) identifies two prototypical structures: (a) *necessary and sufficient condition* and (b) *family resemblance*. Mathematically, the necessary and sufficient condition concept structure is characterized by the classical logic operator ‘‘AND’’ or the intersection in set theory. This structure entails that all the relevant dimensions or attributes of a phenomenon are necessary (i.e., required) and jointly sufficient (i.e., no other dimension is required) for a phenomenon to fit into a concept. The second concept structure, family resemblance, is mathematically modeled by the classical logic operator ‘‘OR’’ or the union in set theory (Goertz, 2006). This structure allows the absence of one dimension to be compensated by the presence of another and is characterized by an ‘‘*m* of *n* rule’’ meaning that *m* dimensions out of *n* are needed to assume that we are in presence of the concept. Different structures can be used for different levels of conceptualization of a same concept. A common combination for concepts is the necessary and sufficient structure at the dimension level and family resemblance at the indicator level (Goertz, 2006). Indeed, we argue that PE is characterized by three necessary and jointly sufficient conditions at the secondary level (i.e., diversity of participants, extent of involvement, and control of the evaluation process) and by a family resemblance structure at the indicator level (i.e., no indicator is individually necessary; only a sufficient number of them).

In addition, Goertz (2006) has argued for considering all concepts as continuous (treating dichotomous concepts as special cases), as this reduces measurement error and allows theorists and researchers to better tackle the *negative pole* of a concept and the problem of borderline cases, which are related to the criteria of differentiation presented earlier. Goertz has contended that concept continuity can be usefully theorized using the tool of fuzzy logic and set theory (see also Ragin, 2000). In a nutshell, fuzzy logic rejects the ‘‘black or white logic’’ (i.e., a yes or no logic) of conventional sets and proposes that most phenomena do not fit perfectly in a clear-cut category. Partial membership is allowed and represented by intermediate scores between .00 (completely out the set) and 1.00

(completely in the set).[21] Continuous dimensions and fuzzy logic allow for more precise and refined measurement. Fuzzy sets are different from ordinal scales because they possess an explicit zero-point and because they combine measurement, theory and interpretation. Indeed, they need to be "calibrated" using substantive knowledge about the cases (Ragin, 2000). We will come back on the issue of indicators and dimensions aggregation using fuzzy logic when applying these tools to the conceptualization and operationalization of PE.

The frameworks and tools presented so far will assist us in proposing a conceptualization of PE that is more parsimonious, has a more consistent structure, and allows one to better differentiate it from neighboring concepts. Before going further, a quick note on our epistemological perspective is warranted. Although PE is often rooted in a constructivist, transformative, or emancipatory epistemology, it does not necessarily entail that all empirical research on this topic should be participatory or rooted in critical epistemologies. PE can indeed be studied from different perspectives. This study rests on a traditional, realist epistemology. Stated otherwise, we believe that there is a phenomenon ''out there'' called PE that can be measured. Furthermore, our focus is exclusively theoretical and empirical, not normative (i.e., we do not take sides with respect to the desirability of PE).

2.2 Toward the Development of an Amended Version of the Cousins and Whitmore (1998) Framework

2.2.1 A Useful Starting Point for Reconceptualizing PE

We have argued that the most serious shortcomings of current PE conceptualizations are their lack of internal coherence, parsimony, and external differentiation. We contend that *the necessary and sufficient condition structure* at the secondary level is the most appropriate aggregation procedure to overcome these challenges. Indeed, this structure emphasizes the constitutive elements that are required for an evaluation to fit into the conceptual category of PE and helps users distinguish it from non-PE. That being said,

[21] The .50 point denotes "maximal ambiguity", that is the point where cases are as much in as out the logical set (Ragin, 2000). It must be stressed that the operationalization of the concept of PE presented here violates this characteristic of fuzzy logic as the participatory threshold is set at .25 rather than .50. However, this infringement on fuzzy logic is transparent and, in practice, does not seem to have serious implications with respect to the measurement of PE.

identifying the fundamental attributes of PE is not an easy matter. To maximize its usefulness in scientific research, the selection of PE's constitutive dimensions should rest as much as possible on the literature so that these dimensions are *familiar* to evaluators and stakeholders. Additionally, these dimensions must be congruent with our basic intuitions about instances of PE. When incongruence occurs, we must either revise our intuitions or modify the selection of dimensions (i.e., the method of reflective equilibrium; see Daniels, 2003).

Fortunately for us, Cousins and Whitmore (1998) have developed a valuable framework to classify various forms of PE and collaborative inquiry in general. Their conceptual framework distinguishes between two streams of PE, namely *practical* (P-PE) and *transformative* (T-PE), that are characterized by different rationales, organizational decision making and problem solving on one hand and empowerment of disadvantaged or oppressed groups on the other. The framework also extends and formalizes three dimensions that have been more or less explicit in Cousins' earlier work (see Cousins et al., 1996; J. Cousins & Earl, 1992, 1995). These process dimensions are conceived as analytical tools used to describe, classify, and rank various types of collaborative inquiry or PE. The first dimension is *control of the evaluation process* (or simply the *control* dimension) which refers to the control of the technical decisions related to the conduct of evaluation as opposed to decisions about whether and when to initiate an evaluation. It is conceived as a continuum ranging from total control by the evaluator to total control by other stakeholders. The second dimension, *stakeholder selection for participation* (or simply the *diversity* dimension), refers to the types of stakeholders involved in the evaluation and ranges from the inclusion of primary users to all legitimate groups. The third dimension is *depth of participation*; it ranges from consultation (implying no decision-making authority) to deep participation, namely involvement in all evaluation tasks (i.e., design, data collection, data analysis, reporting, and decisions about dissemination of findings and use).

We argue that these three dimensions also happen to be PE's fundamental attributes or constitutive dimensions. The fact that the dimensions identified by Cousins and Whitmore (1998) correspond to the basic intuitions that many evaluators have about the nature of PE

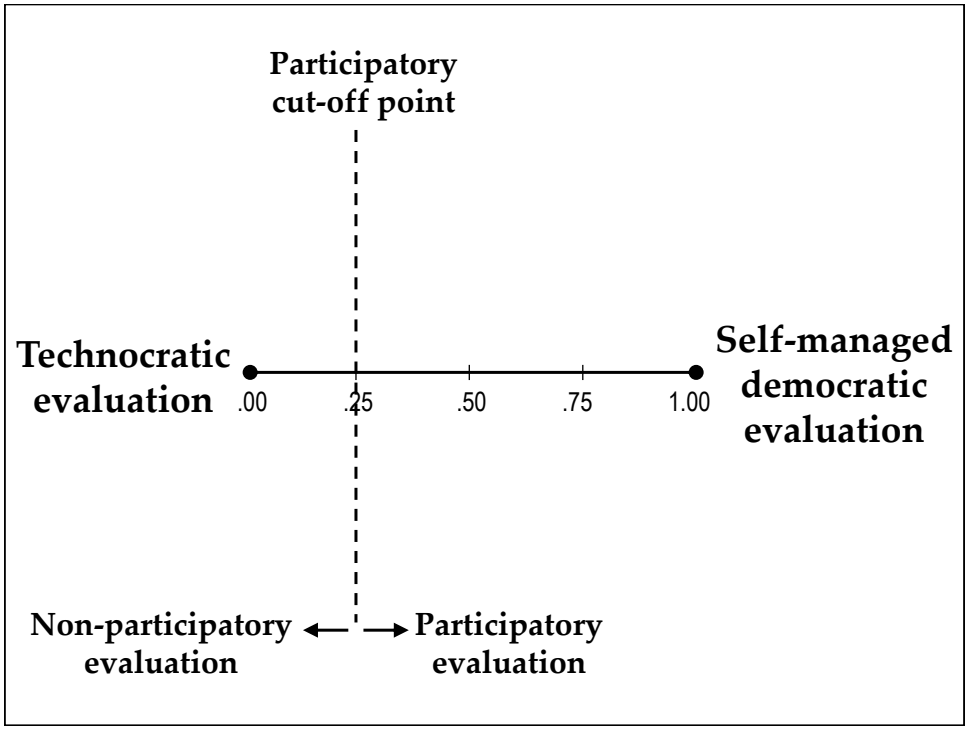
supports that claim. Using selected references, we first show that the three dimensions are more or less implicit in many discussions of participatory approaches to evaluation and already are familiar to many evaluators. We then demonstrate how the framework satisfies the concept structure of necessary and sufficient conditions for PE. We also argue that this framework helps to overcome the limitations of current conceptualizations that have been identified above.

The three constitutive dimensions of PE (diversity of participating stakeholders, involvement in the evaluation process and control) are found in a more or less explicit form in the work of many authors in the field of evaluation. For instance, Mathie and Greene (1997) have put forward the following definition of PE: ''A defining feature of PE is the active engagement of multiple stakeholders'' (p. 279). Other authors have offered very similar accounts of PE, for instance, Rossi, Lipsey, and Freeman (2004): ''The participating stakeholders are directly involved in planning, conducting, and analyzing the evaluation in collaboration with the evaluator whose function might range from team leader or consultant to that of a resource person called on only as needed'' (p. 51). Similarly, King (2005) has stated that ''*participatory evaluation* is an overarching term for any evaluation approach that involves program staff or participants actively in decision making and other activities related to the planning and implementation of evaluation activities'' (p. 291). More or less implicit in these definitions are the facts that people not usually involved in traditional evaluation (i.e., ''multiple stakeholders'', ''participating stakeholders'', ''program staff or participants'') are involved in the evaluation and that their involvement takes the form of meaningful participation (i.e., ''active engagement'', direct or active involvement in decision making and other tasks).

First, who are those people not usually involved in traditional evaluation but who are involved in PE? A fair, yet imprecise, answer is *stakeholders*—a concept defined as the ''people who have a stake or a vested interest in the program, policy, or product being evaluated ... and therefore also have a stake in the evaluation'' (Greene, 2005, p. 397). This answer lacks in precision because evaluators and evaluation sponsors are also stakeholders and have always been involved in evaluation (see Forss, 1993, as cited in King, 2005; Rebien, 1996): evaluation sponsors defined the evaluation terms of reference, provided

payment for the evaluation, and used the findings; evaluators conducted the evaluation per se; and, in some instances, program staff and beneficiaries provided data for the evaluation. In that broad sense, every evaluation is participatory. Nevertheless, PE departs from this picture of traditional evaluation on a major issue. PE directly involves not only the evaluator but also various actors in the process of actually producing the evaluation. In contrast, most stakeholders are not involved in traditional evaluation and when they are, it is indirectly either by planning the evaluation process or by providing data. Traditionally, evaluators and sponsors (typically decision makers) plan the evaluation and evaluators carry it out. PE is thus characterized by the fact that nonevaluative stakeholders play a significant role in the evaluation process, that is, evaluators share their tasks with other stakeholders. One could even imagine an extreme case in which all evaluative tasks are carried out by nonevaluative stakeholders without the support of an individual trained in evaluation theory, methods, and practice (see in Figure 4, *self-managed democratic evaluation*). In this case, stakeholders become de facto evaluators.

Figure 4 : The participation index and two polar constructs



Second, the involvement of nonevaluative stakeholder takes the form of meaningful participation, which is made up of two distinct dimensions. The first is involvement in a number of different evaluation tasks, which is a necessary, but not sufficient, condition for PE. ‘‘Participation’’ might be characterized by evaluator-directed interactions with stakeholders and passive involvement of the latter in the evaluation process. In that case, stakeholders are objects of the evaluation and act as data providers. Participation can also signify that stakeholders are real subjects of the evaluation and have a significant degree of control or decision-making power over a number of issues such as methodological design (King, 2005). Control of the process is therefore the second facet of meaningful participation and the third fundamental dimension of PE.[22] It has a broad intuitive appeal within the evaluation community and is explicitly supported by evaluation literature (Burke, 1998; Greene, 1987, 2005; King, 2005; Murray, 2002; Rebien, 1996; Weiss, 1986, 1998). Back to the Cousins and Whitmore (1998) framework, its wide influence within and outside the field also lends credence to this claim (e.g., Butterfoss, Francisco, & Capwell, 2001; Themessl-Huber & Grutsch, 2003): their article is indeed one of the most cited chapters ever published in *New Directions for Evaluation* (King, 2007). The dimensions of control, diversity, and depth of participation are familiar to many evaluators and satisfy one of the criteria of good conceptualization.

We contend that each of the dimensions presented above—including control—is required to classify an evaluation as participatory. Therefore, stakeholders selected for participation, depth of participation[23] and control of the evaluation process dimensions are all *necessary* conditions for PE (see above, i.e., the section presenting our theoretical framework). Moreover, we argue that these dimensions are *jointly sufficient* for membership in this category. In other words, other ‘‘defining’’ characteristics of PE that are found in the literature are unnecessary or nonessential conditions. For example, Burke (1998) argued that the process of PE ‘‘*should use multiple and varied approaches to codify data*’’ and ‘‘*should explicitly aim to build capacity*’’(p. 46). Indeed, PE might be *more likely* than

[22] When stakeholders are included in the evaluation process but exert no control whatsoever on it, we qualify this situation as one where there is *involvement* but no *participation*. In set-theoretic terms, the concept of participation is a subset of the concept of involvement. Therefore, participation is always a form of involvement but the converse is not necessarily true.

[23] The dimension *depth of participation* is renamed below to avoid potentially contradictory situations such as a case which display a high level of depth of participation but is nevertheless nonparticipatory.

conventional evaluation to use more than one method to codify data or to promote capacity-building, but these elements are not necessary conditions: they are only accompanying characteristics for membership in the PE category.

A few words on the addition of two dimensions to the original Cousins and Whitmore (1998) framework are warranted as it has been argued in Cousins' subsequent works that the diversity dimension was confounded and conceptually inadequate (Cousins, 2005; Weaver & Cousins, 2004). As a result, Weaver and Cousins (2004) have recast this original dimension as an almost identical dimension, namely *diversity among stakeholders selected for participation*, and two other dimensions, *power relations among participating stakeholders* and *manageability of evaluation implementation*. These new dimensions, respectively, address power differentials between participating groups (with power relations ranging from *conflicting* to *neutral*) and the feasibility of evaluation in relation to logistics, time, and resources in particular (from *unmanageable* to *manageable*). Although these dimensions add analytical power to the original framework and thus allow for a more fine-grained and precise description and classification of evaluation approaches, they do not constitute necessary attributes of PE. Manageability of evaluation implementation is a consequence of PE or a corollary of the three original dimensions. Indeed, participatory approaches are frequently seen as time-consuming and associated with higher costs (Butterfoss et al., 2001), but this need not be the case. Some instances of non-PEs are harder to manage than some instances of PE, for instance, a multisite evaluation using randomized control trials. Moreover, stakeholder involvement sometimes *facilitates* the research process by reducing conflict between stakeholders relative to program decision making (Weiss, 1983, p. 12). Thus, the manageability (or unmanageability) of evaluation implementation is not a fundamental attribute of participatory approaches. Similarly, power relations between stakeholders cannot define PE because these relations exist even in the case of non-PE and can be conflicting. Thus, while the revised and expanded framework may be useful for descriptive and classificatory purposes, considering the new dimensions as fundamental attributes of PE is inappropriate. Consequently, we uphold the original framework for the conceptualization of PE.

To sum up our argument, the three process dimensions identified by Cousins and Whitmore (1998) can logically be described as necessary constitutive dimensions of PE. In addition to their familiarity to evaluators, four further advantages of this conceptualization should also be noted. First of all, with only three dimensions required to characterize PE and to distinguish it from nonparticipatory approaches, the framework is relatively parsimonious (i.e., the list of fundamental attributes is as short as it can be for that purpose). Second, the identified fundamental attributes display an impressive degree of internal coherence. The three dimensions are logically related through the necessary and sufficient condition structure and, as process dimensions, are located in the same unit of analysis. Third, these attributes really distinguish PE from more conventional forms of evaluation, that is if one of these attributes is lacking in a given evaluation, it does not feel participatory. Finally, the framework can accommodate both streams of PE (P-PE and T-PE) and various types of evaluation in which stakeholders participate meaningfully (fourth-generation evaluation, democratic evaluation, empowerment evaluation, etc.). As a result, it can be applied in a wide range of instances.

2.2.2 Using the Framework as a Measurement Instrument: Some Difficulties

Although the Cousins and Whitmore (1998) framework has been very useful for theorists and practitioners in its conceptual form since its development, it has limitations with respect to the goal of *measuring* participation. To be sure, the authors should not be faulted for such imperfections as they might not have intended to translate their framework into a fully specified operational instrument to measure participation. Yet, these shortcomings are real and should be tackled through careful operationalization of the secondary-level dimensions. Before turning to this operation, a few general issues related to operationalization need to be examined.

The first issue has to do with the substantive content of each dimension and of the basic concept. At the basic level, PE constitutes the positive pole of the continuum whereas conventional or non-PE is the negative pole. ‘‘Mechanically and numerically the negative pole can be operationalized as zero on all of the secondary-level dimensions that characterize the positive extremes’’ (Goertz, 2006, p. 32). Now, in its current state, the

Cousins and Whitmore (1998) framework does not possess an explicit zero point for all dimensions. The diversity dimension ranges from the inclusion in the evaluation process of primary users to all legitimate groups. On the continuum of stakeholder selection, no other participants may exist except for the evaluation sponsor and the evaluator, as in the case of traditional evaluation. In other words, this should be the negative pole of the diversity dimension. The conceptualization of the depth of participation dimension is incoherent because the indicators used at each pole are not of the same type. At its positive end, the depth of participation dimension is indeed clearly conceptualized as the involvement in all tasks of an evaluation, namely design, data collection, analysis, reporting, dissemination of results, and use. The polar opposite of involvement in *all* evaluation tasks is *no* involvement in any task rather than *consultation*.[24] Because consultation refers more to low decision-making power than involvement in only a few tasks of the evaluation process, this dimension seems to be ''contaminated'' by the control dimension *at its lowest bound*. This is problematic because the control issue is already taken into account by the control dimension, thereby double counting the control attribute. Stakeholders might ''only'' be consulted and yet be involved in all tasks of the evaluation process. Such a case would thus receive a high score on the depth of participation dimension but not on the control dimension. In short, a coherent conceptualization of each dimension is needed to set a clear cut-off point, a threshold, between an evaluation that is participatory and one that is nonparticipatory.

A second issue relates to the specification of indicators. The Cousins and Whitmore (1998) framework gives insufficient attention to the *indicator/data level* where a phenomenon is empirically measured to examine the match between this phenomenon and a given concept (Goertz, 2006). With the exception of the two extreme values of each continuum, the authors have not given any detailed indications as to how one should apply their framework. The justification as to why a given evaluation should get that specific rating—whether numeric, nominal or both—is entirely left to the appreciation of the rater. Even though the five-dimension version of the framework has been applied with apparent

[24] In all justice to the designers of the conceptual framework, it must be stressed that the ''true'' negative pole of PE has received some attention in an earlier article. The lower bound of the depth of participation dimension has indeed been labeled ''No Participation/Consultation Only'' (Cousins et collab., 1996, see also pp. 209-210). Even in that case, however, the use of ''consultation'' is still problematic.

success, the precision and reliability of the ratings seemed to have been a concern (see Ridde, 2006; Weaver & Cousins, 2004). The conceptual framework in its actual form has, thus, a limited usefulness as a *measurement* device. In addition, in a deductive approach to conceptualization and operationalization like the one used here, indicator selection must be guided by theory and be consistent with the structure of the basic concept. An explicit and consistent aggregation procedure must thus link the indicators to the secondary-level dimensions. We argue that PE should be characterized by the necessary and sufficient condition structure at the secondary level and the family resemblance structure at the indicator level.

A third issue is that the Cousins and Whitmore (1998) framework only allows one to rank an evaluation on each dimension (e.g., control) but not according to its general level of participation. It makes no mention of the structure holding dimensions together. A participation index that would allow researchers to precisely measure participation and verify covariation with other constructs such as evaluation use is therefore needed.

Together with the framework for concept formation presented earlier, these issues will inform and guide the operationalization of secondary-level dimensions so that a significant level of concept-measure consistency is achieved. The operationalization of each dimension is examined in the next section.

2.3 From Conceptualization to Measurement: Operationalizing the PE Framework

2.3.1 Extent of Involvement

We have renamed the depth of participation dimension as *extent of involvement* because it better describes and resonates with the idea of stakeholder involvement in a number of evaluation tasks. Stakeholders can be involved throughout the whole process but only superficially (e.g., mere presence without decision-making authority, token participation). In that case, speaking of involvement is more accurate than participation, as the latter term conveys the idea of a certain level of control over the evaluation process. The term *extent* is more appropriate than *depth* to reflect the number of tasks in which stakeholders are involved, as the latter refers more to the quality and intensity of involvement than its

‘‘quantity.’’ Because PE is about stakeholder involvement in the production of the evaluation, we restrict the measurement of this attribute to the *technical* tasks of the evaluation process such as evaluation design and interpretation of findings (seeWeaver & Cousins, 2004). By using the term technical, we stress that these tasks are more technical in nature than the decision to initiate an evaluation and the use of evaluation findings. The technical tasks are those normally considered to be the responsibility of the evaluator. Contrary to the view held by Cousins and Whitmore (1998), the use of evaluation findings is not considered here as an evaluation task per se. We do not consider it a task of evaluators to actually use the evaluation findings to enact change, however desirable the use of evaluation is.

In PE, like in traditional evaluation, involvement in the following key decision points is essential:

1. *Evaluation questions and issues definition/methodological design* is the moment when a decision is made about the framing of the evaluation including the selection of evaluation questions and issues, theoretical framework, methods, techniques, and instruments. Guiding questions: What is the rationale for conducting the evaluation (program improvement, accountability, or knowledge production)? What is the evaluation focus (needs, processes, outputs, or outcomes)? What is the evaluation type (formative, summative, internal, external, impact, implementation, etc.)? What are the informational needs to which the evaluation can and will answer? Which criteria (relevance, effectiveness, efficiency, equity, etc.) should guide normative judgments? What type of research design is chosen for the evaluation (experimental, quasi-experimental, qualitative, quantitative, meta-analysis, etc.)? What is the methodological logic underlying the evaluation (exploratory, confirmatory, etc.)? Which sources of data will be used?

2. *Data collection and analysis* is the moment where a decision about how to concretely collect, assemble, code, and analyze data (documents, interviews, quantitative data pertaining to treatment effect, etc.) is made and when these tasks are actually carried out. Guiding questions: Who will collect, assemble, code, and analyze data? How?

3. *Judgments and recommendations formulation* is the moment where a decision is made about determining the merit and worth of a program on one hand, and formulating suggestions for future action on the other. Guiding questions: With respect to the selected quality criteria, what standard of performance is considered adequate? What is the merit and worth of this program? Why? What will be done about it?

4. *Report and dissemination of evaluation findings* is the moment where a decision is made about the reporting and diffusion of evaluation findings and their implications. Guiding questions: What communication strategy will be used? Who will be targeted?

Each task is considered a dichotomous indicator of the type *involvement of nonevaluative stakeholders in the task* (presence of the indicator) or *no involvement of nonevaluative stakeholders in the task* (absence of the indicator) where no particular indicator (e.g., involvement in a specific evaluation task such as evaluation design) is necessary for membership in the secondary-level dimension. This is a substitutability logic: as long as the number of indicators is sufficient, the extent of involvement dimension is present. Thus, involvement is defined as the presence of stakeholders (excluding the evaluator) during evaluation key moments. The assumption underlying the coding scheme is that the more tasks nonevaluative stakeholders are involved in, the more participatory an evaluation is, all other things being equal. It does not matter how many types of stakeholders are involved in each task of the process for this dimension. The unit of analysis for this dimension is the evaluation process and its tasks, not the participants. To avoid double counting, the number of participating stakeholders is only taken into account by the diversity dimension (see below).

To effectively distinguish between participatory and nonparticipatory approaches to evaluation, a cut-off point needs to be established. We argue that the involvement of nonevaluative stakeholders in one evaluation task is the minimum required for an evaluation to be considered participatory on the extent of involvement (see Table 3). In terms of the *m* of *n* rule, it means that $m = 1$ and that $n = 4$. In Table 3, the coding values represent the level of membership in the logical set of this dimension. Each indicator has thus a weight of .25. This simple scheme improves ease of use and interpretation and facilitates aggregation with other dimensions.

A quick example might help to clarify the selected coding. Say the frontline staff of a program (i.e., implementers and deliverers) is involved in one way or other in two tasks: (a) data collection and analysis; (b) report and dissemination of evaluation findings. Program beneficiaries, another type of stakeholders, are involved in only one task, namely data collection and analysis. Nonevaluative stakeholders are thus involved in two different tasks

inasmuch as the involvement of more than one type of stakeholders in the same task, that is data collection and analysis, should not be counted twice. According to the coding scheme proposed, the extent of involvement of this specific evaluation would thus be coded as .50 and considered moderate.

Table 3 : Coding scheme for extent of involvement

Number of Tasks Nonevaluative Stakeholders are Involved in	Level of membership	
	Intuitive Label	Numerical
0	No involvement	.00
1	Limited/Weak involvement	.25
2	Moderate involvement	.50
3	Substantial/Strong involvement	.75
4	Full involvement	1.00

Some tasks carry more weight for the conduct of evaluation than others in terms of stakeholder influence on the content of evaluation. That being said, involvement in the evaluation process has a substantive importance that is distinct from control. For example, from a learning, empowerment or process use perspective, involving stakeholders in reporting the findings may be more important than involving them in issues definition and methodological design. All tasks are thus considered of equal importance (i.e. indicators are equally weighted).

2.3.2 Diversity of Participants

We renamed the second dimension, namely stakeholder selection for participation, as *diversity of participants*. We assume that the more diverse the types of stakeholders involved, the more participatory an evaluation is, all other things being equal[25]. As used here, the term *stakeholders* refer to nonevaluative stakeholders, as presented in Table 4.

[25] Although in many contexts it is neither desirable nor feasible to involve all stakeholders in the evaluation, this framework focuses on what is logically possible to achieve in general and is thus context-free.

Table 4 : A typology of nonevaluative stakeholders

Types	Description	Examples Drawn From the Extension
Policy makers and decision makers	People politically, legally and organizationally accountable for the program and its evaluation	Elected and appointed officials, high-ranking civil servants, chief executive officers of nonprofit private foundations and think tanks, etc.
Implementers and deliverers	People responsible for the midlevel management and implementation of the program and the delivery of the intervention and/or services	Lower-level program managers; street-level civil servants, front-line staff and professionals (psychologists, nurses, receptionists, international development volunteers, etc.)
Target populations and intended beneficiaries; indirect beneficiaries and injured parties	People toward which the program is directed to modify their behavior and/or improve their well-being; local people indirectly and/or potentially affected by the program, either positively or negatively.	Juvenile offenders, gays and lesbians, psychotic university students, large families with violence problems, K-12 girls, drunk drivers, HIV-infected farmers, tribal council, community members (neighbors, village elders, fellow believers, classmates, local storekeepers, etc.), family members, etc.
Civil society and citizens	People and organizations having a political interest in the program and its evaluation	Interest groups, unions, think tanks, NGOs, professional associations, private firms, intellectuals, political parties, scientists, etc.

Our typology builds on and extends the categories of nonevaluative stakeholders as devised by Greene (2005). Stakeholders are classified with respect to their distance from the program and its management, decision makers being the closest and civil society the farthest.[26] Although nothing is necessary in this relationship (e.g., unions and first-line civil servants could be directly involved in program management in corporatist settings), we contend that it represents a fairly accurate picture of program structure for most settings. Moreover, a four-category typology is convenient for coding and measurement purposes because it easily allows discriminating different evaluations in terms of their participants. The typology is intended to apply to a wide range of contexts, including the evaluation of international development aid projects. According to the proposed conceptualization, the fact that two or more organizations (e.g., the Departments of Commerce and Health)

[26] According to this framework, an evaluator, whether internal or external, does not have to be present in all evaluations. For instance, an evaluation can be entirely conducted by stakeholders such as program implementers who possess evaluation and methodological knowledge.

collaborate to sponsor an evaluation is not sufficient to label it participatory. Nonevaluative stakeholders need to directly contribute to the evaluative process.

The proposed coding scheme considers each category of stakeholders a dichotomous indicator of the type involvement/no involvement in a given evaluation. To be considered participatory, an evaluation must involve at least one type of nonevaluative stakeholder, thus attributing a weight of .25 to each indicator forming this secondary-level dimension (see Table 5). No particular type of nonevaluative stakeholders (e.g., target populations and intended beneficiaries) is necessary for an evaluation to be participatory. In terms of the m of n rule, it means that $m = 1$ and $n = 4$. Involvement of stakeholders is not necessarily progressive in terms of types. For instance, nothing prevents representatives from unions and lobbies (i.e., civil society groups) from being the sole nonevaluative stakeholders participating in an evaluation.

Table 5 : Coding scheme for diversity of participants

Number of Nonevaluative Stakeholders Types Involved	Level of membership	
	Intuitive Label	Numerical
0	No diversity	.00
1	Limited/Weak diversity	.25
2	Moderate diversity	.50
3	Substantial/Strong diversity	.75
4	Full diversity	1.00

We mentioned earlier the need for a framework that adequately captures the dynamic nature of the evaluation process. For instance, frontline staff may be involved in most steps of the evaluation process, whereas the involvement of program beneficiaries is limited to defining the evaluation questions and issues. The unit of analysis for the diversity dimension is the evaluation itself, not the different steps of its process. Whether two types of stakeholders are involved together at one step of the process or separately in two different tasks, the diversity of participants is the same. Furthermore, an evaluation that involves two types of stakeholders at one step in the process is not fundamentally different

from another where involvement is sustained throughout the entire process *with respect to the diversity dimension*. Thus, diversity should measure the total number of nonevaluative stakeholder types involved in a given evaluation, not the diversity at different steps. Scoring this dimension is thus relatively straightforward: one just has to add .25 for each different type of stakeholder involved at one point or the other (excluding evaluators). If, for example, frontline staff and direct beneficiaries of a program participate in a given evaluation (disregarding the particular tasks in which they are involved), this particular evaluation would get a diversity score of .50.

2.3.3 Control of the Evaluation Process

PE is characterized by the fact that nonevaluative stakeholders partially or totally control the evaluation process. The assumption underlying PE's third constitutive dimension, namely control, is that the more control nonevaluative stakeholders have over the various tasks of the evaluation process in which they are involved, the more participatory an evaluation is, all other things being equal. This dimension is theorized and measured in relative terms. Thus, one has to compare the control that participants (taken as a whole) have over the process to the control the evaluator and sponsors have. An important precision has to be made here. Evaluation sponsors can also be participants in this evaluation. For instance, the board of trustees of a private foundation can sponsor an evaluation and be involved directly in the actual production of this evaluation (collecting data, formulating conclusions, etc.). In that case, sponsors are considered participants in the evaluation. Control is thus measured in terms of the share of control nonevaluative stakeholders have (in that case, decision makers that sponsor the evaluation) compared to the evaluator.

The operationalization of this dimension is somewhat less straightforward than the other two dimensions because control varies substantively during the process (Themessl-Huber & Grutsch, 2003). In contrast to diversity of participants, which is measured for the whole evaluation, control is inseparable from the different tasks of the process. Thus, this score must reflect the way control is exercised at different moments of the process even though control is measured for the whole evaluation.

We suggest two main indicators to measure control of the evaluation process. The first indicator is *authority* to make decisions, which refers to the legal and organizational power and legitimacy to decide what to do with respect to a given evaluation task. For instance, a law might force evaluators to involve teachers and parents in decisions related to the methodological design of an educational program evaluation. Whereas the first indicator refers to what could be labelled official or formal power, the second indicator captures a less formal influence of the evaluation process. Thus, *other resources of influence* refer to other resources besides authority that stakeholders can mobilize to influence the evaluation process, such as substantive and methodological expertise; money and other material resources; mobilization power; values, norms and principles; and persuasiveness and social skills. These sources of influence are indirect but, if mobilized, can lead stakeholders to have real control over the evaluation. To pursue with the preceding example, teachers could exert substantial influence over another evaluation task for which the law does not grant them official recognition (e.g., the formulation of recommendations) because they possess expertise on education and can mobilize effectively to voice this expertise. These resources are used to influence decision making directly (i.e., what is done) but also wield influence indirectly through agenda setting (i.e., what evaluation options are discussed at each stage of the process) and control of the operational setting in which the evaluation takes place (i.e., who participates, when and how, etc.).

The multifaceted nature of control makes the process of coding the control indicators and aggregating their scores not a clear-cut operation. We, therefore, propose a more intuitive and flexible coding scheme than for PE's other dimensions. Whereas the secondary-level dimension of control should still use the same scale of measurement as the other dimensions for aggregation purpose (see Table 6), we do not propose any set-in-stone procedure to derive this score from indicators. The upper limit of the control scale (see Table 6) represents the possibility that nonevaluative stakeholders become both evaluators and sponsors of the evaluation. Theoretically, good reasons exist for adhering to the family resemblance structure of aggregation which implies that no particular indicator is required for the secondary-level dimension of control to be present. Indeed, many routes may be used to influence the evaluation process. A scale ranging from .00 to 1.00 for *indicators* would allow for fine-grained measurement, but would also make it cumbersome (especially

as the second indicator takes into account different types of resources: financial, expertise, persuasive, and for last normative). The question of whether one should give different weight to the indicators also makes measurement a tricky issue. Furthermore, measurement should take into account variations in control during the different steps of the evaluation process. Some kind of averaging scheme must then be devised to aggregate the different scores. Consequently, we believe coding the control dimension should be a matter of informed judgment, rather than mechanistic rule following. We do not argue that coding the control dimension should always remain a matter of informed judgment, only that it seems the most sensible option for now. We are confident that subsequent studies involving the application of this framework to real evaluations will contribute to clarifying these issues.

Table 6 : Coding scheme for control of the evaluation process

Level of membership	
Intuitive Label	Numerical
Exclusive control by evaluator and/or nonparticipating evaluation sponsor	.00
Limited/weak control by participants	.25
Shared control between participants and evaluator and/or nonpartcipating evaluation sponsors	.50
Substantial/Strong control by participants	.75
Exclusive control by participants	1.00

Note: "Participants" refer to nonevaluative stakeholders participating in the evaluation. Evaluation sponsors are considered "participants" if they are directly involved in the actual production of the evaluation.

2.3.4 Combining Dimensions to Measure Participation

Now that each dimension has been operationalized, a broader understanding of participation in the evaluation is needed. How do dimensions combine to form the basic concept of PE? Once aggregated, what does PE look like? The presence of *all* three fundamental attributes of participation is required for membership in the category of PE. To be qualified as participatory, an evaluation must therefore have a score equal to or greater

than .25 on each dimension and, consequently, at the basic level. This is the set-membership threshold that we have proposed. Other thresholds are possible but this one has the advantage of being sensitive to low levels of participation while still allowing differentiation with non-PEs. Score calculation at this level is straightforward: the *minimum* score of the secondary-level dimension determines the overall PE score (Goertz, 2006). Suppose an evaluation includes all types of nonevaluative stakeholders (thereby having a score of 1.00 for the diversity dimension). This evaluation would get an overall participation score of 1.00 *only if* stakeholders are involved in all evaluation tasks (extent of involvement = 1.00) *and* if they are in total control of decision making all along the way because of the authority and informal resources they possess (control = 1.00). If, however, the score for control is .50, the same evaluation would instead get a score of .50 at the basic level. If the score on the control dimension is .00, the evaluation would not be considered participatory, even though the evaluation gets a perfect score on the other two dimensions.[27] This is the necessary and sufficient condition logic.

Interestingly, the basic level coding scheme is also a participation index that can be used to rank different evaluations with respect to their level of participation and to distinguish participatory from non-PEs (see Figure 4). Scores reflect membership in the set of PE and their interpretation is similar to what has been done for secondary-level dimensions: .00 stands for no participation in the evaluation, .25 stands for weak/limited participation, .50 for moderate participation, and so on.

Let us now turn to the continuum formed by the basic level concept. The positive end, that is, where all dimensions are scored 1.00, can be conceived of as an *ideal type* (Goertz, 2006). It is not an ideal in the sense of a normative ideal that all evaluators should strive to attain, but rather in the sense that there are no (or very few) empirical cases that fit the type. Still, the ideal type is useful as a standard against which other types of PE could be more or less explicitly compared. We propose the label *self-managed democratic evaluation* for such an ideal type as it conveys the idea that all types or categories of nonevaluative stakeholders are totally in charge of the evaluation from beginning to end. In that case,

[27] For instance, all types of nonevaluative stakeholders could be involved in all steps of an evaluation as passive observers with absolutely no decision making power.

stakeholders not only act as subjects (as opposed to objects) of the inquiry but also as both evaluation sponsors and evaluators. To our knowledge, no empirical instance of self-managed democratic evaluation exists. In regard to the polar opposite of this ideal type, which is also the negative end of the concept, many labels would adequately capture the idea of nonparticipation. Although the terms *conventional* and *traditional* evaluation are appropriate, the label *technocratic* evaluation (Jacob, 2005; Murray, 2002) seems to better convey the idea of nonparticipation in which an evaluation is exclusively realized by evaluators and specialists of various methodological tools and techniques for decision makers. To the extent that evaluators interact with nonevaluative stakeholders, the latter are only considered sources of data, not participants. While no or very few empirical instances of self-managed democratic evaluation exist, technocratic evaluation seems more common.

2.4 Discussion: Promises and Limitations of the Conceptualization and Measurement Instrument

Building on the Cousins and Whitmore (1998) conceptual framework, this article has proposed an expanded and clearer conceptualization of PE. We argued that *diversity of participants*, *extent of involvement*, and *control of the evaluation process* are PE's constitutive dimensions and thus capture the essence of PE. Conceptually, we contend that these attributes are familiar to evaluators, display a high level of internal coherence, are parsimonious, and efficiently differentiate between participatory and nonparticipatory approaches to evaluation[28]. Moreover, the negative and positive poles of each dimension, as well as the continuum uniting them, have been coherently theorized. Operationally, a set of explicit indicators for each dimension and aggregation rules have been have been proposed, although clear measurement procedures of the control dimension still need to be developed. Although the measurement framework still has to be fully validated, we contend that the proposed empirical measures of PE seem to display a high level of concept-measure consistency, as well as a priori validity. On one hand, the proposed indicators seem to have *face validity* inasmuch as they measure what they are intended to measure (Singleton, Straits, & Straits, 1993). On the other hand, the selected empirical measures seem to have *content validity* in that they appear to adequately cover all facets or dimensions of PE, as

[28] Indeed, the necessary and sufficient conditions concept structure possesses a high discriminating power (Gerring, 1999).

defined by experts of the field through their writings (see Babbie & Benaquisto, 2002; Singleton et al., 1993).[29] The measurement device has indeed a strong theoretical base to support its dimensions.

A number of limitations also need to be acknowledged, however, with respect to the proposed framework. First of all, the operationalization of PE is still incomplete. Operationalization is not limited to the specification of indicators; it also entails identifying specific procedures for scoring each indicator (Singleton et al., 1993). This weakness is particularly striking for the control indicators as the specific procedures for scoring each indicator, for averaging the scores across the steps of the evaluation process, and for aggregating those scores into the second-level dimension have not yet been formalized. Consequently, the measurement instrument is not yet totally ready to use. Further studies on PE will be needed to derive a fully operational device. Second, despite what has been said above on content validity, two facets of participation have deliberately been excluded from the framework. With respect to the extent of involvement, the proposed conceptualization does not take into account the *intensity* of participation and, as such, does not exhaust the meaning of participation. Some stakeholders are indeed more enthusiastic, proactive, and affectively and intellectually engaged than others. In addition, we have focused on diversity of *types* of stakeholders as opposed to diversity *within* and *across* particular types of stakeholder groups in terms of values, opinions, socioeconomic characteristics, gender, ethnicity, and language. A last limitation pertains to the neglect of some sites of power. The extent of involvement dimension focuses on the technical tasks of the evaluation process but not on the political task of actually initiating (or not) an evaluation, writing the terms of reference or using the results. These sites of power are excluded from our analyses on participation. At the same time, however, tradeoffs are inescapable in any research endeavor. We have thus strived for a balance between simplicity and precision.

[29] Validity (i.e., the extent to which scores derived from an instrument reflect the construct that this instrument purports to measure) should not be confused with validation (i.e., the process of empirically testing whether an instrument produce scores which are valid). Furthermore, validity is increasingly conceived as a unitary concept (see sect. 1.1.3, supra).

We believe that the proposed conceptualization and instrument are going to be useful to both evaluation practitioners and researchers. From a practice perspective, the proposed framework may help evaluators to be more aware of their practices with respect to participation and foster reflection on them. For instance, the framework can be used in self-evaluation to compare the subjective beliefs of the evaluator (e.g., ‘‘I am deeply committed to PE’’) with a more objective measure of participation (e.g., ‘‘According to the framework, the last evaluation I have conducted was weakly participatory’’). This use of our framework is consistent with point 4 of the competence component of the American Evaluation Association (2004) *Guiding Principles for Evaluators* which states that:

> Evaluators should continually seek to maintain and improve their competencies, in order to pro-vide the highest level of performance in their evaluations. This continuing professional development might include formal coursework and workshops, self-study, evaluations of one's own practice, and working with other evaluators to learn from their skills and expertise.

The framework's utility as a self-evaluation tool is in no way limited to individual evaluators. Because many governments, international organizations, and departments claim that participation is an important ideal in evaluation, assessing whether this reflects the reality of practice would be illuminating.

The potential contribution of the framework for research is also important. A promising area of research relates to predicting and explaining the consequences of participatory approaches. Many studies on the relationship between PE and the use of evaluation and empowerment display a low level of precision in terms of measurement. Low comparability and generalizability of studies are the rule. We contend that the proposed device could foster quality research in this area by offering a more objective and systematic way to measure participation. Many research questions could benefit from the use of the proposed framework: Is the relationship between PE and evaluation use linear (i.e., does more participation always lead to greater use)? What is the effect size of this relationship? Do we observe differences for participatory effects between different streams of PE (P-PE vs. T-PE)? Are the *power relations among participating stakeholders* and *manageability of evaluation implementation* dimensions identified by the mediating factors for evaluation use of Weaver and Cousins (2004)? At a micro level, is evaluation influence a function of

the specific types of nonevaluative stakeholders and the specific steps in which they are involved?

Many calls urging evaluators to undertake further empirical research on evaluation in general (Christie, 2003; Mark, 2001; Smith, 1993) and on PE in particular (Cousins, 2001; Cousins & Earl, 1999; Mark, 2001) have been made in the past. The proposed conceptualization and instrument is a tool that has the potential to help evaluators meet this challenge.

3 Mesurer la participation des parties prenantes à l'évaluation: une validation empirique du Participatory Evaluation Measurement Instrument (PEMI)[30]

Résumé : ***Contexte.*** La participation des parties prenantes (*stakeholders*) est une tendance importante dans le champ de l'évaluation de programme. Même si quelques instruments de mesure de la participation ont été proposés, ils n'ont pas fait l'objet de validation empirique, ou encore ils ne couvrent pas complètement le contenu du concept. ***Objectifs.*** Cette étude consiste en une première validation empirique d'un instrument de mesure couvrant adéquatement le contenu de la participation, soit le *Participatory Evaluation Measurement Instrument* (PEMI). Elle examine en particulier 1) la fidélité intercodeur des scores obtenus par deux assistants de recherche à partir de cas d'évaluation rapportés dans la littérature; 2) la convergence des scores des codeurs et ceux de répondants-clés (c.-à-d., les auteurs des cas); et 3) la convergence entre les scores obtenus par les auteurs des cas d'évaluation sur le PEMI et l'*Evaluation Involvement Scale* (EIS). ***Échantillon.*** Un échantillon non probabiliste de 40 cas tirés de la littérature sur l'évaluation a été utilisé pour vérifier la fidélité. Un auteur par cas de l'échantillon a ensuite été invité à participer à un sondage; 25 questionnaires pleinement utilisables ont été reçus. ***Mesures.*** La participation a été mesurée sur des échelles nominales et ordinales. Le kappa de Cohen, le coefficient de corrélation intraclasse et le rho de Spearman ont été utilisés pour examiner la fidélité et la convergence. ***Résultats.*** Les résultats pour la fidélité vont d'acceptable (*fair*) à excellent. Les résultats de validation convergente pour les scores des codeurs et des auteurs vont de faible (*poor*) à bon. Les scores obtenus sur le PEMI et l'EIS présentent une association modérée. ***Conclusions.*** Les preuves empiriques de cette étude sont élevées dans le cas de la fidélité intercodeur et vont de faibles à élevées dans le cas de la validation convergente. Globalement, cette étude suggère que le PEMI peut générer des scores qui sont à la fois fidèles et valides.

Mots-clés : *évaluation participative; implication des parties prenantes; instrument de mesure; validation empirique.*

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

[30] Le titre original de cet article, tel qu'il a été soumis à *Evaluation Review*, est « Measuring stakeholder participation in evaluation: An empirical validation of the Participatory Evaluation Measurement Instrument (PEMI) » (l'article a été accepté pour publication et est, au moment d'écrire ces lignes, sous presse). Les auteurs désirent remercier tous les répondants pour leur générosité et leur intérêt envers cette étude. Nous désirons également remercier Marvin C. Alkin et Marie Gervais pour leur aide concernant le prétest du questionnaire, ainsi que David Collier et Nathalie Loye pour leurs précisions utiles concernant les différents types de validation. Merci à Stacie Toal pour ses précisions sur l'*Evaluation Involvement Scale*. La révision du manuscrit final a été effectuée par Kristen Leppington à qui nous souhaitons exprimer toute notre reconnaissance. Nous remercions le Conseil de recherches en sciences humaines du Canada (CRSH) pour son soutien financier. Les opinions exprimées dans cet article n'engagent toutefois que les auteurs. Ce projet a été approuvé par le Comité d'éthique de la recherche avec des êtres humains de l'Université Laval (2011-246/28-10-2011) (Annexe 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; measurement instrument; empirical validation.*

3.1 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, 2005a; 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 et al., 2011; Thayer & Fine, 2001).

Stakeholder participation is one of the major constructs that have 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[31] 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 current 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 et al., 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 et al., 2005).

3.2 Research Objectives and Hypotheses

This study consists of an empirical validation of the PEMI. Specifically, three objectives are pursued (see Figure 5):

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.

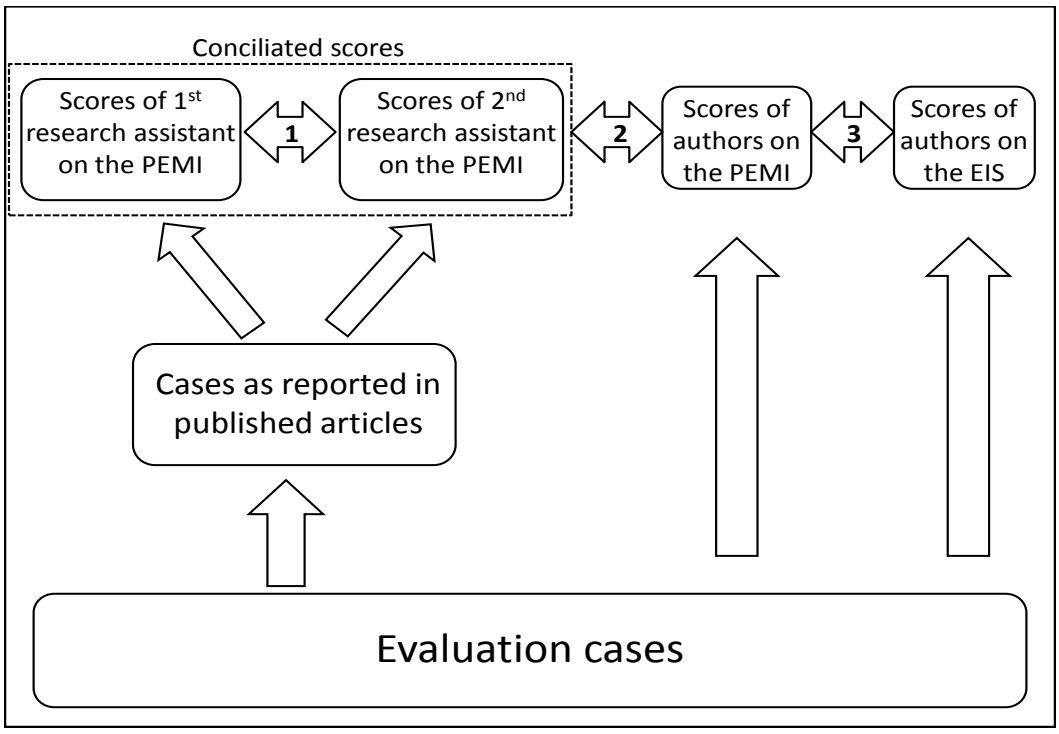
2) *Convergent validation I*. To examine the extent to which the scores achieved by two independent coders, once discrepancies are resolved through discussion, align with those of a key respondent for each case.

3) *Convergent validation II*. To examine the convergence between the scores obtained by key respondents on the PEMI and on the *Evaluation Involvement Scale*

[31] As a matter of fact, it is the *inferences* derived from the instrument which are validated, not the instrument itself. However, the expression "instrument validation" is common in the literature and much shorter so that it is used in this dissertation.

(hereafter EIS, Toal, 2009), a validated—albeit incomplete—measure of stakeholder participation.

Figure 5 : Schematic representation of validation objectives



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

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 (Objective 1, Figure 5), 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, Figure 5). 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 (as 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, Figure 5). First, 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). Second, 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.

3.3 Methods

3.3.1 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.[32]

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).[33] 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 not formally 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 Annexe B). 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: nonparticipatory 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 approach to evaluation 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

[32] 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.

[33] It was assumed that the articles were accurate reports of the evaluation process.

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).[34] 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 (one per case) to an online questionnaire (see Annexes D and E). 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 which 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).[35] In the end, 44 invitations to

[34] The first pilot test was based on 4 cases, not 3. 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.

[35] 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.

participate in the survey were sent, including noncontacts. A total of 25 fully completed surveys were received.[36] 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 56.8%.[37] This response rate compares well to those generally obtained through electronic surveys (Couper, 2000; Kwak & Radler, 2002; Millar & Dillman, 2011).

3.3.2 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. 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 & 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 Annexe C). 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

[36] 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.

[37] At 59.1%, 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 62.5% and 65%, respectively.

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 (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, 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: italic 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 ($\kappa_{DoP} = .00$; $ICC_{DoP} = -.19$) 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 7). 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 (see Annexe E). 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 incomplete measure of stakeholder participation. Why use the EIS if it does not capture all facets of 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.

3.3.3 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 (see Annexe F for individual results). 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 G Theory (*Generalisability 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 non parametric test appropriate for ordinal scales that makes few assumptions about the distribution of data (Newbold, 1995). 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).

3.4 Results

Intercoder reliability assessment. Results from the different rounds of coding are presented in Table 7. 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 inferences about a construct that is assumed to be stakeholder participation.

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 8). Even though the results for extent of involvement are not satisfactory, it must be noted that they are still better than what would be expected by chance alone (i.e., $\kappa = 0$). ICC results for the diversity of participants and extent of involvement dimensions are "poor" and, in the latter case, even failed 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.

Table 7 : 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 8 : 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

Convergent validation II. The authors' scores on the PEMI were compared to their scores on the EIS (see Table 9). 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.

Table 9 : Correlation (Spearman's rho) between scores derived from the PEMI and the EIS

	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

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)

3.5 Discussion

This study aimed at examining 1) whether the PEMI could be used by two coders on published evaluation cases to produce reliable inferences about stakeholder participation; 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.

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 10). 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 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 Annexe C) 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.

Table 10 : 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.

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.

4 Inattendues mais fort bienvenues : les méthodes mixtes pour la validation et la révision du Participatory Evaluation Measurement Instrument[38]

Résumé : Bien que la combinaison des méthodes qualitatives et quantitatives ne soit pas nouvelle, davantage d'études expliquant pourquoi et comment utiliser les méthodes mixtes à des fins de validation sont nécessaires. Cet article présente un cas de validation du Participatory Evaluation Measurement Instrument (PEMI), un instrument qui aspire à mesurer la participation des parties prenantes à l'évaluation. Initialement envisagée comme une étude presque exclusivement quantitative, une composante significative de celle-ci s'est transformée en une étude mixte. Cette transformation a à son tour généré un cycle de révision de l'instrument de mesure et de validation quantitative supplémentaire. Les résultats de validation suggèrent que la version révisée du PEMI est plus en phase avec l'opinion des répondants quant au niveau de participation de cas d'évaluation choisis. La valeur ajoutée des méthodes mixtes à des fins de validation est ensuite discutée à l'aide du raisonnement contrefactuel.

Mots-clés : *participation des parties prenantes, évaluation participative, conceptualisation, mesure et validation, Participatory Evaluation Measurement Instrument (PEMI), analyse thématique.*

***Abstract*:** Although combining methods is nothing new, more contributions about why and how to mix methods for validation purposes are needed. This article presents a case of validation of the Participatory Evaluation Measurement Instrument (PEMI), an instrument that purports to measure stakeholder participation in evaluation. Although the process was intended to be almost exclusively quantitative, a component of it unexpectedly turned into a mixed methods study. This, in turn, spurred on a cycle of instrument revision and further quantitative validation. The validation evidence suggests that the revised version of the PEMI offers a better fit with the respondents' opinions on the participation level of selected evaluation cases. The added value of mixed methods for validation purposes is discussed using counterfactual reasoning.

Keywords: *stakeholder participation, stakeholder involvement, conceptualization, instrument development and validation, Participatory Evaluation Measurement Instrument (PEMI), thematic analysis.*

---

[38] Le titre original de cet article, tel qu'il a été soumis au *Journal of Mixed Methods Research*, est « Unexpected but most welcome: Mixed methods for the validation and revision of the Participatory Evaluation Measurement Instrument ». Les auteurs désirent remercier tous les répondants pour leur générosité et leur intérêt envers cette étude. Nous désirons également remercier Marvin C. Alkin et Marie Gervais pour leur aide concernant le prétest du questionnaire, ainsi que Nathalie Loye pour nous avoir suggéré une référence utile sur la validation mixte. La révision du manuscrit final a été effectuée par Kristen Leppington à qui nous souhaitons exprimer toute notre gratitude. Nous remercions le Conseil de recherches en sciences humaines du Canada (CRSH) pour son soutien financier. Les opinions exprimées dans cet article n'engagent toutefois que les auteurs. Ce projet a été approuvé par le Comité d'éthique de la recherche avec des êtres humains de l'Université Laval (2011-246/28-10-2011) (Annexe A).

While collecting quantitative and qualitative data within the same study is nothing new, their meaningful integration into an explicit and coherent research methodology is a relatively novel research practice (Creswell & Plano Clark, 2007). Indeed, mixed methods research has only established itself recently as a distinct field of research, some would even say paradigm, on an equal footing with qualitative and quantitative research (Johnson, Onwuegbuzie, & Turner, 2007; Small, 2011). The following definition, based on a review of definitions from leaders in the field, clearly highlights that mixed methods research is much more than the mere collection of quantitative and qualitative data in the same study:

> Mixed methods research is the type of research in which a researcher or team of researchers combines elements of qualitative and quantitative research approaches (e.g., use of qualitative and quantitative viewpoints, data collection, analysis, inference techniques) for the broad purposes of breadth and depth of understanding and corroboration. (Johnson, et al., 2007, p. 123)

What holds for the field of mixed methods in general also holds for the specific subfield of empirical validation. Indeed, mixed methods validation requires the *combination* of qualitative and quantitative approaches for assessing the reliability and validity[39] of measurement instruments. Thus, even though "Campbell and Fiske (1959) are being rightfully credited as being the first to explicitly show how to use multiple research methods for validation purposes" (Onwuegbuzie, Bustamante, & Nelson, 2010, p. 114), their research orientation was first and foremost quantitative. In other words, qualitative data were only granted an auxiliary role. By contrast, many of the recent empirical studies which have combined qualitative and quantitative research for validation purposes are explicitly grounded in the mixed methods paradigm (e.g., Arnon & Reichel, 2009; Durham, Tan, & White, 2011).

Assessing the validity and fidelity of measurement instruments is a common rationale advocated by researchers for using mixed methods (Collins, Onwuegbuzie, & Sutton, 2006). At the same time, however, mixed methods validation has been described as an underdeveloped subfield: "…clearly, more publications are needed that outline explicitly ways of optimizing the development of instruments by mixing qualitative and quantitative techniques" (Onwuegbuzie, et al., 2010, pp. 57-58). Philosophical and practical obstacles

[39] On the validation of *inferences* derived from measurement instruments, see the precision presented in n. 29, *supra*.

are indeed credited for having hindered the development of mixed methods for validation purposes (Luyt, 2011). Despite these difficulties, a few theoretical frameworks have been proposed to help mixed methods researchers in their validation endeavours (Dellinger & Leech, 2007; Luyt, 2011; Onwuegbuzie et al., 2010). For instance, Dellinger and Leech (2007) have proposed a unified validation framework for mixed methods research. However, we remain somewhat unconvinced by the value of creating new validity criteria in a domain that is already overwhelmed by a plethora of different concepts and terms. Beyond the particular terms used to talk about validation and validity, we favour a unified concept of validity that spans across research traditions and paradigms.

The specific guidance offered by Luyt (2011) and Onwuegbuzie et al. (2010) about how to conduct mixed methods validation was, by contrast, a much-needed contribution. Both frameworks consider the validation process as made of a number of steps (phases or stages) that are linked together by a cyclical logic. The *Instrument Development and Construct Validation* (IDCV) developed by Onwuegbuzie et al. (2010, p. 10) contains 10 phases:

1. Conceptualize the construct of interest
2. Identify and describe behaviors that underlie the construct
3. Develop initial instrument
4. Pilot-test initial instrument
5. Design and field-test revised instrument
6. Validate revised instrument: Quantitative analysis phase
7. Validate revised instrument: Qualitative analysis phase
8. Validate revised instrument: Mixed analysis phase: Qualitative-dominant crossover analyses
9. Validate revised instrument: Mixed analysis phase: Quantitative-dominant crossover analyses
10. Evaluate the instrument development/construct evaluation process and product.

The framework proposed by Luyt (2011) builds on Adcock and Collier (2001) and contains three interrelated stages: 1) measurement development (i.e., concept definition and operationalization); 2) measurement validation (i.e., assessing the extent to which the scores derived from the measurement instrument are meaningfully related to the concept);

and 3) measurement revision (i.e., limited alterations and extensions to the concept based on the results of the second stage).

Although the validation frameworks by Onwuegbuzie et al. (2010) and Luyt (2011) can be used prescriptively (i.e., they can guide researchers through the validation process), their analytical and heuristic value should not be underestimated. Indeed, they can be used to make sense of the validation process and generate new understanding. These frameworks will therefore be used to describe the process followed in a research program which has aimed at developing a sound conceptualization and measurement instrument of participatory evaluation (see Table 11).

Table 11 : Use of different frameworks to illustrate the steps of a research program on participatory evaluation

Steps in PEMI development and validation	*Phases in Instrument Development and Validation Process (Onwuegbuzie et al., 2010)*	*Stages in Luyt's Framework (2011)*
1. Conceptualization and operationalization of stakeholder participation	1. Conceptualize the construct of interest 2. Identify and describe behaviors that underlie the construct of interest 3. Develop initial instrument	1. Measurement development
2. Applications of the initial instrument	4. Pilot test initial instrument	None specified
3. Quantitative validation of the initial instrument	6. Validate (revised) instrument: Quantitative analysis phase	2. Instrument validation
4. Mixed methods validation of the initial instrument	8. Validate (revised) instrument: Mixed analysis phase: Qualitative-dominant crossover analyses	2. Instrument validation
5. Instrument revision	5. Design (and field-test) revised instrument	3. Instrument revision
6. Quantitative validation of the revised instrument	6. Validate revised instrument: Quantitative analysis phase	2. Instrument validation

Note: PEMI stands for Participatory Evaluation Measurement Instrument. Words within parentheses indicate the original formulation of the step.

4.1 Research Purpose

This article presents a case of empirical validation of an instrument that purports to measure stakeholder participation in evaluation, namely the *Participatory Evaluation Measurement Instrument* (PEMI). Although the validation process was intended to be

conducted almost exclusively within the quantitative research paradigm, a significant part of it unexpectedly turned into a mixed methods study. Indeed, the unexpected richness of the respondents' comments to an open-ended question and informal email correspondence has spurred on an unplanned cycle of mixed-methods analysis, instrument revision and further quantitative analysis. The purpose of this article is twofold. First of all, it reports substantive results pertaining to the PEMI's validity and proposes a revised version of this instrument that is more in line with the data. In addition, it shows and discusses how mixed methods research can play an unexpected but valuable role in a rigorous research program.

4.2 Conceptualization and Operationalization of Stakeholder Participation

In the field of program evaluation, participatory approaches such as utilization-focused, empowerment and democratic evaluation are clearly on the rise: "One of the larger trends in evaluation theory and practice is an increased focus on stakeholder participation" (Mark, 2001, p. 462). The downside of this period of expansion is that the concept has been used inconsistently by theorists, researchers and practitioners (Daigneault & Jacob, 2009). The meaning of participation is ambiguous, the terms used to describe it are multiplying and, in practice, the demarcation line between participatory and nonparticipatory evaluation is frequently blurred.

To remedy these problems, Daigneault and Jacob (2009) have proposed a "mixed methods conceptualization" of stakeholder participation, which draws on and adapts the Cousins and Whitmore (1998) framework. This conceptualization was developed in two stages. First, the structure of the Cousins and Whitmore framework was improved and operationalized using the qualitative tools of classical logic (i.e., necessary and sufficient conditions) and fuzzy logic (i.e., quantification of membership to sets which represent constitutive dimensions and the concept itself) (see Goertz, 2006) and a set of criteria for assessing concepts (Gerring, 1999). Second, an argument based on a qualitative literature review on stakeholder participation was developed to demonstrate the value of the new conceptualization.

The conceptualization developed by Daigneault and Jacob (2009) was operationalized in an instrument in which stakeholder participation is measured on an ordinal scale (.00 = no participation; .25 = limited participation; .50 = moderate participation; .75 = substantial participation; 1.00 = full participation). Two ideal types were also proposed for the negative and positive ends of the continuum, namely *technocratic evaluation* and *self-managed democratic evaluation*. A similar five-point scale was used to measure each of the three constitutive dimensions of participation: *diversity of participants* (i.e., Which non-evaluative stakeholders are involved in the evaluation?), *extent of involvement* (i.e., At which steps of the evaluation process does involvement occur?), and *control of the evaluation process* (i.e., What is the distribution of power between the evaluator and stakeholders?). Whereas the coding of the last dimension involves a subjective judgment as to the respective level of control of evaluators and participants, the other two dimensions' scores are each determined by four dichotomous indicators. These indicators measure the presence/absence of four types of participating stakeholders and involvement/non involvement in four steps of the evaluation process. Since each of these dimensions is considered necessary to the concept of participation, the overall level of participation of an evaluation is determined by the dimensions' *minimum* score (i.e., the lowest score) (Goertz, 2006). For instance, if an evaluation is attributed a score of 1.00 for both diversity of participants and extent of involvement, as well as .00 for control, it is considered nonparticipatory (i.e., it has a score of .00 for the overall level of participation).

It is important to stress that this conceptualization/instrument does not have a normative orientation. It is a neutral instrument that empirical researchers can use to test various hypotheses involving stakeholder participation. For instance, the instrument can be used to test the relationship between participation and various "negative" outcomes such as increased conflict between stakeholders or an evaluation of lower methodological quality.

4.3 Applications of the Initial Instrument

The conceptualization developed by Daigneault and Jacob (2009) appears to rest on solid theoretical foundations and gave birth to a few promising applications (Connors & Magilvy, 2011; Jacob, Ouvrard, & Bélanger, 2011; Laudon, 2010). Although not specifically intended for that purpose, these applications are somewhat equivalent to an

informal, qualitative, pilot test of the initial instrument (see Table 11). The commonalities between these applications and the formal pilot test described by Onwuegbuzie et al. (2010) should not be overstated, however:

> This field test optimally would represent a mixed research study to assess the appropriateness of each item. In particular, each item should be assessed for clarity, esthetics, relevancy, tone, length of time needed for a response, and, above all, cultural competence […] [T]he focus at this phase should be more on the content-related validity (i.e., face validity, sampling validity, item validity) and two elements of construct-related validity (i.e., outcome validity, generalizability) of the initial instrument... (p. 64)

Since this "step" generated exclusively positive comments (see Connors & Magilvy, 2011), no need for instrument revision was felt. It was thus decided that the authors would pursue a full-scale quantitative validation of Daigneault and Jacob's framework.

4.4 Quantitative Validation of the Initial Instrument

The next logical step was to assess whether the PEMI could produce reliable and valid scores related to the concept of stakeholder participation. This was done by Daigneault, Jacob & Tremblay (forthcoming) using a purposive sample of evaluation cases published in peer-reviewed journals (final sample: n = 40). The cases were coded independently by two research assistants. The authors of the articles reporting the evaluation cases were then invited by email to participate in a survey (fully usable questionnaires: n = 25). The survey had two main sections. The first section focused on the PEMI. Respondents had to check boxes about types of participants, steps in which they were involved and their level of control on the evaluation process. Five-point indices were derived from respondents' answers to the PEMI for each dimension and for the overall level of participation. These scores were presented to the authors of the evaluation cases for reactions in the first part of the survey. Respondents' opinions were measured on an ordinal scale of agreement and an open-ended question asked respondents to justify their choice (i.e., "Why?"). The second part of the survey contained 11 questions from which the calculation of the Evaluation Involvement Scale was derived (Toal, 2009).

The respondents' scores on the PEMI were compared with those of the coders for agreement, on the one hand, and with their scores on the Evaluation Involvement Scale for

correlation, on the other. Based on the results, the authors concluded that the evidence was strong for reliability and ranged from weak (but positive) to strong for convergent validation. The authors concluded that the PEMI appears to generate scores which are both reliable and valid (see Chapitre 3).

4.5 Turning to Mixed Methods for the Validation of the Initial Instrument

The validation study discussed above was entirely conducted from a postpositivist and quantitative perspective. Although the survey included one open-ended question, it was not exploited in the article that reported the results of this study (see Chapitre 3) for reasons that will shortly become clearer. In retrospect, the purpose of including this qualitative question into a quantitative questionnaire was probably aimed at "elaboration, enhancement, illustration, clarification of the results from one method with the results from the other method" (Greene, Caracelli, & Graham, 1989, p. 259). However, we expected that only a few respondents would answer the open-ended question since they could easily skip it and only very brief comments for those who would choose to answer. The fact that no sophisticated method had been planned beforehand to analyze open-ended answers also testifies to our low expectations toward the richness of data that would be collected from this question. Indeed, the importance of the open-ended question was exclusively envisioned in connection to the quantitative score of agreement. Provided that the open-ended question would have been exploited, this case would probably fit the "study employing minimum qualitative research" category according to the definitions of mixed methods research put forward by Creswell and Plano Park (2007, p. 11). Although clearly a borderline case, it nevertheless satisfies the minimum criteria to be considered mixed methods research. In our opinion, this study would probably have fallen between the "Pure Quantitative" and "Quantitative Mixed" types on the continuum proposed by Johnson et al. (2007), and probably nearer the former than the latter.

The "barely qualitative" component of this study unexpectedly turned into a mixed methods study when it was decided to analyze thematically qualitative data. It must be stressed that the quantitative and qualitative components were not of equal status, however. Despite an expanded qualitative component, this study would now be located a bit to the

left along the Johnson and colleagues' (2007) continuum and now clearly considered of the "Quantitative Mixed" type. Indeed, the open-ended question (i.e., "Why?") connected to the quantitative score of agreement and generated abundant qualitative data. Quantitatively, a strong majority of respondents (91.6%) formulated comments. Two respondents also substantively supplemented their answers by email. Taken together, these two sources of data generated about 1,400 words of data (M = 60 words; Min = 9; Max = 242). A first inspection clearly revealed that, despite its limited size as a qualitative corpus, the data's content was rich and varied; a superficial browse through the data would clearly prove insufficient to extract all the meaning out of it. It was thus decided that a full-scale thematic analysis would be conducted. Indeed, this pragmatic approach which aimed at identifying patterns within qualitative data appeared to be a good fit in our situation: "Through its theoretical freedom, thematic analysis provides a flexible and useful research tool, which can potentially provide a rich and detailed, yet complex, account of data" (Braun & Clarke, 2006, p. 78).

4.5.1 Quantitative Component

When questioned about the participation scores that were generated from their answers to the PEMI ("Do not agree at all" = 1, "Agree to some extent" = 2, "Totally agree" = 3)[40], respondents indicated on average that they "somewhat agree" with it (M = 2.04, Mdn = 2, n = 24). Although not as strong as expected, this result still provides positive evidence with respect to the validity of the PEMI. Agreement with the PEMI-generated score can indeed be interpreted as the fact that the instrument "truly" measures stakeholder participation. Since it is impossible to directly observe a concept — if this were the case, measurement instruments such as the PEMI would be useless — the validation process involves relying on indirect evidence. In that regard, the favorable opinion of key respondents on the participation score generated by the PEMI for their evaluation case plays an important role in ensuring that the instrument really measures stakeholder participation. That being said, it is worth stressing that subjective opinion of respondents is not the only source of evidence used for validation purpose.

[40] Respondents could also choose "I don't know/I don't want to answer". One respondent chose that option and was not included in the averaged score.

4.5.2 Qualitative Component: Approach and Methods

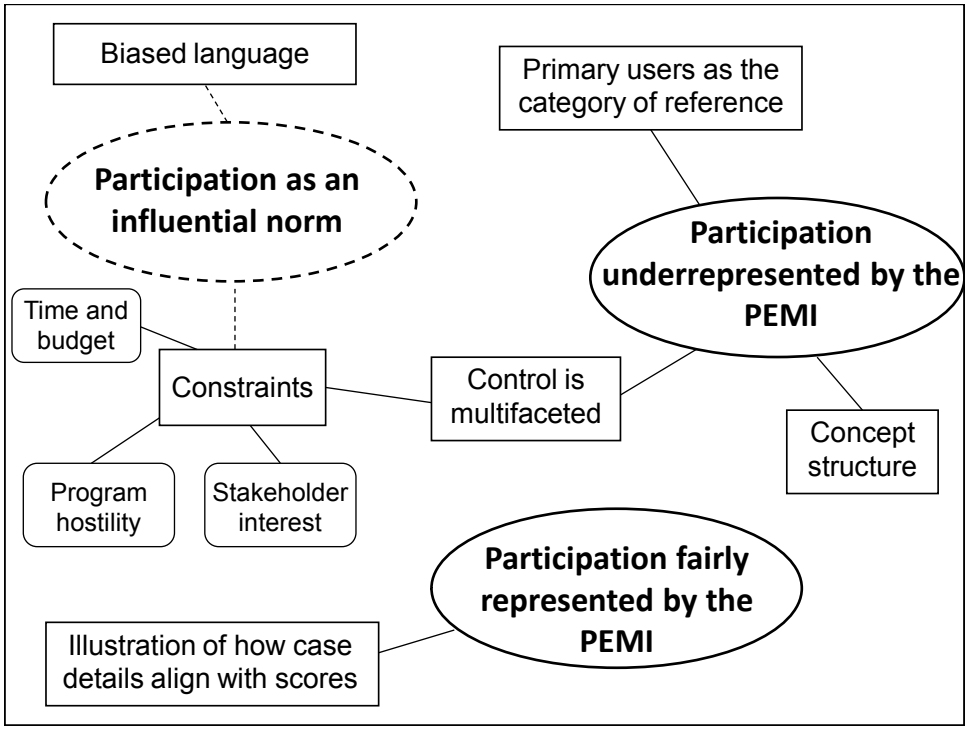
The thematic analysis conducted here closely follows the guidelines and steps proposed by Braun and Clarke (2006, p. 87): 1) Familiarization with the data; 2) Initial coding; 3) Searching for themes; 4) Reviewing themes; 5) Defining and naming themes; and 6) Reporting. Furthermore, we clarified our epistemological stance and analytical strategy before data analysis, as recommended by Braun and Clarke (2006). The analysis first followed an essentialist/realist stance according to which meaning and experience are assumed to be directly mediated by language. Second, the thematic analysis was primarily guided by a semantic approach which focused on the explicit meaning of the data. We say "primarily" because the analysis also identified a latent theme. Third, the analysis was conducted from both a theoretical (i.e., top-down) and inductive (i.e., bottom-up) perspective. While the coding was guided by a set of initial codes, they evolved during the coding process and additional codes and themes were identified from the data.

Data from open-ended comments and email correspondence were first transcribed in a single text document and edited. Weft QDA (Fenton, 2006), a free open-source software for qualitative analysis, was then used by the first author to analyze the data. After perusing the document a few times, a set of initial codes (i.e., expression of agreement/disagreement, justification for agreement/disagreement) were used to code the data. The process of searching for, reviewing and defining themes involved many iterative cycles of analysis; "A theme captures something important about the data in relation to the research question, and represents some level of patterned response or meaning within the data set" (Braun & Clarke, 2006, p. 82). While overarching themes satisfied both of these criteria (i.e., significance and prevalence), themes and subthemes were mainly identified on the basis of their significance in the context of this study due to the limited size of our data set. To ensure that the results from the thematic analysis were a faithful and complete representation of the data, the second author read through the data set and systematically reviewed the first author's coding and interpretation. His thorough review contributed to a clarification of the coding categories and a refinement of the themes. This also stimulated discussion about data interpretation and led to a revised version of the thematic map (see Figure 6 for final version).

Qualitative Component: Results

Three overarching themes were identified from the data: (1) Participation fairly represented by the PEMI; (2) Participation underrepresented by the PEMI; (3) Participation as an influential norm (Figure 6). These overarching themes and their associated themes and subthemes will be discussed sequentially.

Figure 6 : Final thematic map based on respondents' reactions to PEMI scores



Participation fairly represented by the PEMI. The first overarching theme clusters data which indicate a positive alignment between the respondents' scores on the PEMI and their opinion about the level of stakeholder participation for their case. Explicit comments such as "I believe that the representation is quite accurate", "The score feels right to me […] So yes, I agree with the score" and "It's about right" suggest that the measurement instrument faithfully reflects their opinion about stakeholder participation. It must be stressed, however, that the comments indicate partial rather than total agreement. In that respect, this overarching theme corroborates the quantitative finding according to which agreement was not perfect. Respondents whose answers cluster around this overarching theme justified their answers for the most part by citing details from their case which support the scores

they obtained on the PEMI (i.e., *illustration of how case details align with scores*). For instance, a respondent stated, "We work collaboratively with all stakeholder parties, our goal is partnership and sharing control in all phases". Another respondent—who obtained a score of .50 (moderate participation) on the instrument for overall participation and the control dimension—commented:

> The process employed involved a deliberate sharing of control – stakeholders were the decision-makers regarding the substantive foci of the evaluation. The evaluators retained control over the technical and methodological aspects of the evaluation. The evaluators also planned and implemented the participatory processes.

Participation underrepresented by the PEMI. It is interesting to note that respondents who expressed disagreement with their participation score unanimously thought that the latter was too low. In other words, the PEMI underestimated their opinion about the level of stakeholder participation for their case. Comments such as "Participants more influential in outcomes than this scale indicates" and "I would tend to rate the participation higher through most of the evaluation process" are clear illustrations of this disagreement. Similarly, another respondent stated: "I don't get it. I was in the role of input and consultation at the level of workshop facilitator to folks who had projects on the go. The projects were highly participatory from the point of view of nonevaluator stakeholder participation".

To paraphrase the famous opening sentence from Leo Tolstoy's Anna Karenina, happy respondents resemble one another, but each unhappy respondent is unhappy in its own way. Of course, all disagreeing respondents thought that the score somewhat underrepresents the "true" level of participation of their case but, beyond that, they all gave different reasons as to why this was so. A first theme related to underrepresentation of stakeholder participation highlights the fact that the *concept structure* underlying the PEMI's scores is counterintuitive. Indeed, a respondent could not understand why he or she got such a low overall score for participation when, at the same time, the scores obtained on other dimensions were so high: "I don't get why the overall score was .25 with a 1 and a 0.75 in the mix".

Another theme derived from the data to justify the respondents' disagreement, *primary users as the category of reference*, refers to how the diversity of participants dimension is conceptualized and operationalized. A respondent argued: "Stakeholder participation was limited to primary users. I don't really think involving, for example, intended program beneficiaries makes the process more participatory". Another respondent—who did not agree at all with the overall participation score obtained—explained: "Our evaluation was done with the full participation of those who were being evaluated". In effect, this theme calls into question one of the PEMI's fundamental assumptions: "The more diverse the types of stakeholders involved, the more participatory an evaluation is, all other things being equal" (Daigneault & Jacob, 2009, p. 341). These comments can be interpreted as calling for a reconceptualization of stakeholder participation that would not be based on all possible stakeholder types, but only on primary users (i.e., generally managers and program personnel).

Control is multifaceted is the last theme derived from the respondents' answers to justify their opinion regarding the PEMI's underrepresentation of participation. This theme brings attention on a conception of control that differs from the one underlying the measurement instrument. A first facet of control is the fact that the various steps of the evaluation process are not equally important to all respondents. Whereas the PEMI assumes equal weighting of each step, a respondent argued that the potential for influence is greater at the beginning and at the end of the evaluation:

> Stakeholders had considerable input into framing questions that then directed the entire study. They also were involved in interpreting the results of the study. I give these evaluation components considerably more weight than the actual conduct of the study.

The concept of "critical junctures" raised by another respondent is also related to this idea of the disproportionate importance of some evaluation steps. Moreover, the influence exerted by people *not* directly involved in the evaluation as an important facet of control was raised as illustrated by the following comments by two respondents:

> …[S]ome of the decision-making was limited from the very beginning since the program was funded by the federal government and included a prescribed cross-site evaluation component...

> There is the question of how the task was developed and how the stakeholders' meetings were shaping the results. Though they had a limited control in deciding the topics for discussion, the results of the evaluation were continuously discussed and negotiated.

Similarly, another respondent emphasized that the format of an evaluation is often determined by evaluation sponsors: "The external 'technocratic' evaluation was a political request…". Influence is exerted not only before the actual conducting of an evaluation but also after: "… [T]he city officials were the ones who agreed to the process, supported it, *and had to decide how to use the results*, so even if it appeared to be shared participation, those in power had ultimate control in the end" (emphasis added). These last comments all point to the importance of considering the other, less visible, face of control, that is what is decided before and after the evaluation by evaluation sponsors and institutional features.

Participation as an influential norm. This third overarching construct identified is latent. Latent themes go beyond what has been said and "examine the *underlying* ideas, assumptions or conceptualizations" (Braun & Clarke, 2006, p. 84) contained in data. Many respondents explicitly mentioned or alluded to the normative power of stakeholder participation. Participation was either embraced or criticized by respondents. Two themes were related to this latent construct: *biased language* and *constraints*. The first theme relates to data criticizing what is perceived as a democratic orientation in the PEMI:

> Having become familiar with [the survey], I need to emphasize that it does appear either to be biased in favor of complete or least very substantial participation by program offices. One even senses a democratic theme in the background. This is just too idealistic.

The second theme underlines the constraints faced by evaluators who want to put forward a participatory approach. The following comment is particularly relevant in that regard: "…[I]deally there could have been more participation, particularly by end users—but this wasn't possible in the time frame/budget". The lack of interest of stakeholders was also considered a significant constraint on participatory practices:

> I had originally envisioned that the consumers would take on more of an active role in designing, implementing, and reporting on the evaluation. The consumers were not very interested in the above, but they were very interested in expressing their opinions about the evaluand and in discussing and interpreting the results of the report.

Similarly, another respondent stated: “Stakeholders' participation was merely incidental, there was no interest at all in promoting organizational change”. If lack of interest from stakeholders can naturally hinder participatory evaluation, outright hostility can be an even more important constraint on participation:

> The score vastly—I emphasize vastly—underestimates the volatile and hostile nature of the relationship between the evaluation function and the program function. Thirty years ago when this evaluation was undertaken it was considered a wildly risky thing for the evaluation function to allow the program function to be present from design through analysis and recommendations. […] Most often the program office being evaluated is exceedingly hostile. Not just a little. […]. Your methodology in the first part classifies this part as not being very democratic or some such. Naive in the extreme. Thirty years ago this was radical.

A last constraint on participation mentioned by respondents related to the multifaceted nature of control presented earlier. Indeed, sponsors and institutional rules can orient and limit the extent to which an evaluation is participatory. Taken together, these citations are indicative of the normative power of participation.

If one takes a look at the thematic map from a “big picture” perspective, what general conclusions can be formulated with respect to the validity of the measurement instrument? A first point to mention is that stakeholder participation seems to be somewhat underestimated by the PEMI. A significant percentage of respondents indeed believed that the level of stakeholder participation in their case was higher than indicated by the scores derived from the measurement instrument. Second, the content of the stakeholder participation construct appears to be well-represented by the dimensions of extent of involvement, diversity of participants and control of the evaluation process. Indeed, no respondent argued that one of the dimensions was superfluous, nor did any respondent recommend adding extra dimensions to the instrument. This is an important finding even though respondents were not specifically prompted on this topic. Of course, criticisms were voiced—almost exclusively by respondents who thought that their score was too low—on the specifics of each dimension’s conceptualization and operationalization, but they did not represent a fundamental challenge to the relevance of the dimensions themselves. Third, one cannot disregard the normative power of stakeholder participation. Even a measurement instrument like the PEMI that is aimed at producing valid inferences for empirical purposes is not immune to the fact that participation represents a powerful norm.

In the end, the thematic analysis suggested that the PEMI appears to provide valid measures of stakeholder participation but, at the same time, the instrument could be revised to better take into account certain limitations raised by respondents.

4.6 Instrument Revision

It must be noted that the process of instrument quantitative validation, mixed methods validation and revision was not as linear as suggested by the order of presentation of the different sections. In fact, the issue of replacing the minimum rule for determining the overall level of participation had been raised earlier in the article in which the full results of the quantitative validation phase are reported (see Chapitre 3). The two research assistants in charge of coding then suggested that averaging the three dimensions of the PEMI would better reflect their intuition about stakeholder participation than using the minimum which was overly restrictive. Yet, the results generated in the mixed methods phase provide additional evidence of the necessity of revising the PEMI. Quantitative data indicated that the alignment between the scores derived from the PEMI and the respondents' opinions with respect to the level of participation was only partial. Qualitative data revealed that an overarching theme derived from the respondents' answers was the underrepresentation of stakeholder participation by the PEMI's scores, thus suggesting the need to find a less conservative concept structure.

The PEMI was therefore revised with respect to how the overall level of participation is derived from the dimensions of diversity of participants, extent of involvement and control of the evaluation process. In the original version, stakeholder participation is simply the minimum or lowest score of the three dimensions because each of the latter is considered a necessary condition for participation (for the full argument, see Daigneault & Jacob, 2009). In the revised version, determining the overall score is a two-step process which corresponds to a hybrid concept structure (see Goertz, 2006, p. 36). First of all, each dimension is still considered a necessary condition for distinguishing participatory evaluation from nonparticipatory evaluation. This premise implies that if any dimension is missing (i.e., has a score of .00) in a case, this case is not an instance of participatory evaluation. Once this minimal criterion is satisfied, the overall participation score is simply derived from the average score on the three dimensions (and, if needed, rounded *down* to

the nearest score). If we have, for instance, a case with full diversity and full involvement but limited control of the evaluation process by stakeholders (i.e., scores of 1.00, 1.00 and .25 respectively), the overall participation score would be .75 under the revised version instead of .25 under the original version. The revised rule of aggregation does not always affect the overall participation score, but it nevertheless mitigates the conservative bias of the PEMI. In fact, a similar rule had been suggested in the earlier work of Daigneault and Jacob (Daigneault & Jacob, 2007).

We have already mentioned that the PEMI is an objective measurement instrument. Yet, it seems that respondents have high expectations toward their score for overall participation that cannot only be explained by a conservative bias in the PEMI. The dominant discourse on stakeholder participation—which has even been qualified as "participatory orthodoxy" by some (i.e., Biggs, 1995 as cited in Gregory, 2000, p. 180) —might be an explanation for these expectations. In addition, a score below 1.00 on the scale might naturally orient respondents toward reflexivity and self-improvement. That being said, we admit that the two ideal types that have been proposed for the negative and positive ends of the continuum (i.e., *technocratic evaluation* and *self-managed democratic evaluation*) have strong connotations and, as a result, we think they should be abandoned. The language used in the PEMI will thus be more neutral and objective.

Other revisions to the PEMI were considered on the basis of the themes derived from the qualitative data but were rejected in the end. A first option was to refocus the conceptualization of the diversity of participants dimension exclusively on primary users instead of all potential types of users. This was rejected. For one thing, we argue that it is fundamental for an instrument that purports to measure participation (such as the PEMI) to be applicable to various evaluation contexts and rationales for participatory evaluation, whether political, pragmatic or epistemological (Cousins & Whitmore, 1998). For another, defining primary users is context-specific and involves judgement. If primary users have to be identified on a case-by-case basis, this would significantly complicate the application of the instrument. By contrast, the objective and well-delineated typology of stakeholders that actually characterizes the instrument contributes, in our opinion, to a reliable and valid coding.

Other considered but rejected revisions were linked to the control dimension. We sympathize with respondents' grievances about the operationalization of control but, at the same time, we feel that differential weighting of control based on which steps of the process control is exerted is unrealistic. The most important obstacle would be context: a specific step such as data collection and analysis can represent an occasion to exert more control in one case and less in another. Besides, the actual operationalization of control is much simpler to use. Similarly, the idea of adjusting control to take into account the influence of evaluation sponsors and institutions is tempting but difficult to apply in practice. For these reasons, we argue that the actual operationalization of control is satisfactory.

4.7 Quantitative Validation of the Revised Instrument

The revised version of the PEMI was then subjected to a phase of "pure" quantitative validation (see Johnson, et al., 2007). A first step was to examine how the *original* overall participation scores of respondents would be affected by the proposed revision to the instrument. For the 24 respondents who expressed an opinion on their scores, the revised version of the PEMI would leave 6 (20%) scores unchanged and would increase 18 (80%) of them. In the latter group, 17 respondents would see their score increase by one point on the scale while the other would see a two-point increase.

The original scores of these two groups of respondents, respectively those who would be affected by a revision and those who would not, can be compared *before* the actual revision of the PEMI to check if they are equally likely to agree with their scores. These two groups are formed retrospectively on the basis of the anticipated impact of a revision to the PEMI and the analysis was conducted on the original data. According to the qualitative findings of the mixed methods phase, the level of agreement is expected to be lower for the affected group than for the unaffected group. A Mann-Whitney U test was used to evaluate this hypothesis. The results of the test were in the expected direction and statistically significant ($U = 28$, $p = .017$, exact significance, one-tailed). The group that would *not* be affected by the PEMI's revision ($M = 2.5$, $Mdn = 3$, $n = 6$) had an average rank of 16.83 while the would-be affected group ($M = 1.88$, $Mdn = 2$, $n = 18$) had an average rank of 11.06. This finding can thus be interpreted as supporting the proposed revision to the PEMI.

A follow-up was conducted informally by email from the 11th to the 21st of January 2012 (two reminders were sent to nonrespondents) to assess whether the respondents who would see their participation score increase with the revised version of the instrument would prefer the revised score over the original score. The 18 authors in the "would-be affected group" were contacted and presented with the new overall participation score for their case and were asked two questions[41]:

> (1) Does this new score correspond better, equally or less to your opinion about the level of stakeholder participation for this particular case?
>
> (2) Using the following scale (1 = Do not agree at all, 2 = Agree to some extent, 3 = Totally agree, 4 = Don't know / Don't want to answer), to what extent would you say the new score corresponds to your opinion about the level of stakeholder participation for this case?

The first two questions may look redundant but they actually played complementary roles. To maximize comparability, it was important to use the same ordinal scale of agreement as in the original study. However, the original scale was somewhat "crude," that is, it could not capture as many nuances in the opinion of respondents as traditional five-point or seven-point Likert scales. For instance, a respondent who "partly agrees" with both the original and the new scores could nevertheless prefer one score over the other. We hypothesized that respondents would prefer the revised score over the original score and would therefore be more likely to agree with it.

All contacted persons but one replied to this targeted follow-up survey. However, only 16 questionnaires were usable (i.e., one respondent chose the "don't know/don't want to answer" option. In a few cases, clarifications on respondents' answers were sought by email. In any case, the logic, type of data collected and analysis strategy were all quantitative. The participation rate (88.9%), which was calculated according to the American Association for Public Opinion Research's *Standard Definitions* RR1 (AAPOR, 2011), was high.

---

[41] A third question was posed to contextualize previous answers but, in the end, was not exploited: "If you are *not* in total agreement with the new score, which option would offer the best fit for your opinion (0 = no participation, 0.25 = limited participation, 0.50 = moderate participation, 0.75 = substantial participation, 1.00 = full participation)?"

The comparison of the original results and follow-up results conforms to a before-and-after, quasi-experimental logic. Whereas results go in the expected direction for both hypotheses, they fail to attain statistical significance. In terms of the direction of preference, 10 respondents indicated that they preferred the revised score to the original (+), 5 indicated that they preferred the original (-) and 2 said that they equally liked both scores (=). However, the result from a sign test on the hypothesis that respondents prefer the revised score over the original score was not statistically significant (p = .151, exact one-tailed). Statistically, this should be interpreted as "no evidence of effect" rather than "evidence of no effect" (see Littell, Corcoran & Pillai, 2008), especially as this result is based on a small sample size. In terms of the average intensity of preference, respondents seemed to agree more with the score derived from the revised version than with the original score on average (M = 2.21 vs. M = 1.89). A Wilcoxon signed rank test was used to assess the hypothesis that respondents prefer the revised score over the original. Again, the results were not significant (z = -1.136, p = .17, exact one-tailed). The sum of the ranks in favour of the original score was 21.00 while the sum of the ranks in favour of the revised score was 45.00 (n = 11) [42]. To sum up, the respondents seemed to favour the revised score over the original score on average, but it cannot be ruled out that this improvement could be due to chance (once again, this result cannot be interpreted as evidence of no effect, see Littell, Corcoran & Pillai, 2008). Indeed, there is a high probability that the limited sample size has prevented the attainment of a statistically significant result.

A last statistical test was conducted to check whether the level of agreement with the revised participation scores for the unaffected and affected groups was now equal.[43] The hypothesis was that there is no significant difference between the two groups with respect to their opinion of the scores derived from the revised version of the PEMI. A Mann-Whitney U test was used to evaluate this hypothesis. The results of the test confirmed that there is no statistically significant difference between the level of agreement of the two groups (U = 36.5, p = 0.304, exact significance, two-tailed). The group that would be unaffected by the PEMI's revision (M = 2.5, Mdn = 3, n = 6) had an average rank of 14.42 while the affected group (M = 2.21, Mdn = 2, n = 17) had an average rank of 11.12. This

[42] Although 17 cases were analyzed, ties are automatically discarded by the Wilcoxon test.
[43] By definition, the revised scores of the unaffected group are identical to the original scores.

result means that the revised version of the PEMI did not create two categories of respondents in terms of agreement level. As a result, it can be inferred that the revised version mitigates the conservative bias in the measurement instrument.

4.8 Discussion and Conclusion

4.8.1 Measuring Stakeholder Participation

This article has presented substantive results on the validation of the PEMI, an instrument that purports to measure stakeholder participation. In the end, is the validity of the evidence credible and convincing? We indeed think that the evidence presented in this article lends some support as to the validity of the original version of the instrument and therefore corroborates previous evidence (see Chapitre 5). The PEMI appears to adequately cover the content of the concept and align with the respondents' opinions about the level of participation of their evaluation. A thematic analysis of the qualitative data revealed that, while some respondents believed that the PEMI well represented the "true" level of stakeholder participation (i.e., their opinion of it), others thought that the PEMI was too conservative and thus underrepresented participation. Consequently, the measurement instrument was revised. The evidence points to the fact that this revised version seems to offer an improved fit with the respondents' opinions. We therefore recommend using the revised version in future endeavours involving the measurement of stakeholder participation in evaluation.

It is important to stress that this study is not without its limitations. The most important one is certainly the fact that it is based on a small, purposive sample. For one thing, it is difficult to obtain statistically significant results with a sample size of about twenty respondents, even with nonparametric tests. For another, it is impossible to be certain that the findings derived from this study are generalizable to other respondents and settings. For instance, using the PEMI on a sample made of evaluation sponsors might reveal that the instrument does not produce scores which align with respondents' opinions. Another limitation is the lack of sensitivity of the three-point quantitative scale used to measure agreement. It would certainly have been desirable to use a more sophisticated scale. Thus, further empirical studies are certainly needed to establish the robustness of the findings presented in this article.

4.8.2 The Value Added of Mixed Methods for Validation Purposes

This article has shown how mixed methods research can play an unexpected but significant role in a rigorous study aimed at validating a measurement instrument. While the claim that mixed methods have made a positive contribution to the current study is plausible, this should be demonstrated and not be assumed.

Counterfactual thinking can be used to assess whether mixing qualitative and quantitative methods have added value to this study. Counterfactual propositions "take the generic form of 'If it had been the case that C (or not C), it would have been the case that E (or not E).' Counterfactuals make claims about events that did not actually occur." [where C = cause and E = event of interest] (Fearon, 1991, p. 169). In this context, it involves building a convincing argument as to how the study would have been different in the absence of mixed methods.

Suppose our counterfactual case (i.e., our basis of comparison) is the same validation study except for the open-ended question and thematic analysis of qualitative data. In that hypothetical case, the study would have been exclusively quantitative. In both this counterfactual case and the actual mixed methods study, the quantitative results on the convergence between respondents' scores on the PEMI and their opinion on the level of participation of their case would have been identical. However, the absence of qualitative data would have not permitted triangulation or elaboration, illustration or clarification of the quantitative results (see Greene, et al., 1989). By contrast, the actual mixed methods study used qualitative data for these purposes. Furthermore, it is possible—but not certain—that the PEMI would have been revised in the counterfactual scenario. True, the procedure to derive the overall participation score in the PEMI has been described earlier as conservative and counterintuitive (see Chapitre 5). At the same time, the quantitative results indicated that, on average, respondents partly agreed with their score. Despite the fact that the agreement was not perfect, this is hardly a strong justification for instrument revision. Further studies and applications of the instrument would have certainly been needed before a strong case for revision could be built. Collecting qualitative data and analyzing them thematically provided, by contrast, a strong justification for instrument revision. Finally, since the second phase of quantitative validation is built on instrument

revision, it is highly improbable that such a type of validation would have been conducted in the counterfactual case, at least within the same study.

To sum up, we feel that the fact of mixing qualitative data and analysis with quantitative data significantly and positively contributed to the validation and improvement of the PEMI. This study has presented an original and explicit way to combine research methods for instrument development and, as such, has tackled the need outlined by Onwuegbuzie et al. (2010). To conclude, we humbly suggest that more researchers consider using mixed methods in their validation study.

Conclusion

L'évaluation participative représente, en définitive, une tendance lourde du domaine de l'évaluation de programme. Les parties prenantes telles que les décideurs, opérateurs des programmes et bénéficiaires sont de plus en plus sollicitées pour qu'elles s'impliquent dans la réalisation des évaluations. La popularité croissante du concept de participation ne comporte toutefois pas que des avantages. D'une part, la multiplication des termes et des acceptions a fait de la participation un concept ambigu et vague dont l'utilité est limitée dans le cadre de recherches empiriques rigoureuses. D'autre part, la charge normative (positive) du concept de participation a donné lieu à des écrits abondants au sein desquels il est parfois difficile de distinguer ce qui relève de la réflexion, de la croyance et de l'expérience anecdotique, d'un côté, et de la recherche empirique rigoureuse, de l'autre.

Retour sur les questions de recherches

À travers cette thèse, nous avons tenté de combler les lacunes précédentes en proposant une analyse conceptuelle systématique de l'ÉP et en développant un instrument de mesure de ce concept dont nous avons empiriquement testé la fidélité et la validité des mesures. Nous nous tournons à présent vers un examen récapitulatif de nos conclusions qui est structuré autour des trois grandes questions de recherche présentées en introduction.

1) Qu'est-ce que la participation à l'évaluation?

Une conceptualisation tridimensionnelle de la participation a été développée, dans la foulée des travaux de Cousins et Whitmore (1998) et d'autres écrits pertinents du domaine de l'évaluation, à l'aide des outils d'analyse conceptuelle développés par Gerring (1999) et Goertz (2006). Le concept de participation possède ainsi trois dimensions constitutives : la diversité des participants, l'étendue de l'implication et le contrôle du processus évaluatif. Si une structure conceptuelle fondée sur la logique de la nécessité et de la suffisance avait été initialement proposée, une structure hybride a finalement été jugée plus en phase avec les intuitions des évaluateurs. Selon cette structure hybride, les trois dimensions constitutives (essentielles, fondamentales, etc.) doivent d'abord être « présentes » (c.-à-d., elles doivent atteindre un niveau minimal) pour considérer qu'il y a participation (relation de nécessité). Aucune dimension supplémentaire n'est toutefois requise à l'ÉP (relation de suffisance).

Une fois passé le seuil minimal, le niveau global de participation est ensuite déterminé par la *moyenne* des niveaux des trois dimensions (arrondie au besoin à l'entier inférieur). Cela signifie que la diversité, l'étendue et le contrôle ne sont pas des conditions nécessaires absolues. Un score élevé sur une dimension peut en effet « compenser » un score faible sur une autre. Le principal avantage de cette conceptualisation de l'ÉP est qu'elle représente un savant compromis entre un pouvoir discriminant élevé (c.-à-d., elle permet de distinguer efficacement les cas participatifs des cas non participatifs) et, comme nous le verrons ci-dessous, une meilleure congruence avec les intuitions des évaluateurs en matière de participation.

2) Comment traduire ce concept en un instrument de mesure opérationnel?

La conceptualisation précédente a été opérationnalisée dans un instrument baptisé *Participatory Evaluation Measurement Instrument* (PEMI) afin de pouvoir servir à la mesure de la participation. Quatre indicateurs dichotomiques ont ainsi été proposés pour la diversité des participants (présence ou absence de quatre types de parties prenantes) et pour l'étendue de l'implication (implication ou non dans quatre étapes du processus). Quant au codage du contrôle du processus évaluatif, il est fondé sur le jugement éclairé de l'utilisateur du PEMI. En outre, une dizaine de conventions de codage ont été développées afin de permettre l'assignation aux différentes catégories de l'instrument de mesure (Annexe C). Les assistants de recherche ayant utilisé l'instrument de mesure ont confirmé son caractère opérationnel et ce malgré quelques difficultés liées aux définitions des indicateurs et à la disponibilité des données dans le matériel codé.

3) Est-ce que cet instrument mesure la participation de manière fidèle et valide?

L'appréciation empirique de la fidélité et de la validité des scores générés par le PEMI a été effectuée en deux temps. Dans un premier temps, une étude exclusivement quantitative a été menée sur la version originale du PEMI (c.-à-d., celle où le score global de participation est déterminé par le score minimum des dimensions). Dans un second temps, une étude de validation combinant méthodes quantitatives et qualitatives a été réalisée sur la version

originale du PEMI nous a conduit à réviser l'instrument en adoptant une structure conceptuelle hybride.

Validation quantitative

La fidélité et la validité des scores du PEMI ont été appréciées à l'aide d'un échantillon non probabiliste de 40 cas d'évaluation rapportés dans des revues du domaine. Ces cas ont été codés en double aveugle par deux assistants de recherche. Les auteurs des articles rapportant ces cas ont ensuite été sollicités pour participer à un sondage portant sur le niveau de participation. Au total, 25 questionnaires pleinement utilisables ont été reçus.

La fidélité des scores dichotomiques et ordinaux a d'abord été appréciée à l'aide du kappa de Cohen et du coefficient de corrélation intraclasse, respectivement. Les résultats étaient acceptables (*fair*) pour le niveau global de participation, bons pour ce qui est du contrôle et de la diversité des participants et excellents pour l'étendue de l'implication. De plus, tous les résultats étaient hautement significatifs du point de vue statistique ($p = .001$), ce qui signifie que la probabilité que de tels résultats soient dus à la chance est infime.

La validité a ensuite été appréciée de manière convergente en comparant les scores quantitatifs obtenus sur le PEMI par les assistants de recherche et par les auteurs (kappa de Cohen et coefficient de corrélation intraclasse). Les résultats, qui vont dans la direction observée mais ne sont pas tous statistiquement significatifs, indiquent un accord allant de faible (*poor*) à bon. Enfin, une analyse corrélationnelle (rho de Spearman) des scores obtenus par les auteurs sur le PEMI et l'EIS, un instrument alternatif de mesure de l'implication des parties prenantes (Toal, 2009), suggère que la validité des scores de notre instrument est relativement bonne. D'une part, tel qu'attendu, une association modérée indiquant à la fois des éléments convergents et divergents a été observée entre les scores sur les deux instruments. D'autre part, nous avons constaté une corrélation modérée entre l'étendue de l'implication telle que mesurée par les deux instruments. Bien qu'une association plus forte ait été attendue, ce résultat nous laisse croire que les deux instruments mesurent des construits relativement semblables.

De manière générale, les résultats de cette étude suggèrent que le PEMI est en mesure de générer des scores qui sont à la fois fidèles et valides.

Validation mixte

Les auteurs qui ont répondu au sondage (n = 25) devaient indiquer leur niveau d'accord avec leurs résultats sur le PEMI et pouvaient, s'ils le voulaient, justifier leur réponse par des commentaires. En moyenne, les auteurs ont indiqué être partiellement en accord avec les scores du PEMI pour leur cas. Toutefois, nous avons constaté, suite à l'analyse thématique des commentaires des répondants, qu'il fallait réviser la logique d'agrégation de l'instrument de mesure pour une structure conceptuelle hybride. Lors de l'analyse des niveaux d'accord des répondants, nous avons constaté que les répondants dont le score global de participation obtenu pour leur cas augmentait avec la révision avaient une opinion plus défavorable du score obtenu que les autres répondants (test de Mann-Whitney). Un suivi a ensuite été effectué auprès de ces répondants pour examiner s'ils préféraient le score de la version révisée du PEMI à celui de la version originale. Si une majorité de répondants a affirmé préférer le score généré par la version révisée du PEMI, ces résultats n'étaient pas statistiquement significatifs (tests du signe et de Wilcoxon). Par contre, une fois la révision appliquée, le niveau d'accord des répondants affectés par celle-ci ne présentait plus de différence avec celui des autres répondants.

En définitive, la version originale du PEMI a généré des scores relativement en phase avec les intuitions des répondants quant au niveau de participation des cas. La révision apportée à l'instrument a permis d'augmenter l'alignement entre les scores de l'instrument et les intuitions des évaluateurs. Les résultats de validation convergente suggèrent que le PEMI est en mesure de produire des scores valides, c'est-à-dire qui traduisent adéquatement le niveau de participation de cas d'évaluation.

Révision du PEMI

Il est à noter que les études de validation quantitative et mixte ont été réalisées de manière séquentielle. Cela signifie que les résultats de l'étude quantitative portent sur la version *originale* du PEMI. Est-ce que ces résultats seraient différents si la version *révisée* de l'instrument était utilisée? D'une part, la logique hiérarchique de codage sous-jacente au PEMI (c.-à-d., les scores des indicateurs déterminent les scores des dimensions qui, à leur tour, déterminent le score global) fait en sorte que la révision n'aurait un impact que sur le score global de participation (PART). L'impact potentiel de la révision serait donc limité à

cet élément. D'autre part, pour les scores modifiés par la révision, l'impact a été presque toujours limité à une hausse d'une unité sur l'échelle de mesure. Si nous avons des raisons de croire que le recours à la version révisée n'aurait pas d'impact majeur sur nos résultats de validation, il n'en demeure pas moins qu'il s'agit d'abord et avant tout d'une question empirique.

Nous avons donc mené une série d'analyses rétrospectives afin de déterminer l'impact réel de la révision du PEMI sur nos résultats de validation. Il ressort d'abord de ces analyses que la révision exerce un impact positif modéré sur la fidélité intercodeur.[44] La révision entraîne toutefois une légère détérioration de l'accord entre les scores des assistants de recherche et les auteurs, d'une part, et de la convergence entre les scores des auteurs sur le PEMI et l'EIS, d'autre part.[45] Ainsi, les changements découlant de la révision de l'instrument de mesure ne semblent pas présenter de patron commun et semblent somme toute d'une magnitude limitée. Nous soutenons par conséquent que les conclusions de l'étude quantitative de validation de la version originale du PEMI sont transférables à sa version révisée.

Limites

Les conclusions d'une étude sont tributaires de la qualité du raisonnement et de la méthodologie qui les sous-tendent. Cette thèse ne faisant pas exception, il importe d'en rappeler les principales limites.

Conceptualisation et instrument de mesure

La conceptualisation de l'ÉP qui a été proposée est théorisée de manière cohérente, est solidement ancrée dans les écrits du domaine et correspond de surcroît aux intuitions de plusieurs théoriciens et praticiens du champ de l'évaluation, incluant les nôtres. Il serait toutefois surprenant que cette conceptualisation suscite l'adhésion de tous les théoriciens, chercheurs et praticiens du domaine de l'évaluation. Certains pourraient ainsi soutenir que

[44] L'ICC_{PART} passe ainsi de 0,53 (p = .000) à 0,77 (p = .000). Selon les critères de Cichetti (1994), cette amélioration représente le passage de résultats acceptables à bons.

[45] On constate une diminution de l'ICC_{PART} de 0,4 (p = 0,24) à 0,32 (p = 0,61) pour l'accord entre les assistants de recherche et les auteurs. Alors que le premier résultat se situait au début de la catégorie « acceptable », le second se situe plutôt dans la catégorie « faible ». En ce qui a trait à la convergence entre les scores des auteurs sur le PEMI et l'EIS, le résultat est presque inchangé. La corrélation (r_s) passe ainsi de 0,44 à 0,43 et demeure statistiquement significatif (p = 0,25 et 0,27).

la conceptualisation proposée néglige une dimension importante de la participation ou, au contraire, que sa couverture est trop large. Cela étant dit, les résultats de l'analyse thématique des commentaires générés lors du sondage nous laissent croire que les remises en question porteront davantage sur le calcul du score global de participation et la manière dont les dimensions sont opérationnalisées plutôt que sur des aspects conceptuels fondamentaux. Il semble en effet probable que les participants trouvent que le calcul de la participation soit trop conservateur et ce, malgré la révision qui a été apportée à l'instrument. En outre, l'analyse a fait ressortir certaines insatisfactions quant à la façon d'opérationnaliser la diversité des participants (c.-à-d., doit-on limiter la mesure aux principaux utilisateurs?) et le contrôle du processus évaluatif (p. ex., doit-on pondérer certaines étapes?). Dans l'éventualité où une alternative conceptuelle serait proposée, il faudrait en déterminer les conséquences en ce qui a trait à la délimitation de l'extension du concept et à l'alignement avec les intuitions des évaluateurs.

Résultats empiriques de validation

Les deux études empiriques de validation du PEMI présentent certaines limites méthodologiques qui peuvent affecter la validité des conclusions. Dans le présent contexte, les « menaces à la validité » (Shadish et collab., 2002) sont principalement de trois ordres : (1) biais de sélection; (2) biais de mesure et d'instrumentation; (3) validité des conclusions statistiques.

Biais de sélection

Premièrement, la sélection des cas a été effectuée « par choix raisonné ». Bien que nous ayons tenté d'assembler un échantillon de cas d'évaluation représentatifs en termes de niveau de participation et d'approches théoriques, nous ne pouvons prétendre qu'à une représentativité analytique, par opposition à statistique (Yin, 2003). Par ailleurs, il est possible que nous ayons introduit (à notre insu) certains biais lors du processus de sélection. Le critère selon lequel les cas devaient présenter un niveau suffisant d'information sur le processus évaluatif et son application ont ainsi pu favoriser certains types d'approches évaluatives (p. ex., les évaluations participatives de type pratique) et en défavoriser d'autres (p. ex., les évaluations habilitatives). De même, les auteurs des cas qui ont été invités à participer au sondage étaient, bien entendu, libres de participer ou non à

l'étude. Il est donc possible que les répondants diffèrent des non répondants par un processus d'autosélection. À titre d'exemple, les répondants pourraient être plus au fait et plus intéressés par les approches participatives que les non répondants, ou encore avoir vécu des expériences plus positives avec ces approches. Cela pourrait avoir pour effet de biaiser à la hausse les réponses des participants. Nous croyons que si biais de sélection il y a, il constitue d'abord et avant tout une menace à la validité *externe*, c'est-à-dire qu'il affecte négativement la possibilité de généraliser nos conclusions à d'autres répondants et à d'autres contextes.

Ceci étant dit, recourir à un échantillon non probabiliste semble justifié dans le cas d'une première tentative de validation empirique d'un nouvel instrument de mesure, surtout que dans le cas présent il n'existait pas de « population » bien délimitée de cas d'évaluation participatifs et non participatifs. De plus, le fait de recourir à des cas d'évaluation rapportés dans des articles peut être considéré à la fois comme un « test exigeant » et un « test aisé » pour le PEMI ce qui fait en sorte que la direction du biais n'est pas claire (sur les types de tests, voir George & Bennett, 2005). D'une part, le fait que plusieurs des cas sélectionnés contenaient beaucoup de détails sur leur processus évaluatif et que plusieurs d'entre eux aient été rédigés par des chefs de file en évaluation participative pourrait avoir contribué à l'obtention de résultats de validation plus favorables. D'autre part, le recours à des sources documentaires a occasionné des problèmes de disponibilité des données susceptibles d'avoir accru la mesure dans laquelle les assistants de recherche qui ont réalisé le codage ont dû interpréter certains aspects des cas. En ce sens, l'application du PEMI à des articles constitue un « test exigeant » par rapport à une application en situation réelle.

Biais de mesure et d'instrumentation

La deuxième catégorie de menaces à la validité des conclusions concerne la qualité de la mesure et l'instrument utilisé. Nous avons pris des mesures afin de minimiser ce type de biais, soit un prétest des questions du sondage par deux professeurs du domaine de l'évaluation, mais aucun instrument n'est à l'abri des biais, particulièrement lorsqu'il n'a jamais été utilisé auparavant, comme c'est le cas ici.

Il faut d'abord mentionner que le fait de répondre aux questions relatives au PEMI en première partie du sondage peut avoir contribué à « contaminer » les réponses de la

seconde partie sur l'EIS (voir Durand & Blais, 2006). À titre d'exemple, le fait de questionner les participants sur le niveau de contrôle des parties prenantes à l'évaluation pourrait avoir influencé leur façon d'interpréter les questions de l'EIS et donc, leurs réponses.

Ensuite, la charge normative positive du concept de participation a pu, à travers le mécanisme de la désirabilité sociale et du désir de plaire au chercheur, avoir amené certains répondants à surestimer la mesure dans laquelle leur évaluation était participative (Champagne, Brousselle, Contandriopoulos, & Hartz, 2009; Durand & Blais, 2006). C'est là un risque réel qui a été notamment discuté lors de l'analyse thématique des commentaires des répondants (chapitre 4). Indépendamment de la charge normative de la participation, le langage du questionnaire n'était pas aussi neutre qu'on aurait pu le souhaiter. En effet, les étiquettes de l'absence de participation (*évaluation technocratique*) et d'une évaluation totalement participative (*évaluation démocratique autogérée*) étaient fortement connotées. Il a donc été décidé de les retirer de la version révisée de l'instrument. Dans tous les cas, ce biais a probablement exercé un impact négatif marginal sur le niveau d'accord des scores obtenus par les assistants de recherche et les auteurs sur le PEMI.

Une dernière menace à la validité a trait à la mémoire des répondants. La publication de nombreux cas de notre échantillon datait de plusieurs années. Certains participants potentiels ont d'ailleurs refusé de répondre au sondage parce qu'ils ne se rappelaient pas suffisamment les détails du cas pour lesquels nous les avons contactés. À deux ou trois reprises, nous avons dû transférer à des auteurs une copie électronique de l'article visé. Il est donc probable que les résultats des auteurs aient été affectés par des problèmes de mémoire. Il est toutefois difficile d'estimer la direction et la magnitude de ce biais.

Validité des conclusions statistiques

Une dernière limite concerne la capacité de détecter des relations statistiquement significatives dans les données (Shadish et collab., 2002). D'une part, la faible taille des échantillons utilisés ($17 < n < 40$) constitue clairement une limite de cette étude. D'autre part, les tests non paramétriques, en particulier les méthodes de « décompte du vote » (*vote counting procedures*) combinées au « test du signe » (*sign test*), possèdent un faible pouvoir statistique (Bushman, 1994; Cooper, 1998; Littell Corcoran & Pillai, 2008;

Newbold, 1995). Ainsi, lorsque nous avons comparé les niveaux d'accord des répondants avec le score de participation obtenu avec la version originale et révisée du PEMI, nous n'avons pas été en mesure d'écarter l'hypothèse selon laquelle l'amélioration observée était due à la chance. Toutefois, nous pensons que la magnitude des résultats observés traduit plutôt le fait que la taille limitée de notre échantillon nous a empêché de déceler une relation statistiquement significative.

Contribution et pistes de recherche future

Malgré les limites soulevées à la section précédente, le PEMI est un instrument qui semble en mesure de générer des scores fidèles et valides de la participation à l'évaluation. Qui plus est, le potentiel pratique et scientifique de l'instrument semble très prometteur.

Pour la pratique

En premier lieu, le potentiel pratique de l'instrument est indéniable tant pour l'évaluateur que pour les autres parties prenantes à l'évaluation. Dans un contexte où l'ÉP gagne en popularité, le PEMI peut être utilisé comme cadre de référence pour préparer le cahier des charges d'une évaluation (c.-à-d., les spécifications attendues) que le commanditaire veut participative. Il pourrait en outre servir de « test de réalité » aux parties prenantes quant au véritable niveau de participation d'une évaluation et, dans le cas de dissonance cognitive entre croyance et réalité, offrir des pistes pour rendre l'évaluation plus participative. L'idée du test de réalité est loin d'être inutile; il a en effet été démontré qu'évaluateurs et praticiens présentent des opinions significativement différentes sur la nature participative d'une *même* évaluation (Cousins, 2001).

Pour la recherche

Nous croyons que le potentiel scientifique de l'instrument est probablement encore plus grand que son potentiel pratique. L'ÉP étant l'une des tendances les plus importantes du domaine de l'évaluation, les études théoriques et empiriques dont elle fait l'objet actuellement se multiplient à un rythme effréné (p. ex., Cousins & Chouinard, à paraître; Cullen et collab., 2011; King et collab., 2011; Rodrígues-Campos, 2012; Smits et collab., 2011; Toal, 2009; Toal & Gullickson, 2011). Or, malgré cette effervescence sur le plan de la recherche, nos connaissances sur l'ÉP demeurent étonnamment limitées. Suite à une

recension exhaustive et à une synthèse des études empiriques portant sur l'ÉP, Cousins et Chouinard (à paraître) sont en effet arrivés à la conclusion suivante :

> We argue that despite preliminary efforts to delineate what participatory evaluation is, what effects it has, and under what circumstances it works best, contemporary discourse is largely based on theoretical musings and abstract reflections on multifaceted and varied experiences within the community of evaluation practice. We contend that knowledge about participatory evaluation through systematic empirical inquiry is at an early stage of development, yet sufficiently well developed as to warrant serious consideration by evaluation practitioners and scholars alike. (Chap. 1, s.p.)

Le problème se situe notamment au niveau du type d'étude qui est publié dans les ouvrages et les revues du domaine. Une part importante des études sur l'évaluation en général et la participation en particulier prend en effet la forme de réflexions théoriques et anecdotiques, ainsi que de cas narratifs fondés sur la pratique (Cai, 1996; Cousins, 2001; Cousins & Chouinard, à paraître; Cummins, 1997; Fleischer & Christie, 2009; Peck & Gorzalski, 2009; Poth, 2008; Robinson & Cousins, 2001; Smits & Champagne, 2008; Thayer, 2006). S'il ne faut en aucun cas minimiser la contribution des études théoriques ou quasi empiriques de ce type au développement des connaissances sur l'ÉP, on constate cependant un déséquilibre au détriment des études empiriques systématiques. En outre, Cousins et Chouinard (à paraître) ont constaté que la proportion d'études quantitatives parmi les études empiriques était très faible.

Ce double déséquilibre – réflexif/empirique et qualitatif/quantitatif – signifie que nos connaissances sur la relation qu'entretient la participation avec d'autres construits tels ses déterminants et ses conséquences (voir Annexe G), reposent sur des bases qui ne sont pas aussi solides qu'elles pourraient l'être. En effet, la force des études qualitatives et des récits fondés sur la pratique réside principalement dans la production d'inférences descriptives[46], dans le développement d'hypothèses et de théories et dans l'interprétation du sens qu'attribuent les acteurs aux événements complexes qu'ils vivent (Cousins et Chouinard, à paraître; George & Bennett, 2005; King, Keohane & Verba, 1994; Yin, 2003). Quant aux

[46] L'inférence descriptive, entendue au sens de la formulation de conclusions relatives à un phénomène qui n'est pas directement observable à partir de données (observables), ne se limite pas à la généralisation statistique à partir d'un échantillon vers une population (King, Keohane & Verba, 1994). Elle peut également prendre la forme d'une généralisation vers les concepts effectuée à partir d'observations individuelles ou, encore, établir une distinction entre les composantes systématique et aléatoire d'un phénomène (Collier, Seawright & Munck, 2004, pp. 23-24).

études quantitatives, en particulier celles de type expérimental ou quasi expérimental, elles permettent de répondre avec beaucoup plus de certitude à des questions portant sur des relations causales (Hansen & Rieper, 2009; Petticrew & Roberts, 2006; Shadish, Cook & Campbell, 2002). Ces deux traditions de recherche sont donc nécessaires à la compréhension du réel, un argument qui a d'ailleurs été repris et adapté par les tenants de la recherche ayant recours aux méthodes mixtes (p. ex., Creswell & Plano Clark, 2007). Ceci dit, il faut rétablir le déséquilibre en produisant plus d'études empiriques quantitatives sur l'ÉP.

Vers de meilleures recherches sur l'hypothèse participative

Si les approches participatives représentent un sujet de plus en plus populaire en évaluation, le problème de l'utilisation de l'évaluation demeure sans aucun doute la principale préoccupation des évaluateurs depuis les années soixante. L'une des premières personnes à sonner l'alarme à cet égard a été Carol Weiss en 1967 (cité dans Weiss, 1998). Les évaluateurs de l'époque se désolaient alors du fait que leurs rapports ne se traduisaient pas directement en décisions portant sur les politiques et programmes évalués (*utilisation instrumentale*). Ce sombre tableau a été progressivement remis en question à mesure que des recherches ont démontré que les évaluations sont bel et bien utilisées, mais pas toujours de manière instrumentale et directe. Les évaluations contribuent également à modifier les croyances, connaissances et attitudes des utilisateurs sans que cela ne se traduise nécessairement par des actions (*utilisation conceptuelle*). Les évaluateurs se sont ainsi défaits d'une conception hyperrationaliste et idéaliste de la prise de décision pour se tourner vers une conception plus réaliste, ce qui a contribué à réviser à la baisse leurs attentes en ce qui a trait à l'utilisation.

L'utilisation de l'évaluation demeure encore aujourd'hui *le* sujet de débat et de recherche pour les théoriciens, chercheurs et praticiens du champ de l'évaluation. L'utilisation est en effet considérée comme un concept central en évaluation (Henry & Mark, 2003; Kirkhart, 2000; Shadish, Cook, & Leviton, 1991), voire comme un « dogme » (Vedung, 1997). Après tout, l'utilisation est ce qui justifie le travail de l'évaluateur, c'est sa raison d'être (Mark & Henry, 2004). Un rapport d'évaluation qui dort dans le tiroir d'un décideur sans jamais avoir été lu représente un gaspillage de ressources et une occasion manquée d'améliorer la

prise de décision relative au programme concerné. Considérant la force normative du concept d'utilisation, il n'est pas étonnant de constater qu'il s'agit du thème qui canalise le plus de recherches en évaluation (Brandon, 2011; Christie, 2003, 2007; Kirkhart, 2000).

Les nombreuses recherches sur l'utilisation ont notamment contribué à identifier et à raffiner les types d'utilisation (instrumentale, conceptuelle, symbolique, processuelle, etc.) ainsi qu'à identifier ses facteurs et déterminants. Sur ce dernier point, on ne compte plus les revues de littérature et tentatives de synthèse à caractère empirique ou théorique (p. ex., Cousins & Leithwood, 1986; Hofstetter & Alkin, 2003; Johnson, et collab., 2009; Leviton, 2003; Leviton & Hughes, 1981; Shulha & Cousins, 1997). Malgré le fait que la liste cumulative de facteurs affectant l'utilisation soit d'une longueur impressionnante, on en connaît relativement peu sur l'influence individuelle de facteurs comme la participation. Depuis plus de quinze ans, plusieurs chercheurs ont ainsi dénoncé le nombre limité d'études empiriques qui établissent une relation entre la participation et l'utilisation de l'évaluation ou ce qu'il conviendrait de désigner par l'expression *hypothèse participative* (Cousins, Donohue, & Bloom, 1996; Cousins, et collab., 2011; Fleischer & Christie, 2009; Thayer, 2006; Turnbull, 1999). S'il est vrai que la base empirique de nos connaissances sur l'ÉP était limitée il y a une trentaine d'années, il semble que celle-ci ait pris beaucoup d'expansion depuis, comme en témoignent les nombreuses recherches récentes sur le sujet (voir p. ex., Cullen, Coryn, & Rugh, 2011; King, et collab., 2011; Laudon, 2010; Lawrenz, King, & Ooms, 2011; Poth & Shulha, 2008; Toal & Gullickson, 2011). Certains vont même jusqu'à affirmer que « [l]'influence d'une implication accrue des parties prenantes dans tous les aspects du processus évaluatif a *dominé* la recherche sur l'utilisation de l'évaluation dans les années 90 » (Poth, 2008, p. 36 : italiques ajoutés).

Contrairement à ce qu'on pourrait croire, la multiplication des recherches empiriques sur l'hypothèse participative ne se traduit pas nécessairement par une réduction de l'incertitude à son propos. Il convient en effet de rappeler que le recours à des conceptualisations différentes de la participation – sans parler de l'utilisation – signifie que les connaissances ne s'accumulent pas mais se juxtaposent. À titre d'exemple, King, Ross, Callow-Heusser, Gullickson, Lawrenz et Weiss (2011) spécifient qu'ils ont examiné la relation entre l'implication (*involvement*) et l'utilisation, mais pas la participation. La conceptualisation et

l'instrument de mesure développés dans le cadre de cette thèse, en fournissant un cadre commun, pourraient sans aucun doute contribuer à une nouvelle génération de recherches sur l'hypothèse participative.

Un thème de recherche qui nous semble extrêmement prometteur est l'étude systématique de la relation ontologique et causale qu'entretient l'utilisation processuelle (*process use*) avec la participation. *Utilisation processuelle* est une expression inventée par Patton dans les années 90 qui réfère à un phénomène connu des évaluateurs depuis le début des années 80 (Amo & Cousins, 2007). Patton (2007) en offre la définition suivante : « Process use refers to changes in attitude, thinking, and behavior that result from participating in an evaluation. Process use includes individual learnings from evaluation involvement as well as effects on program functioning and organizational culture » (p. 99). Il a également identifié six manifestations de ce type d'utilisation (développement de la capacité évaluative, clarification des buts, etc.). Comme son nom l'indique, l'utilisation processuelle – qui serait d'ailleurs désignée de manière plus appropriée par le terme *influence* (Kirkhart, 2000) – découle du processus d'évaluation, par opposition aux conclusions qui sont contenues dans le rapport d'évaluation. Si les évaluateurs ont progressivement abaissé leurs attentes quant à l'impact que peuvent avoir les rapports d'évaluation qu'ils produisent sur la prise de décisions, ils se rendent de plus en plus compte de l'impact que peuvent avoir leurs actions durant le processus d'évaluation (Patton, 2008).

Quelle est la relation entre l'utilisation processuelle et la participation? Ce type d'utilisation découle-t-il exclusivement de pratiques participatives? D'un côté, la conceptualisation originale de Patton suggère que la participation constitue une condition nécessaire de l'utilisation processuelle : « When we take people through a process of evaluation—at least in any kind of stakeholder involvement or participatory process—they are in fact learning things about evaluation culture and often learning how to think in these ways. » (Patton, 1998, p. 226). D'un autre côté, d'autres sources (Amo & Cousins, 2007; Baptiste, 2010) laissent croire que la relation ne serait pas nécessaire : l'utilisation processuelle pourrait donc survenir dans un contexte non participatif. Cependant, même Patton (2007) semble se contredire à ce propos au sein du même article. Alors que le premier passage cité définit

l'utilisation processuelle comme conséquence de la participation, le second ouvre la porte à l'interprétation quant au caractère nécessaire de cette relation :

> Process use refers to changes in attitude, thinking, and behavior that result *from participating* in an evaluation. (Patton, 2007, p.99 : italiques ajoutés)
>
> It says things are happening to people and changes are taking place in programs and organizations as evaluation takes place, *especially when stakeholders are involved in the process* ». (Patton, 2007, p. 103 : italiques ajoutés)

La participation contribue-t-elle à l'utilisation processuelle? Cette question peut à première vue sembler circulaire puisque, si l'utilisation processuelle découle par définition de la participation, il est impossible d'y répondre par la négative.[47] Or, il est possible de traiter la conceptualisation de l'utilisation processuelle comme une hypothèse de travail et de vérifier empiriquement si les six manifestations identifiées par Patton (2007) surviennent exclusivement dans un contexte évaluatif participatif. L'instrument de mesure que nous avons développé peut s'avérer utile à ce titre. Deux principaux cas de figure sont alors logiquement possibles.

Dans le cas où ces manifestations surviendraient également dans des contextes *non* participatifs – c'est d'ailleurs ce scénario que nous privilégions – il faudrait d'abord réviser la conceptualisation de l'utilisation processuelle en effaçant toute mention à la participation. Le concept de participation serait alors remplacé par un concept moins exigeant tel que l'implication dans le processus évaluatif ou l'interaction évaluateurs-parties prenantes. Rappelons que selon la conceptualisation que nous avons proposée, il est tout à fait possible pour une partie prenante non évaluative d'être impliquée dans une évaluation à titre de client, d'observateur ou de répondant sans que cette évaluation ne soit pour autant participative. Même dans ce cas, il est possible que le processus de l'évaluation – par exemple, le fait pour l'évaluateur de poser des questions sur un système de mesure de la performance – génère des effets sur les personnes et organisations impliquées. Dans ce contexte, il serait tout à fait pertinent d'étudier empiriquement si le niveau de participation influe sur le niveau d'utilisation processuelle.

[47] Nous sommes reconnaissant à Mathieu Ouimet d'avoir attiré notre attention sur cet élément.

Dans le cas où ces manifestations identifiées par Patton (2007) seraient uniquement observées dans le cadre d'évaluations participatives, il pourrait être utile de déterminer si, d'une part, la participation constitue une condition nécessaire absolue ou relative du concept d'utilisation processuelle (p. ex., voir la structure conceptuelle hybride présentée au chapitre quatre) et, d'autre part, si d'autres conditions sont requises pour qu'on observe le phénomène d'utilisation processuelle. En effet, ce n'est pas parce qu'une condition est nécessaire à un phénomène qu'elle est pour autant suffisante (Goertz, 2006; Ragin, 2000). À titre d'exemple, la présence de nuages dans le ciel (visibles ou non) est nécessaire à la pluie mais elle n'est pas suffisante. D'autres conditions comme la pression atmosphérique, la température et le niveau d'humidité des nuages, sont requises pour qu'il y ait pluie. Pour revenir à l'utilisation processuelle, la compétence de l'évaluateur et l'intérêt des participants à l'évaluation pourraient constituer des conditions nécessaires qui, une fois combinées à la participation, seraient collectivement suffisantes au concept.

De l'importance des concepts

Au-delà de la conceptualisation et de la mesure de l'évaluation participative, et même au-delà du champ de la science politique, nous espérons avoir contribué à sensibiliser les scientifiques de toutes les disciplines à l'importance des concepts et de leur analyse. Comme nous l'avons affirmé ailleurs :

> En science politique comme dans les autres sciences sociales, des tâches comme la théorisation, la recherche empirique, la diffusion des recherches et l'enseignement seraient tout simplement impossibles en l'absence de concepts. Puisque les concepts représentent les « briques » avec lesquelles les théories et les recherches sont construites, les négliger pose des risques évidents pour la validité et l'utilité du travail scientifique […] il existe en effet un risque sérieux que l'édifice des sciences sociales ne s'écroule. (Daigneault, à paraître, 2012, s.p. : traduction libre)

Il ne faudrait toutefois pas conclure que *tous* les scientifiques doivent absolument procéder à l'analyse systématique et à l'opérationnalisation de *chacun* des concepts qu'ils utilisent. D'une part, la science étant une entreprise collective, il est tout à fait légitime d'avoir recours aux conceptualisations développées par d'autres, si celles-ci sont adéquates bien sûr. D'autre part, les scientifiques peuvent, dans certains contextes, notamment lorsqu'il s'agit d'orienter le travail de terrain et de développer des hypothèses, avoir recours à des concepts *sensibilisants* dont les contours moins bien définis :

> Sociologist Herbert Blumer (1954) is credited with originating the idea of "sensitizing concept" to orient fieldwork. Sensitizing concepts include notions like victim, stress, stigma, and learning organization that can provide some initial direction to a study as one inquires into how the concept is given meaning in a particular place or set of circumstances (Schwandt, 2001). The observer moves between the sensitizing concept and the real world of social experience, giving shape and substance to the concept and elaborating the conceptual framework with varied manifestations of the concept. Such an approach recognizes that although the specific manifestations of social phenomena vary by time, space, and circumstance, the sensitizing concept is a container for capturing, holding, and examining these manifestations to better understand patterns and implications. (Patton, 2007, p. 102)

Un concept aux contours flous peut ainsi, dans une démarche de recherche inductive et exploratoire, contribuer à l'avancement des connaissances. Un phénomène pour lequel on note une absence partielle ou totale de connaissances peut bien évidemment fort bien s'accommoder d'une telle approche de recherche. Encore faut-il être en mesure d'identifier les concepts, de connaître leur nature et d'apprécier leur valeur dans le cadre d'un projet de recherche donné. Le but de la science est la production d'inférences descriptives et causales.[48] Or, les inférences que les scientifiques génèrent à travers leurs recherches empiriques sont dépendantes de la validité de la mesure. Et nous ne pouvons pas mesurer ce qui n'a pas d'abord été conceptualisé (Sartori, [1970], 2009). Théorisation, description, explication et vérification sont des taches pour lesquelles le scientifique ne saurait se passer des concepts. À quand une formation de base en analyse conceptuelle pour les scientifiques du social? Les concepts jouent un rôle essentiel dans l'entreprise scientifique et il est grand temps pour ceux qui étudient le social de s'en rendre compte et d'agir en conséquence.

[48] King, Keohane et Verba (1994) affirment à ce sujet : « Even if explanation—connecting causes and effects—is the ultimate goal, description has a central role in all explanations, and is fundamentally important in and of itself. It is not description versus explanation that distinguishes scientific research from other research; it is whether systematic inference is conducted according to valid procedures. Inference, whether descriptive or causal, quantitative or qualitative, is the ultimate goal of all good social science » (p.34).

Bibliographie[49]

[s.a.]. (2005). Transdiscipline. In S. Mathison (Dir.), *Encyclopedia of evaluation*. Thousand Oaks : Sage.

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

Alkin, M. C. (Ed.). (2004). *Evaluation roots: Tracing theorists' views and influences*. Thousand Oaks, CA: Sage.

Alkin, M. C., & Christie, C. A. (2004). Evaluation tree revisited. In M. C. Alkin (Ed.), *Evaluation roots: Tracing theorists' views and influences* (pp. 381-392). Thousand Oaks, CA: Sage.

American Evaluation Association. (2004). American Evaluation Association guiding principles for evaluators. Page consultée à l'adresse suivante : http://www.eval.org/GPTraining/GP%20Training%20Final/gp.principles.pdf

Amo, C., & Cousins, J. B. (2007). Going through the process: An examination of the operationalization of process use in empirical research on evaluation. *New Directions for Evaluation, 2007*(116), 5-26.

Arnon, S., & Reichel, N. (2009). Closed and open-ended question tools in a telephone survey about "The good teacher". *Journal of Mixed Methods Research, 3*(2), 172-196. doi: 10.1177/1558689808331036

Arnsperger, C., & van Parijs, P. (2003). *Éthique économique et sociale*. Paris : La Découverte.

Babbie, E. R., & Benaquisto, L. (2002). *Fundamentals of social research* (1ère éd. canadienne). Scarborough, Ontario : Nelson Thomson Learning.

Bachelard, G. ([1951], 2001). *La dialectique de la durée* (3e éd.). Paris : PUF.

Baptiste, L. J. C. (2010). *Process use across evaluation approaches: An application of Q methodology in program evaluation.* Thèse de doctorat non publiée, Kent State University College, Ohio.

Baron, G., & Monnier, É. (2003). Une approche pluraliste et participative : coproduire l'évaluation avec la société civile. *Informations sociales* (110), 120-129.

Bemelmans-Videc, M.-L. (1989). Dutch experience in the utilization of evaluation research: the procedure of reconsideration. *Knowledge in Society, 2*(4), 31-48.

Bickman, L., & Reich, S. (2004). Profession of evaluation. In S. Mathison (Dir.), *Encyclopedia of evaluation*. Thousand Oaks: Sage.

[49] Il est à noter que les références qui ont servi à la constitution de la base de données pour l'étude de validation (cf., chap. 3) sont présentées séparément à l'Annexe B.

Bouveresse, J. (1998). Les sots calent. *Le Monde de l'Éducation*(255), 54-55. Page consultée à l'adresse suivante : http://peccatte.karefil.com/SBPresse/LeMondeEducJB.html

Brandon, P. R. (2011). Reflection on four multisite evaluation case studies. *New Directions for Evaluation, 2011*(129), 87-95.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. doi: 10.1191/1478088706qp063oa

Brisolara, S. (1998). The history of participatory evaluation and current debates in the field. *New Directions for Evaluation*(80), 25-41.

Brunner, I., & Guzman, A. (1989). Participatory evaluation: A tool to assess projects and empower people. *New Directions for Evaluation, 1989*(42), 9-18.

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

Bushman, B. J. (1994). Vote-counting procedures in meta-analysis. In H. Cooper & L. V. Hedges (Dir.), *The Handbook of research synthesis* (pp. 193-213). New York: Russell Sage Foundation.

Butterfoss, F. D., Francisco, V., & Capwell, E. M. (2001). Stakeholder participation in evaluation. *Health Promotion Practice, 2*(2), 114-119.

Cai, M. (1996). *An empirical examination of participatory evaluation: Teachers' perceptions of their involvement and evaluation use.* Thèse de doctorat non publiée, State University of New York at Albany, New York.

Carlsson, J., Ericksson-Baaz, M., Fallenius, A. M., & Lövgren, E. (1999). *Are evaluation useful? Cases from Swedish development co-operation*. Stockholm.

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

Champagne, F., Brousselle, A., Contandriopoulos, A.-P., & Hartz, Z. (2009). L'analyse des effets. In A. Brousselle, F. Champagne, A.-P. Contandriopoulos & Z. Hartz (Dir.), *L'évaluation : concepts et méthodes* (pp. 161-186). Québec : Les Presses de l'Université de Montréal.

Chang, H. (2009). Operationalism. In E. N. Zalta (Dir.), *The Stanford Encyclopedia of Philosophy* (Éd. de l'automne 2009). Page consultée à l'adresse suivante : http://plato.stanford.edu/archives/fall2009/entries/operationalism/

Chevrier, J. (2006). La spécification de la problématique. In B. Gauthier (Dir.), *Recherche sociale : de la problématique à la collecte des données* (4e éd., pp. 51-84). Québec : Presses de l'Université du Québec.

Christie, C. A. (2003). What guides evaluation? A study of how evaluation practice maps onto evaluation theory. *New Directions for Evaluation*(97), 7-35.

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

Collier, D., Seawright, J., & Munck, G. L. (2004). The quest for standards: King, Keohane, and Verba's *Designing Social Inquiry*. In H. E. Brady & D. Collier (Dir.), *Rethinking social inquiry: Diverse tools, shared standards* (pp. 21-50). Lanham, MD: Rowman & Littlefield.

Collins, K. M. T., Onwuegbuzie, A. J., & Sutton, I. L. (2006). Model incorporating the rationale and purpose for conducting mixed-methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal, 4*(1), 67-100.

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.

Cooper, H. M. (1998). *Synthesizing research: A guide for literature reviews* (3e éd.). Thousand Oaks, CA: Sage.

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

Cousins, J. B. (2001). Do evaluator and program practitioner perspective converge in collaborative evaluation? *Revue canadienne d'évaluation de programme, 16*(2), 113-133.

Cousins, J. B. (2003). Utilization effects of participatory evaluation. In T. Kellaghan & D. L. Stufflebeam (Dir.), *International handbook of educational evaluation* (vol. 9, pp. 245-265). Dordrecht: Kluwer Academic.

Cousins, J. B. (2005). Will the real empowerment evaluation please stand up? A critical friend perspective. In D. M. Fetterman & A. Wandersman (Dir.), *Empowerment Evaluation in Practice* (pp. 183-208). New York, NY: Guilford Press.

Cousins, J. B., & Chouinard, J. A. (à paraître). *Participatory evaluation up close: A review and integration of research-based knowledge*. Charlotte, NC: Information Age Press.

Cousins, J. B., Donohue, J. J., & Bloom, G. A. (1996). Collaborative Evaluation in North America: Evaluators' Self-reported Opinions, Practices, and Consequences. *American Journal of Evaluation, 17*(3), 207-226.

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

Cousins, J. B., & Earl, L. M. (1995). Participatory evaluation in education: What do we know? Where do we go? In J. B. Cousins & L. M. Earl (Dir.), *Participatory evaluation in education: Studies in evaluation use and organizational learning* (pp. 159-180). London: Falmer Press.

Cousins, J. B., & Earl, L. M. (1999). When the boat gets missed: Response to M.F. Smith. *American Journal of Evaluation, 20*(2), 309-317. doi: 10.1177/109821409902000212

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., & Leithwood, K. A. (1986). Current empirical research on evaluation utilization. *Review of Educational Research, 56*(3), 331-364.

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

Creswell, J. W., & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage.

Cullen, A. (2009). *The politics and consequences of stakeholder participation in international development evaluation.* Thèse de doctorat non publiée, Western Michigan University, Michigan (É.-U.). Page consultée à l'adresse suivante : 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.

Cummings, R. (1997). *The influence of stakeholder involvement in evaluation studies on the use of evaluation information: A longitudinal study.* Thèse de doctorat non publiée, Murdoch University, Perth.

Dahl, R. A. (2005). What Political Institutions Does Large-Scale Democracy Require? *Political Science Quarterly, 120*(2), 187-197.

Daigneault, P.-M. (2010). L'examen de la qualité des évaluations fédérales : une méta-évaluation réussie? *Revue canadienne d'évaluation de programme, 23*(2, automne 2008), 191-224.

Daigneault, P.-M. (2011). Les approches théoriques en évaluation : état de la question et perspectives. *Les approches théoriques en évaluation, Cahiers de la performance et de l'évaluation, Printemps 2011*(4), 2-6.

Daigneault, P.-M. (à paraître, 2012). Introduction to the Symposium 'Conceptual Analysis in Political Science and Beyond'. *Social Science Information – Information sur les sciences sociales, 51*(2).

Daigneault, P.-M., & Jacob, S. (2007, novembre). *Rethinking participatory evaluation's conceptualization: Toward the development of a full-blown, useful, concept*. Communication présentée au 21e colloque annuel de l'American Evaluation Association, "Evaluation 2007: Evaluation and Learning", Baltimore, MD.

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. (à paraître, 2012). Les concepts souffrent-ils de négligence bénigne en sciences sociales? Éléments d'analyse conceptuelle et examen exploratoire de la littérature francophone à caractère méthodologique. *Social Science Information – Information sur les sciences sociales, 51*(2).

Daniels, N. (2003). Reflective equilibrium. *The Stanford encyclopedia of philosophy* (Éd. : été 2003). Page consultée à l'adresse suivante : http://plato.stanford.edu/entries/reflective-equilibrium/

DeLeon, P. (1997). *Democracy and the policy sciences*. Albany: State University of New York Press.

Dellinger, A. B., & Leech, N. L. (2007). Toward a Unified Validation Framework in Mixed Methods Research. *Journal of Mixed Methods Research, 1*(4), 309-332. doi: 10.1177/1558689807306147

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

Díaz-Puente, J. M., Montero, A. C., & de los Ríos Carmenado, I. (2009). Empowering communities through evaluation: Some lessons from rural Spain. *Community Development Journal, 44*(1), 53-67. doi: 10.1093/cdj/bsm008

Dubois, N., & Marceau, R. (2005). Un état des lieux théoriques de l'évaluation: une discipline à la remorque d'une révolution scientifique qui n'en finit pas. *Revue canadienne d'évaluation de programme, 20*(1), 1-36.

Durand, C., & Blais, A. (2006). La mesure. In B. Gauthier (Dir.), *Recherche sociale : de la problématique à la collecte des données* (4e éd., pp. 185-209). Québec : Presses de l'Université du Québec.

Durham, J., Tan, B.-K., & White, R. (2011). Utilizing mixed research methods to develop a quantitative assessment tool. *Journal of Mixed Methods Research, 5*(3), 212-226. doi: 10.1177/1558689811402505

Duval, J. (2004). Les concepts comme instruments et comme objets. Éléments sur l'usage et l'analyse de concepts en sociologie. In P. Robert-Demontrond (Dir.), *L'analyse de concepts* (pp. 131-159). Rennes : Apogée - IREIMAR.

Fearon, J. (1991). Counterfactuals and hypothesis testing in political science. *World Politics, 43*(2), 169-195.

Fenton, A. (2006). Weft QDA (Version 1.0.1.). Page consultée à l'adresse suivante : http://www.pressure.to/qda/.

Fetterman, D. M. (2000). *Foundations of empowerment evaluation*. Thousand Oaks, CA: Sage.

Fetterman, D. M. (2004). Fieldwork. In S. Mathison (Dir.), *Encyclopedia of evaluation*. Page consultée à l'adresse suivante : http://www.sage-ereference.com/view/evaluation/n214.xml?rskey=cKeNAh&result=3&q=naturalistic

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*, (3e éd., pp. 598-626). Hoboken, NJ: Wiley.

Forss, K., Rebien, C. C., & Carlsson, J. (2002). Process use of evaluations: types of use that precede lessons learned and feedback. *Evaluation, 8*(1), 29-45.

Fournier, D. M. (2004). Evaluation. In S. Mathison (Dir.), *Encyclopedia of evaluation*. Thousand Oaks: Sage.

Garaway, G. B. (1995). Participatory evaluation. *Studies in Educational Evaluation, 21*(1), 85-102.

Garon, S., & Roy, B. (2001). L'évaluation des organismes communautaires. L'exemple d'un partenariat avec l'État: entre l'espoir et la désillusion. *Nouvelles pratiques sociales, 14*(1), 97-110.

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Gerring, J. (1999). What makes a concept good? A criterial framework for understanding concept formation in the social sciences. *Polity, 31*(3), 357-393.

Geva-May, I., & Pal, L. A. (1999). Good fences make good neighbours. *Evaluation, 5*(3), 259-277. doi: 10.1177/13563899922208986

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

Goertz, G. (2009). Point of departure: Intension and extension. In D. Collier & J. Gerring (Dir.), *Concepts and method in social science: The tradition of Giovanni Sartori* (pp. 181-202). London: Routledge.

Goyette, M. (2009). Le développement de l'évaluation de programme. In M. Alain & D. Dessureault (Dir.), *Élaborer et évaluer les programmes d'intervention psychosociale* (pp. 29-42). Québec : Presses de l'Université du Québec.

Greene, J. C. (1987). Stakeholder participation in evaluation design: Is it worth the effort? *Evaluation and Program Planning, 10*(4), 379-394.

Greene, J. C. (1988). Communication of results and utilization in participatory program evaluation. *Evaluation and Program Planning, 11*(4), 341-351.

Greene, J. C. (2005). Stakeholders. In S. Mathison (Dir.), *Encyclopedia of evaluation* (pp. 397-398). Thousand Oaks, CA: Sage.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274. doi: 10.3102/01623737011003255

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

Guba, E. G., & Lincoln, Y. S. (1989). *Fourth generation evaluation*. Newbury Park, CA: Sage.

Hansen, H. F., & Rieper, O. (2009). The evidence movement: The development and consequences of methodologies in review practices. *Evaluation, 15*(2), 141-163. doi: 10.1177/1356389008101968

Hayward, C., Simpson, L., & Wood, L. (2004). Still left out in the cold: Problematising participatory research and development. *Sociologia Ruralis, 44*(1), 95-108. doi: 10.1111/j.1467-9523.2004.00264.x

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314.

Hofstetter, C. H., & Alkin, M. C. (2003). Evaluation use revisited. In T. Kellaghan, D. L. Stufflebeam & L. A. Wingate (Dir.), *International handbook of educational evaluation* (pp. 197-222). Dordrecht; Boston; London: Kluwer Academic.

House, E. R., & Howe, K. R. (2000). Deliberative democratic evaluation. *New Directions for Evaluation, 2000*(85), 3-12.

Howlett, M. P., & Ramesh, M. (2003). *Studying public policy: Policy cycles and policy subsystems* (2e éd.). Don Mills, Ontario: Oxford University Press.

Huberman, M. (1995). The many modes of participatory evaluation. In J. B. Cousins & L. M. Earl (Dir.), *Participatory evaluation in education: Studies in evaluation use and organizational learning* (pp. 103-111). London: Falmer Press.

Jackson, E. T., & Kassam, Y. (1998). Introduction. In E. T. Jackson & Y. Kassam (Dir.), *Knowledge shared: Participatory evaluation in development cooperation* (pp. 1-20). Ottawa / West Hartford, CT: International Development Research Centre / Kumarian Press.

Jacob, S. (2004). Évaluation. In L. Boussaguet, S. Jacquot & P. Ravinet (Dir.), *Dictionnaire des politiques publiques* (pp. 201-208). Paris : Les Presses de Science Po.

Jacob, S. (2005). Réflexions autour d'une typologie des dispositifs institutionnels d'évaluation. *Revue canadienne d'évaluation de programme, 20*(2), 49-68.

Jacob, S. (2010). Évaluation. In L. Boussaguet, S. Jacquot & P. Ravinet (Dir.), *Dictionnaire des politiques publiques*, (2e éd., pp. 257-265). Paris : Les Presses de Science Po.

Jacob, S., & Boisvert, Y. (2010). To Be or Not to Be a Profession: Pros, Cons and Challenges for Evaluation. *Evaluation, 16*(4), 349-369. doi: 10.1177/1356389010380001.

Jacob, S., & Daigneault, P.-M. (2011). La gouvernance et l'implication des parties prenantes dans l'évaluation des politiques : panacée ou boîte de pandore? In C. Rouillard & N. Burlone (Dir.), *L'État et la société civile sous le joug de la gouvernance* (pp. 217-242). Québec : Presses de l'Université Laval.

Jacob, S., & Ouvrard, L. (2009). L'évaluation participative : avantages et difficultés d'une pratique innovante. *Cahiers de la performance et de l'évaluation, automne 2009*(1), 82.

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

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*(2), 112-133. doi: 10.1177/1558689806298224

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

Joint Committee on Standards for Educational Evaluation. (1994). *The Program Evaluation Standards: How to assess evaluations of educational programs* (2e éd.). Thousand Oaks, CA: Sage.

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

King, J. A. (2005). Participatory evaluation. In S. Mathison (Dir.), *Encyclopedia of evaluation* (pp. 291-294). Thousand Oaks, CA: Sage.

King, J. A. (2007). Making sense of participatory evaluation. *New Directions for Evaluation*(114), 83-86.

King, J. A., Ross, P. A., Callow-Heusser, C., Gullickson, A. R., Lawrenz, F., & Weiss, I. R. (2011). Reflecting on multisite evaluation practice. *New Directions for Evaluation, 2011*(129), 59-71. doi: 10.1002/ev.355

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Kirkhart, K. E. (2000). Reconceptualizing evaluation use: An integrated theory of influence. *New Directions for Evaluation* (88), 5-23.

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.* Mémoire de maîtrise non publié, York University, Toronto.

Lawrenz, F., King, J. A., & Ooms, A. (2011). The role of involvement and use in multisite evaluations. *New Directions for Evaluation, 2011*(129), 49-57. doi: 10.1002/ev.354

Leeuw, F. L. (1992). Performance auditing and policy evaluation: Discussing similarities and dissimilarities. *Revue canadienne d'évaluation de programme, 7*(1), 53-68.

Leeuw, F. L. (2009). Evaluation: A booming business but is it adding value? *Evaluation Journal of Australasia, 9*(1), 3-9.

Lemieux, V. (2002). *L'étude des politiques publiques : Les acteurs et leur pouvoir* (2e éd. revue et augmentée). Québec : Presses de l'Université Laval.

Lemieux, V. (2006). Évaluation de programmes et analyse de politiques. *Télescope : Revue d'analyse comparée en administration publique, 13*(1), 1-8.

Lennie, J. (2005). An evaluation capacity-building process for sustainable community IT initiatives: Empowering and disempowering impacts. *Evaluation, 11*(4), 390-414. doi: 10.1177/1356389005059382

Leviton, L. C., & Hughes, E. F. X. (1981). Research on the utilization of evaluations: A review and synthesis. *Evaluation Review, 5*(4), 525-548.

Leviton, L. C. (2003). Evaluation use: Advances, challenges and applications. *American Journal of Evaluation, 24*(4), 525-535.

Lhérété, É. (2011). La solitude du thésard de fond. *Sciences humaines, 10*(230), 10.

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford: Oxford University Press.

Lincoln, Y. S. (2004). Fourth-generation evaluation. In S. Mathison (Dir.), *Encylopedia of evaluation*. Thousand Oaks: Sage.

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.

Luyt, R. (2011). A framework for mixing methods in quantitative measurement development, validation, and revision: A case study. *Journal of Mixed Methods Research*. doi: 10.1177/1558689811427912

Mark, M. M. (2001). Evaluation's future: Furor, futile, or fertile? *American Journal of Evaluation, 22*(3), 457-479.

Mark, M. M., & Henry, G. T. (2004). The mechanisms and outcomes of evaluation influence. *Evaluation, 10*(1), 35-57.

Mathie, A., & Greene, J. C. (1997). Stakeholder participation in evaluation: How important is diversity? *Evaluation and Program Planning, 20*(3), 279-285.

Mathison, S. (2005a). Preface. In S. Mathison (Dir.), *Encyclopedia of evaluation* (pp. 33-35). Thousand Oaks, CA: Sage.

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

Mathison, S. (2008). What is the difference between evaluation and research—and why do we care? In N. L. Smith & P. R. Brandon (Dir.), *Fundamental issues in evaluation* (pp. 183-196). New-York, NY: Guilford.

Mayne, J. (2006). Audit and evaluation in public management: Challenges, reforms, and different roles. *Revue canadienne d'évaluation de programme, 21*(1), 11-45.

McDonald, M. P. (2005). Validity, Data Sources. In K.-L. Kimberly (Dir.), *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

Morgan, L. L. (1996). *Utilization-focused evaluation: A case study of a diversity program evaluation.* Thèse de doctorat non publiée: University of California, Los Angeles, California. Page consultée à l'adresse suivante : http://proquest.umi.com/pqdweb?did=741878091&Fmt=7&clientId=9268 &ampRQT=309&VName=PQD Dissertation Abstracts database

Murray, R. (2002). Citizens' control of evaluations: Formulating and assessing alternatives. *Evaluation, 8*(1), 81-100. doi: 10.1177/1358902002008001488

Newbold, P. (1995). *Statistics for business and economics* (4e éd.). Upper Saddle River, NJ: Prentice Hall.

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

O'Sullivan, R. G., & D'Agostino, A. (2002). Promoting evaluation through collaboration: Findings from community-based programs for young children and their families. *Evaluation, 8*(3), 372-387. doi: 10.1177/135638902401462466

Onwuegbuzie, A. J., Bustamante, R. M., & Nelson, J. A. (2010). Mixed research as a tool for developing quantitative instruments. *Journal of Mixed Methods Research, 4*(1), 56-78. doi: 10.1177/1558689809355805

Organisation de Coopération et de Développement Économiques. (1999). *Vers de meilleures pratiques de l'évaluation : guide de meilleures pratiques à suivre pour l'évaluation et guide de référence*. Paris : OECD.

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

Pal, L. (2001). *Beyond policy analysis: Public issue management in turbulent times* (2e éd.). Scarborough, On: Nelson Thomson Learning.

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.

Patton, M. Q. (1998). Discovering Process Use. *Evaluation, 4*(2), 225-233.

Patton, M. Q. (2007). Process use as a usefulism. *New Directions for Evaluation, 2007*(116), 99-112.

Patton, M. Q. (2008). *Utilization-focused evaluation* (4e éd.). Los Angeles: Sage.

Peck, L. R., & Gorzalski, L. M. (2009). An evaluation use framework and empirical assessment. *Journal of MultiDisciplinary Evaluation, 6*(12), 139-156.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA: Blackwell.

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

Poth, C.-A. N. (2008). *Promoting evaluation use within dynamic organizations: A case study examining evaluator behaviour*. Thèse de doctorat non publiée, (69), US: ProQuest Information & Learning. PsycInfo.

Power, M. ([1997], 2005). *La société de l'audit : l'obsession du contrôle*. Paris : La Découverte.

Preskill, H., & Caracelli, V. (1997). Current and developing conceptions of use: Evaluation use TIG survey results. *American Journal of Evaluation, 18*(1), 209-225. doi: 10.1177/109821409701800122

Preskill, H., Zuckerman, B., & Matthews, B. (2003). An exploratory study of process use: Findings and implications for future research. *American Journal of Evaluation, 24*(4), 423-442.

Québec (Province). Secrétariat du Conseil du trésor. Sous-secrétariat aux politiques budgétaires et aux programmes. (2002). *L'évaluation de programme : document destiné aux dirigeants et dirigeantes de ministères et d'organismes*. Québec : Secrétariat du Conseil du trésor.

Ragin, C. C. (2000). *Fuzzy-set social science*. Chicago: University of Chicago Press.

Rebien, C. C. (1996). Participatory evaluation of development assistance: Dealing with power and facilitative learning. *Evaluation, 2*(2), 151-171. doi: 10.1177/135638909600200203

Ridde, V. (2006). Suggestions d'améliorations d'un cadre conceptuel de l'évaluation participative. *Revue canadienne d'évaluation de programme, 21*(2), 1-23.

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

Robert-Demontrond, P. (2004). Premiers repérages. In P. Robert-Demontrond (Dir.), *L'analyse de concepts* (pp. 15-39). Rennes : Apogée – IREIMAR.

Robinson, T. T., & Cousins, J. B. (2004). Internal participatory evaluation as an organizational learning system: A longitudinal case study. *Studies in Educational Evaluation, 30*(1), 1-22.

Rodrígues-Campos, L. (2012). Stakeholder involvement in evaluation: Three decades of the American Journal of Evaluation. *Journal of MultiDisciplinary Evaluation, 8*(17), 57-79.

Rossi, P. H. (2004). My views of evaluation and their origins. In M. C. Alkin (Dir.), *Evaluation roots: Tracing theorists' views and influences* (pp. 122-131). Thousand Oaks: Sage.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A systematic approach* (7e éd.). Thousand Oaks, CA: Sage.

Rouillard, C., & Burlone, N. (Dir.). (2011). *L'État et la société civile sous le joug de la gouvernance*. Québec : Presses de l'Université Laval.

Rubin, H. J., & Rubin, I. S. (2011). *Qualitative interviewing: The art of hearing data* (3e éd.). Thousand Oaks, CA: Sage.

Sandelowski, M., & Barroso, J. (2006). *Handbook for synthesizing qualitative research*. New York, NY: Springer.

Sartori, G. ([1970] 2009). Concept misformation in comparative politics. In D. Collier & J. Gerring (Dir.), *Concepts and method in social science: The tradition of Giovanni Sartori* (pp. 13-43). London: Routledge.

Sartori, G. ([1975] 2009). The Tower of Babel. In D. Collier & J. Gerring (Dir.), *Concepts and method in social science: The tradition of Giovanni Sartori* (pp. 61-96). London: Routledge.

Sartori, G. ([1984] 2009). Guidelines for concept analysis. In D. Collier & J. Gerring (Dir.), *Concepts and method in social science: The tradition of Giovanni Sartori* (pp. 97-150). London: Routledge.

Schedler, A. (2011). Concept formation. In B. Badie, D. Berg-Schlosser & L. A. Morlino (Dir.), *International encyclopedia of political science* (pp. 370-382). Thousand Oaks: Sage.

Schwandt, T. A. (2004). Auditing. *Encyclopedia of evaluation*. Page consultée à l'adresse suivante : http://www.sage-ereference.com/view/evaluation/n39.xml

Schwartz, R., & Mayne, J. (2005a). Assuring the quality of evaluative information: Theory and practice. *Evaluation and Program Planning, 28*(1), 1-14.

Schwartz, R., & Mayne, J. (Dir.). (2005b). *Quality matters: Seeking confidence in evaluating, auditing, and performance reporting*. New Brunswick, NJ: Transaction.

Scriven, M. (1997). Empowerment Evaluation Examined. *American Journal of Evaluation, 18*(1), 165-175.

Scriven, M. (2005). Book Review: Empowerment Evaluation Principles in Practice. *American Journal of Evaluation, 26*(3), 415-417.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation: Theories of practice*. Newbury Park, CA: Sage.

Shadish, W. R., & Luellen, J. K. (2004). History of evaluation. *Encyclopedia of evaluation*. Page consultée à l'adresse suivante : http://www.sage-ereference.com/view/evaluation/n251.xml

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.

Sheppard, E. (2004). Problème public. In L. Boussaguet, S. Jacquot & P. Ravinet (Dir.), *Dictionnaire des politiques publiques* (pp. 347-354). Paris : Presses science po.

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

Shulha, L. M., & Cousins, J. B. (1997). Evaluation use: theory, research, and practice since 1986. *American Journal of Evaluation, 18*(3), 195-208.

Singleton, R., Straits, B. C., & Straits, M. M. (1993). *Approaches to social research* (2e éd.). New York, NY: Oxford University Press.

Small, M. L. (2011). How to conduct a mixed methods study: Recent trends in a rapidly growing literature. *Annual Review of Sociology, 37*(1), 57-86. doi: doi:10.1146/annurev.soc.012809.102657

Smith, H. J., Kuseck, J. Z., & Rist, R. C. (2004). Performance-based monitoring. *Encyclopedia of evaluation*. Page consultée à l'adresse suivante : http://www.sage-ereference.com/view/evaluation/n407.xml

Smith, N. L. (1993). Improving evaluation theory through the empirical study of evaluation practice. *American Journal of Evaluation, 14*(3), 237-242. doi: 10.1177/109821409301400302

Smits, P. A., & Champagne, F. (2008). An assessment of the theoretical underpinnings of practical participatory evaluation. *American Journal of Evaluation, 29*(4), 427-442.

Smits, P. A., Champagne, F., & Brodeur, J.-M. (2011). A mixed method study of propensity for participatory evaluation. *Evaluation and Program Planning, 34*(3), 217-227.

Springett, J., & Wallerstein, N. (2008). Issues in participatory evaluation. In M. Minkler & N. Wallerstein (Dir.), *Community-based participatory research for health: From process to outcomes* (2e éd., pp. 199-220). San Francisco, CA: Jossey-Bass.

Stake, R. E., & Abma, T. A. (2005). Responsive evaluation. In S. Mathison (Dir.), *Encyclopedia of evaluation* (pp. 376-379). Thousand Oaks, CA: Sage.

Thayer, C. E. (2006). *Participatory evaluation in nonprofit organizations: Rhetoric or reality?* Thèse de doctorat non publiée, The George Washington University, District of Columbia.

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. Page consultée à l'adresse suivante : http://www.aapor.org/AM/Template.cfm?Section=Standard_Definitions2&Template=/CM/ContentDisplay.cfm&ContentID=3156

Themessl-Huber, M. T., & Grutsch, M. A. (2003). The shifting locus of control in participatory evaluations. *Evaluation, 9*(1), 92-111. doi: 10.1177/1356389003009001006

Thoenig, J.-C. (2004). Politique publique. In L. Boussaguet, S. Jacquot & P. Ravinet (Dir.), *Dictionnaire des politiques publiques* (pp. 326-333). Paris : Presses de science po.

Toal, S. A. (2007). *The development and validation of an evaluation involvement scale for use in multi-site evaluations*. Thèse de doctorat non publiée, 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

Toal, S. A., & Gullickson, A. R. (2011). The upside of an annual survey in light of involvement and use: Evaluating the Advanced Technological Education program. *New Directions for Evaluation, 2011*(129), 9-15. doi: 10.1002/ev.349

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.

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

Turnbull, B. (1999). The mediating effect of participation efficacy on evaluation use. *Evaluation and Program Planning, 22*(2), 131-140.

VanderPlaat, M., Samson, Y., & Raven, P. (2001). The politics and practice of empowerment evaluation and social interventions: Lessons from the Atlantic Community Action Program for Children regional evaluation. *Revue canadienne d'évaluation de programme, 16*(1), 79-98.

Varone, F., & Jacob, S. (2004). Institutionnalisation de l'évaluation et Nouvelle Gestion Publique : Un état des lieux comparatif. *Revue internationale de politique comparée, 11*(2), 271-292.

Vedung, E. (1997). *Public policy and program evaluation*. New Brunswick, NJ: Transaction.

Weaver, L., & Cousins, J. B. (2004). Unpacking the participatory process. *Journal of MultiDisciplinary Evaluation (JMDE)*(1), 19-40.

Weiss, C. H. (1983). The stakeholder approach to evaluation: Origins and promise. *New Directions for Evaluation* (17), 3-14.

Weiss, C. H. (1986). The stakeholder approach to evaluation: Origins and promise. In E. R. House (Dir.), *New Directions in Educational Evaluation* (pp. 145-157). London: Falmer Press.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation? *American Journal of Evaluation, 19*(1), 21-33.

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

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

Yin, R. K. (2003). *Case study research: Design and methods* (3e éd.). Thousand Oaks, CA: Sage.

Annexe A : Approbation éthique

UNIVERSITÉ LAVAL

Vice-rectorat à la recherche et à la création
Comité d'éthique de la recherche

APPROBATION DE L'ÉTHIQUE

Projet de recherche impliquant des êtres humains ou
la consultation de renseignements personnels

Ce projet de recherche a été examiné en conformité avec les
Modalités de gestion de l'éthique de la recherche sur des êtres humains de l'Université Laval,
par le Comité plurifacultaire d'éthique de la recherche

Projet intitulé :	*Validation d'un instrument de mesure de la participation à l'évaluation*
Nom du chercheur :	Monsieur Pierre-Marc Daigneault
Nom du directeur de recherche :	Monsieur Steve Jacob
Numéro d'approbation :	2011-246 / 28-10-2011
Date de décision :	28 octobre 2011
Date d'expiration de l'approbation :	1er novembre 2012

Après examen des informations et des documents qui lui ont été transmis, le Comité a constaté que ce projet respecte les principes d'éthique de la recherche avec des êtres humains. Il prend acte de la confirmation écrite du chercheur à l'effet qu'il a pris connaissance des mesures de suivi[1] associées à l'émission de l'approbation éthique de son projet et qu'il accepte de les appliquer. Par conséquent, le Comité approuve ce projet pour un an.

Jocelyn Lindsay

Jocelyn Lindsay, président
Comité plurifacultaire d'éthique de la recherche

Date

[1] Rappel des mesures de suivi au verso

Maison Michael-John-Brophy
2241, chemin Sainte-Foy
Québec (Québec) G1V 0A6
CANADA

418 656-2131, poste 4506
Télécopieur : 418 656-2840
cer@vrr.ulaval.ca
www.cerul.ulaval.ca

Annexe B : Base de données constituée pour l'étude de validation

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 cas)**.

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. *Revue canadienne d'évaluation de programme, 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. *Revue canadienne d'évaluation de programme, 4*(1), 1-10.

Cousins, J. B. (1996). Consequences of researcher Involvement in participatory evaluation. *Studies in Educational Evaluation, 22*(1), 3-27 **(3 cas)**.

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 cas)**.

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 cas)**.

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. *Revue canadienne d'évaluation de programme*, *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, printemps), 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 cas)**.

Ridde, V. (2003). The experience of a pluralist approach in a country at war: Afghanistan. *Revue canadienne d'évaluation de programme*, *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 collab. (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. *Revue canadienne d'évaluation de programme*, *10*(1), 159-162.

Somers, C. (2005). Evaluation 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 Evaluation, 1989*(41), 35-48.

Annexe C : Conventions de codage (version finale)

Instructions générales

G – 1. Le codage de toutes les dimensions est fondé sur le concept d'implication *significative*. Cela signifie que les parties prenantes doivent jouer un rôle significatif dans la conception et/ou la réalisation de l'évaluation (incluant la diffusion des conclusions). Être une source de données ou un observateur passif n'est donc pas suffisant pour qu'on considère qu'un type de partie prenante est impliqué de manière significative (ce qui signifie un score nul pour l'ensemble des dimensions).

G – 2. Le score pour le niveau global de participation est toujours le *minimum* (c.-à-d, le score le plus bas) des trois dimensions. Par exemple, des scores de 0,25, 0,50 et de 0,75 pour la diversité des participants, l'étendue de l'implication et le contrôle du processus évaluatif pour un cas donné signifient que le score global est de 0,25.

Étendue de l'implication (EoI)[50]

EoI – 1. Lorsque certaines étapes du processus évaluatif ne sont pas mentionnées (p. ex., la diffusion des conclusions), les codeurs doivent inférer le score d'étendue à partir des informations disponibles pour les autres étapes.

Diversité des participants (DoP)

DoP – 1. Les premier et second indicateurs ont été respectivement rebaptisés « Décideurs, concepteurs de politiques et gestionnaires » et « Personnes directement responsables de la prestation des programmes ». Le personnel cadre de tous les niveaux hiérarchiques est pris en compte par le premier indicateur tandis que le second indicateur s'applique exclusivement aux professionnels et fonctionnaires de première ligne responsables de la prestation des services.

DoP – 2. Le quatrième indicateur, « société civile et citoyens », s'applique aux individus et groupes qui défendent des intérêts qui sont d'une portée plus large que le programme et/ou l'organisation faisant l'objet de l'évaluation. Supposons qu'un programme éducatif pour élèves doués est mis en œuvre dans une commission scolaire. Les enseignants et leurs délégués syndicaux locaux sont pris en compte par le second indicateur alors que des représentants de la Fédération des syndicats de l'enseignement sont pris en compte par le quatrième.

DoP – 3. Les évaluateurs professionnels et leur personnel de soutien (adjoint administratif, auxiliaire de recherche, etc.), peu importe leur affiliation organisationnelle, ne sont *jamais* pris en compte dans le codage de la diversité. Ainsi, les universitaires qui sont engagés à titre d'évaluateurs externes ne seraient pas pris en compte dans le codage de la diversité. En

[50] Pour plus de cohérence avec l'article du chapitre 3, les acronymes anglais sont utilisés pour les dimensions.

revanche, les universitaires agissant à titre d'experts sur un comité consultatif d'évaluation le seraient.

Contrôle du processus évaluatif (CoEP)

CoEP – 1. Le codage de cette dimension ne relève d'aucun indicateur mais relève plutôt d'une appréciation subjective de la distribution relative du contrôle entre l'évaluateur, d'une part, et les parties prenantes non évaluatives, d'autre part. Le codage de cette dimension est ***exclusivement fondé sur les étapes auxquelles les parties prenantes non évaluatives sont impliquées***. Il est donc théoriquement possible d'obtenir un score élevé sur cette dimension même si l'implication se limite à une seule étape.

CoEP – 2. Lorsque le contrôle varie durant le processus évaluatif, le score pour cette dimension doit être *représentatif* de la part de contrôle que les parties prenantes non évaluatives possèdent durant les étapes où elles sont impliquées. Pour ce faire, le score moyen de contrôle par étape peut être calculé. Contrairement aux échelles de Likert, cependant, le score doit être arrondi au besoin à l'une des bornes de l'échelle telles que 0,25 ou 0,50 (et non 0,378 ou 0,516).

CoEP – 3. Les scores de contrôle doivent être fondés sur ce que les auteurs affirment à propos de sa distribution et comment ils le qualifient. Par exemple, des termes et des expressions tels que « partagé », « également », « de manière collaborative », « conjointement », « ont travaillé ensemble » et « décisions mutuelles » peuvent être des indicateurs utiles d'un contrôle partagé entre évaluateur et participants (0,50). De la même manière, « consultées », « inspiré de », « donner une voix à », « ont facilité » sont généralement des indicateurs d'une distribution inégale mais non exclusive du contrôle (0,25 ou 0,75). Enfin, « garder le contrôle », « totalement ou complètement géré par », « pleinement responsable » et « décidé unilatéralement par » sont des expressions indiquant un contrôle exclusif (score de 0 ou de 1). Même si ces termes et expressions sont utiles pour le codage du contrôle, elles ne doivent pas être appliquées mécaniquement; le codage est une question de jugement informé.

CoEP – 4. Quand on n'a aucune indication quant à la distribution du contrôle, on doit présumer que le contrôle est partagé également entre l'évaluateur et les autres parties prenantes (0,50). De plus, seule la distribution effective du contrôle est pertinente, pas le nombre d'acteurs dans chaque catégorie. En effet, le contrôle peut être totalement dans les mains de l'évaluateur et ce, même si l'équipe d'évaluation compte seulement un évaluateur pour dix participants (et vice-versa).

CoEP – 5. Un score de 1 sur cette dimension indique que les parties prenantes non évaluatives ont un plein contrôle sur le processus évaluatif. Cela signifie qu'il n'y a pas d'évaluateurs professionnels impliqués dans l'évaluation (les parties prenantes agissent à titre d'évaluateurs) ou que le rôle des évaluateurs professionnels est confiné à celui de simple exécutant des décisions des parties prenantes.

Annexe D : Annonce de recrutement

Cette annexe présente une version générique, en français et en anglais, du courriel d'invitation envoyé aux répondants potentiels.

Version française

[**Salutations personnalisées**],

Je sollicite votre participation dans le cadre d'une étude intitulée « Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation ». Cette recherche, qui est réalisée dans le cadre de ma thèse de doctorat, vise à valider un instrument de mesure de l'évaluation participative.

Vous avez été sélectionné(e) sur la base de votre participation à l'évaluation suivante : [**insérer référence complète**].

La participation à cette étude implique de répondre à un questionnaire en format électronique d'une durée totale de 15 à 20 minutes portant sur le niveau de participation à l'évaluation de ce cas spécifique. Vous êtes libre d'accepter ou non de participer à cette étude et, pour éclairer votre décision, vous êtes invité(e) à consulter le texte qui introduit le questionnaire disponible au lien suivant :

[**insérer lien sécurisé vers le questionnaire**]

Ce projet a été approuvé par le Comité d'éthique de la recherche de l'Université Laval (no d'approbation 2011-246/28-10-2011). Dans l'éventualité où vous désireriez des précisions sur ce projet de recherche, n'hésitez pas à contacter le chercheur responsable du projet.

Cordialement,

Pierre-Marc DAIGNEAULT
Candidat au doctorat
Département de science politique
Faculté des sciences sociales
Université Laval.
Tél. : (1) 418-656-2131, poste 14994
Courriel : pierre-marc.daigneault.1@ulaval.ca

Version anglaise

[Salutations personnalisées],

I am requesting your participation to a study entitled "Validation of a Measurement Instrument of Stakeholder Participation in Evaluation". This research, conducted as part of my doctoral dissertation, aims to validate an instrument to be used to measure participatory evaluation.

You have been selected on the basis of your participation in the following evaluation: [**insérer référence complète**].

Participation in this study will involve filling out an electronic questionnaire on your level of participation in that specific evaluation. The questionnaire should only take 15 to 20 minutes to complete. You are free to accept or refuse to participate in this study and, to better inform your decision, we invite you to consult the introduction to the questionnaire available at the following link:

[**insérer lien sécurisé vers le questionnaire**]

This project was approved by the Université Laval Research Ethics Committee (Approval No. 2011-246/28-10-2011). If you have any questions regarding this research project, please do not hesitate to contact the researcher responsible for this project.

Best regards,

Mr. Pierre-Marc DAIGNEAULT
Ph.D. Candidate
Department of Political Science
Faculty of Social Sciences
Université Laval
Ph: (1) 418-656-2131, Ext. 14994
Email: pierre-marc.daigneault.1@ulaval.ca

Annexe E : Questionnaire

Version française

Page 1

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

FORMULAIRE DE CONSENTEMENT À L'INTENTION DES RÉPONDANTS AU QUESTIONNAIRE

Présentation

CHERCHEUR : M. Pierre-Marc DAIGNEAULT, candidat au doctorat, Département de science politique, Faculté des sciences sociales, Université Laval. Cette étude est réalisée dans le cadre de la thèse de doctorat du chercheur.

DIRECTEUR DE RECHERCHE : M. Steve JACOB, professeur agrégé, Département de science politique, Faculté des sciences sociales, Université Laval.

NOTE : avant d'accepter de participer à cette étude, veuillez prendre le temps de lire et de comprendre les renseignements qui suivent. Ce document vous explique le but de ce projet de recherche, ses procédures, avantages, risques et inconvénients. Nous vous invitons à poser toutes les questions que vous jugerez utiles au chercheur dont les coordonnées se trouvent à la page deux du présent formulaire.

Nature de l'étude

Cette étude vise à valider empiriquement un instrument développé par le chercheur et son directeur de recherche pour mesurer le caractère plus ou moins participatif d'une évaluation.[1]

DÉFINITIONS : l'**évaluation** cherche à porter un jugement sur la valeur d'une politique ou d'un programme en fonction de certains critères (pertinence, efficacité, etc.) et de générer des connaissances utiles à la prise de décision à partir d'une méthodologie rigoureuse. L'évaluation est **participative** dans la mesure où des parties prenantes telles que les employés de première ligne, les bénéficiaires d'un programme et/ou de groupes de la société civile (syndicats, groupes d'intérêts, comités de citoyens, etc.) sont impliquées dans les différentes tâches de l'évaluation (détermination des questions, choix des méthodes, etc.) et où le contrôle décisionnel est partagé entre celles-ci et l'évaluateur.

Déroulement de la participation

Votre implication au sein de cette recherche consiste à compléter un questionnaire accessible électroniquement sur le niveau de participation de votre évaluation. La durée prévue de votre participation est de 15 à 20 minutes.

Avantages, risques ou inconvénients possibles liés à la participation

BÉNÉFICES ANTICIPÉS POUR LE PARTICIPANT : le fait de participer à cette recherche vous offre une occasion de réfléchir à vos pratiques d'évaluation. Dans tous les cas, la participation à cette étude favorisera une meilleure compréhension de la nature de la participation et de ses dimensions.

Page 1(suite)

BÉNÉFICES ANTICIPÉS POUR LA COMMUNAUTÉ : le fait de participer à cette recherche constitue une occasion de contribuer au développement de connaissances sur la nature des pratiques participatives, en particulier sur sa mesure. Les connaissances à ce niveau permettront à terme d'améliorer la pratique de l'évaluation participative et contribuer à une meilleure prise en compte du point de vue du citoyen par les gouvernements.

INCONVÉNIENTS : il faut compter de 15 à 20 minutes pour compléter le questionnaire.

Participation volontaire et droit de retrait

Vous êtes libre de participer à ce projet de recherche. Vous pouvez aussi mettre fin à votre participation sans préjudice et sans avoir à justifier votre décision. Vous pouvez en outre refuser de répondre à certaines questions sans conséquence négative pour vous. Si vous décidez de mettre fin à votre participation, il est important d'en prévenir le chercheur dont les coordonnées sont incluses dans ce document. Tous les renseignements vous concernant seront alors détruits.

Confidentialité et gestion des données

- Aucun renseignement personnel ne sera colligé lors de cette étude. La nature de l'étude requiert cependant que le chercheur soit en mesure d'associer les données du questionnaire à chaque répondant;
- les données seront conservées sous format électronique pendant deux ans par le chercheur après quoi elles seront détruites. Seul ce dernier et le directeur de recherche auront accès à ces données;
- la recherche fera l'objet de publications scientifiques, et aucun participant ne pourra y être directement identifié car les données seront agrégées. Toutefois, malgré les mesures prises pour assurer la confidentialité de vos réponses, il se peut que vous soyez identifié comme un participant à cette recherche puisque la référence de votre article figurera dans la bibliographie de l'étude;
- les participants qui souhaitent recevoir un résumé des résultats, qui devraient être disponibles à partir du printemps 2012, sont invités à en informer le chercheur par courriel.

Pour des renseignements supplémentaires

Si vous avez des questions sur la recherche ou sur les implications de votre participation, veuillez communiquer avec M. Pierre-Marc DAIGNEAULT, candidat au doctorat, Département de science politique, Faculté des sciences sociales, Université Laval au numéro de téléphone suivant : 1 (418) 656-2131 poste 14994, ou à l'adresse courriel suivante : pierre-marc.daigneault.1@ulaval.ca.

Remerciements

Votre collaboration est précieuse pour nous permettre de réaliser cette étude et nous vous remercions d'y participer.

Consentement

Le fait de remplir le questionnaire sera considéré comme l'expression de votre consentement à participer au projet.

Plainte ou critique

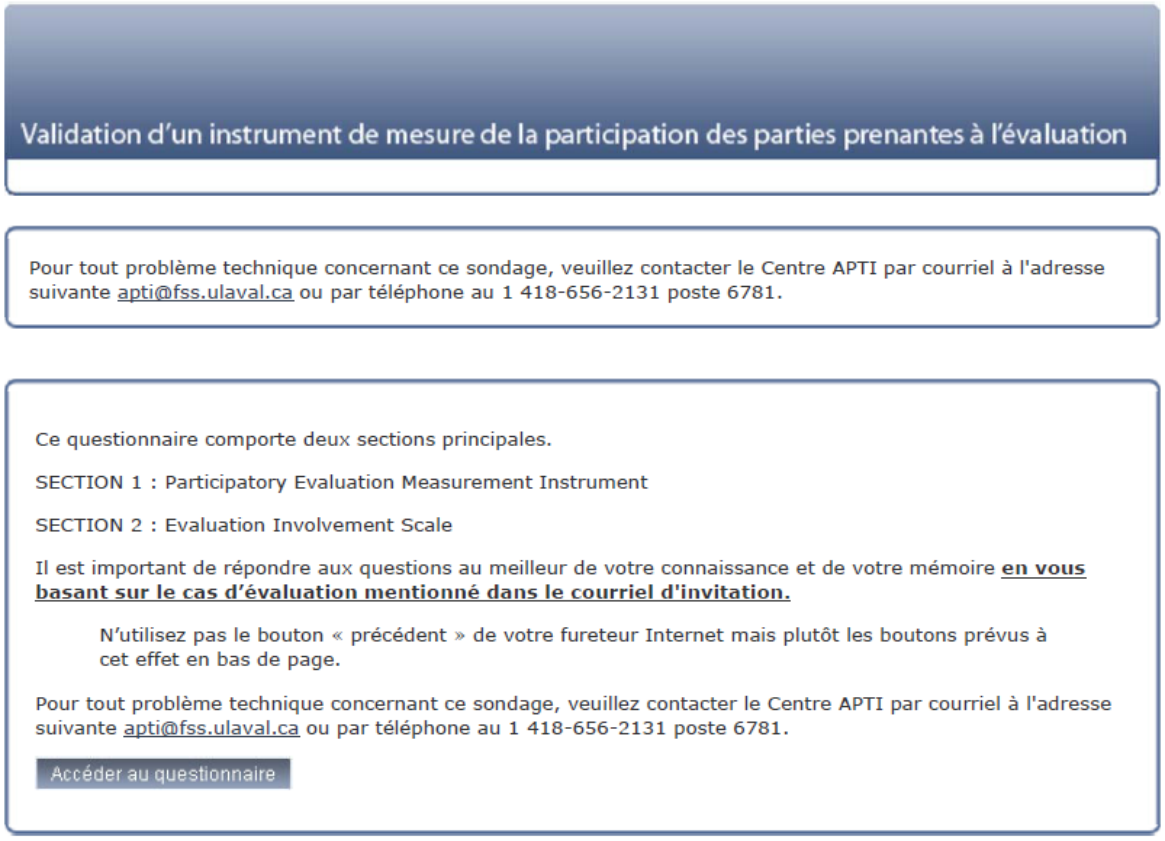
Toute plainte ou critique sur cette étude pourra être adressée au Bureau de l'Ombudsman de l'Université Laval :
Pavillon Alphonse-Desjardins, bureau 3320
2325, rue de l'Université
Université Laval
Québec (Québec) G1V 0A6
Renseignements - Secrétariat : 1 (418) 656-3081
Courriel : info@ombudsman.ulaval.ca

Ce projet a été approuvé par le Comité d'éthique de la recherche de l'Université Laval : no d'approbation 2011-246 / 28-10-2011.

Accepter

[1] Référence : 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.

Page 2



Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

Ce questionnaire comporte deux sections principales.

SECTION 1 : Participatory Evaluation Measurement Instrument

SECTION 2 : Evaluation Involvement Scale

Il est important de répondre aux questions au meilleur de votre connaissance et de votre mémoire **en vous basant sur le cas d'évaluation mentionné dans le courriel d'invitation.**

N'utilisez pas le bouton « précédent » de votre fureteur Internet mais plutôt les boutons prévus à cet effet en bas de page.

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

Accéder au questionnaire

Page 3

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

État d'avancement : Étendue de l'implication et diversité

1) Cochez les étapes de l'évaluation lors desquelles au moins un type de parties prenantes est impliqué de manière significative (c.-à-d., plus qu'à titre de source de données). N.B. : Chaque étape du processus comprenant plusieurs tâches (p. ex., colliger et analyser les données), participer à l'une de ces tâches est suffisant pour considérer qu'il y a implication d'un type de parties prenantes. Pour cocher une case, vous n'avez qu'à déplacer le curseur devant celle-ci et cliquer.

	Type 1 : Décideurs, concepteurs des politiques et gestionnaires	**Type 2 :** Responsables de la livraison du programme	**Type 3 :** Bénéficiaires directs et indirects, tiers lésés	**Type 4 :** Société civile et citoyens
1- Définir les questions et enjeux de l'évaluation et/ou préparer le devis évaluatif	☐	☐	☐	☐
2- Colliger et/ou analyser les données	☐	☐	☐	☐
3- Formuler les jugements et/ou les recommandations	☐	☐	☐	☐
4- Rapporter et/ou diffuser les résultats	☐	☐	☐	☐

Section suivante

Page 4

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

État d'avancement : Contrôle

2) Le contrôle du processus évaluatif réfère à la part de pouvoir que les parties prenantes possèdent par rapport aux évaluateurs relativement aux décisions affectant la réalisation de l'évaluation. Le contrôle est par conséquent appréhendé de manière relative. Il est à noter que le niveau de contrôle peut varier au cours de l'évaluation. Le score choisi devrait être représentatif de l'ensemble des étapes du processus au sein desquelles il y a implication des parties prenantes. Veuillez choisir l'option qui correspond le mieux à l'évaluation dans laquelle vous êtes impliqué(e).

Contrôle exclusif par l'évaluateur	○
Contrôle limité des participants à l'évaluation	○
Contrôle partagé également entre, d'un côté, les participants à l'évaluation et, de l'autre côté, l'évaluateur	○
Contrôle substantiel des participants à l'évaluation	○
Contrôle exclusif par les participants à l'évaluation	○

Section précédente | Reprendre plus tard | Section suivante

Page 5[51]

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

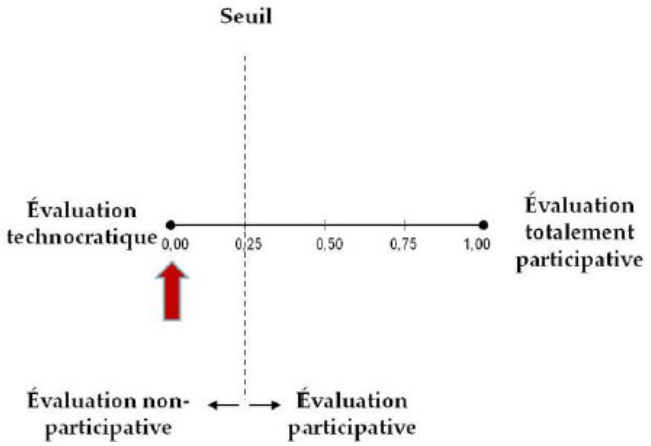
Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

État d'avancement : Score global de participation et réactions des répondants

Voici maintenant les résultats concernant le niveau de participation pour le cas d'évaluation auquel vous avez pris part.

	Scores
1 - Étendue de l'implication	0
2 - Diversité des participants	0
3 - Contrôle du processus	0.25
Niveau de participation global	0

Votre évaluation n'est pas participative ; elle est de type *technocratique*.



[51] Les résultats présentés sont hypothétiques. Ils varient en fonction des réponses au questionnaire.

Page 5 (suite)

3) Dans quelle mesure ces résultats correspondent-ils à votre opinion concernant le caractère plus ou moins participatif de ce cas d'évaluation ?

- Ne correspondent pas du tout
- Correspondent en partie
- Correspondent totalement
- Ne sais pas / Ne veut pas répondre

4) Pourquoi ?

Section précédente | Reprendre plus tard | Section suivante

Page 6

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

SECTION 2 : ÉCHELLE D'IMPLICATION DANS L'ÉVALUATION

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

État d'avancement : ÉCHELLE D'IMPLICATION DANS L'ÉVALUATION

Pour chaque question, s'il vous plaît choisir la réponse qui décrit le mieux la mesure dans laquelle des parties prenantes (*stakeholders*) autres que l'évaluateur ont été impliquées dans cette activité de l'évaluation.

1) Les parties prenantes ont été impliquées dans les discussions qui ont servi à déterminer la perspective de l'évaluation.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

2) Les parties prenantes ont été impliquées dans l'identification des membres de l'équipe d'évaluation.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

Page 6 (suite)

3) Les parties prenantes ont été impliquées dans le développement du plan d'évaluation.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

4) Les parties prenantes ont été impliquées dans le développement des instruments de collecte de données.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

5) Les parties prenantes ont été impliquées dans le développement des procédures de collecte de données.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

6) Les parties prenantes ont été impliquées dans la collecte de données.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

7) Les parties prenantes ont été impliquées dans l'examen du caractère exact et complet des données colligées.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

8) Les parties prenantes ont été impliquées dans l'analyse des données.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

9) Les parties prenantes ont été impliquées dans l'interprétation des données.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

Page 6 (suite)

10) Les parties prenantes ont été impliquées dans la rédaction des rapports d'évaluation.

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

11) Les parties prenantes ont été impliquées dans la présentation des conclusions de l'évaluation (p. ex.., au personnel et aux parties prenantes du programme, ou à un public externe).

- Non
- Oui, un peu
- Oui, dans une certaine mesure
- Oui, de manière extensive
- Je ne pense pas que cette activité a eu lieu

Section précédente | Reprendre plus tard | Soumettre

Page 7

Validation d'un instrument de mesure de la participation des parties prenantes à l'évaluation

Remerciements

Pour tout problème technique concernant ce sondage, veuillez contacter le Centre APTI par courriel à l'adresse suivante apti@fss.ulaval.ca ou par téléphone au 1 418-656-2131 poste 6781.

Nous vous remercions d'avoir pris le temps de participer à cette étude !

Version anglaise

Page 1

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

CONSENT FORM FOR QUESTIONNAIRE RESPONDENTS

Introduction

RESEARCHER: Mr. Pierre-Marc DAIGNEAULT, PhD Candidate, Department of Political Science, Faculty of Social Sciences, Université Laval. This research project is conducted as part of the researcher's doctoral dissertation.

RESEARCH SUPERVISOR: Mr. Steve JACOB, Associate Professor, Department of Political Science, Faculty of Social Sciences, Université Laval.

NOTE: Before accepting to participate in this study, please take the time to read and understand the following instructions. This document will explain the goal of this research project, its procedures, advantages, risks and disadvantages. We invite you to ask the researcher any and all questions you deem helpful to better inform your choice of participating in this study; his contact information can be found on Page 2 of this form.

Nature of the Study

This study aims to empirically validate an instrument developed by the researcher and his research supervisor to measure the participatory nature of a given evaluation.[1]

DEFINITIONS: **Evaluation** seeks to place a judgement on the value of a policy or program according to certain criteria (relevance, efficiency, etc.) and to generate useful knowledge for decision-making using a rigorous methodology. An evaluation is deemed **participatory** in the measure in which stakeholders such as front-line employees, program beneficiaries and/or civil society groups (unions, interest groups, citizen committees, etc.) are involved in the different tasks of evaluation (determination of questions, methodological choices, etc.) and where control of the process is shared between these stakeholders and the evaluator.

Participation

Your involvement in this research project would consist of completing a questionnaire available online and relating to the level of stakeholder participation of your evaluation. The estimated time commitment of this task is between 15 and 20 minutes.

Possible Advantages, Risks or Disadvantages Linked to Participation

ANTICIPATED BENEFITS FOR THE PARTICIPANT: Participation in this research project will give you the opportunity to reflect on your evaluation practices. In all cases, participation in this study will favour a better understanding of the nature of participation and its dimensions.

ANTICIPATED BENEFITS FOR THE COMMUNITY: Participating in this research will give you the opportunity to contribute to the development of knowledge on the nature of participatory practices, particularly on their measurement. Knowledge at this level, will, in the long run, improve evaluation practice and contribute to a better consideration of citizens' points of view by the government.

Page 1 (suite)

DISADVANTAGES: It will take 15 to 20 minutes to complete the questionnaire.

Voluntary Participation and Right to Withdraw

You are free to participate in this research project. You may also end your participation without prejudice and without needing to justify your decision. Additionally, you may refuse to answer particular questions without negative consequence. If you decide to end your participation, it is important that you inform the researcher whose contact information is included in this document. All information concerning your participation will then be destroyed.

Confidentiality and Data Management

- No personal information will be collected during this study. However, the nature of the study demands that the researcher be able to associate questionnaire data to each participant;
- The data will be conserved in an electronic format for two years by the researcher, after which the data will be destroyed. Only the researcher and his research supervisor will have access to this data;
- The research will be featured in scientific publications and no participant will be able to be directly identified since the data will be aggregated. However, despite the measures taken to assure the confidentiality of your responses, it is possible that you may be identified as a participant in this study since the reference of your article will figure into the bibliography of this study;
- Participants who wish to receive a summary of the study's results, which should be available in spring of 2012, are invited to inform the researcher by email.

For Further Information

If you have any questions on the research or the implications of your participation, please contact Mr. Pierre-Marc DAIGNEAULT, PhD Candidate, Department of Political Science, Faculty of Social Sciences, Université Laval at the following phone number: 1 (418) 656-2131 Ext. 14994 or at the following email address: pierre-marc.daigneault.1@ulaval.ca.

Acknowledgements

Your collaboration is invaluable to this study and we thank you for your participation.

Consent

The act of filling out this questionnaire will be considered to be an expression of your consent to participate in this project.

Complaints or Comments

All complaints or comments pertaining to this study can be addressed to:

Office of the Ombudsman (Université Laval)
Pavillon Alphonse-Desjardins, Office #3320
2325, rue de l'Université
Université Laval, Québec (Québec) G1V 0A6
Information: 1 (418) 656-3081
Email: info@ombudsman.ulaval.ca

This project was approved by the Université Laval Research Ethics Committee (Approval No. 2011-246 / 28-10-2011).

Accept

[1] Reference: 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.

Page 2

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

This questionnaire has two main sections.

SECTION 1 : Participatory Evaluation Measurement Instrument

SECTION 2 : Evaluation Involvement Scale

It is important that you answer each question to the best of your knowledge and memory. **Your answers to this questionnaire should be based on the evaluation case cited in the invitation email**.

Please do not use the "Back" button from your web browser: use instead the appropriate buttons in the bottom of each page.

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Access the questionnaire

Page 3

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Status : Extent of Involvement and Diversity of Participants

1) Please check off the evaluation steps in which at least one type of participant is meaningfully involved (i.e., they are more than a data source). N.B. : Since each step of the process includes more than one task (e.g., Collecting and analyzing data), one task is sufficient to consider that a particular type of stakeholder is involved. To check a box, simply place the cursor in front of the box and click.

	Type 1 : Policy makers, decision makers and managers	**Type 2 :** Those Responsible for Program Delivery	**Type 3 :** Direct and Indirect Beneficiaries & Affected Third Parties	**Type 4 :** Civil Society & Citizens
1- Defining the questions and issues of the evaluation and/or preparing the evaluation design	☐	☐	☐	☐
2- Collecting and/or analyzing the data	☐	☐	☐	☐
3- Formulating judgements and/or recommendations	☐	☐	☐	☐
4- Findings reporting and/or dissemination	☐	☐	☐	☐

Next section

Page 4

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Status : Control

2) Control of the evaluation process refers to the share of power that the stakeholders possess in relation to evaluators in what concerns the decisions affecting the conducting of the evaluation. Consequently, control is grasped in a relative manner. Your judgement should take into consideration the fact that the level of control can vary during the course of the evaluation. The chosen score should be representative of all steps in which stakeholders were involved. Please choose the option that best corresponds to the evaluation in which you are involved.

Exclusive control by the evaluator	○
Limited control by the evaluation's participants	○
Control shared equally between the evaluation's participants and the evaluator	○
Substantial control on the part of participants of the evaluation	○
Exclusive control on the part of participants of the evaluation	○

Previous section | Resume later | Next section

Page 5

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

SECTION 1 : Participatory Evaluation Measurement Instrument

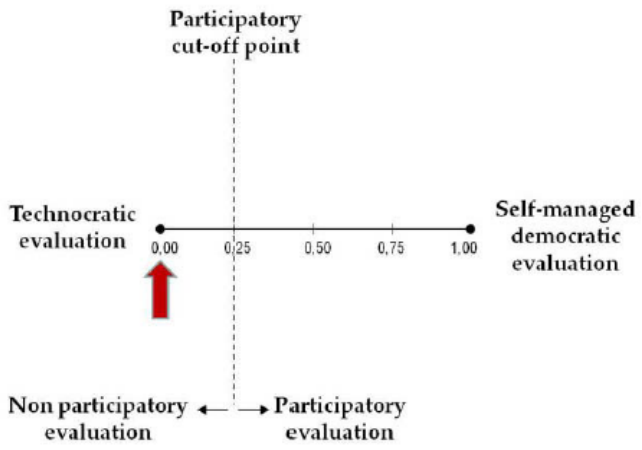
For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Status : Overall Participation Score and Reactions from Respondents

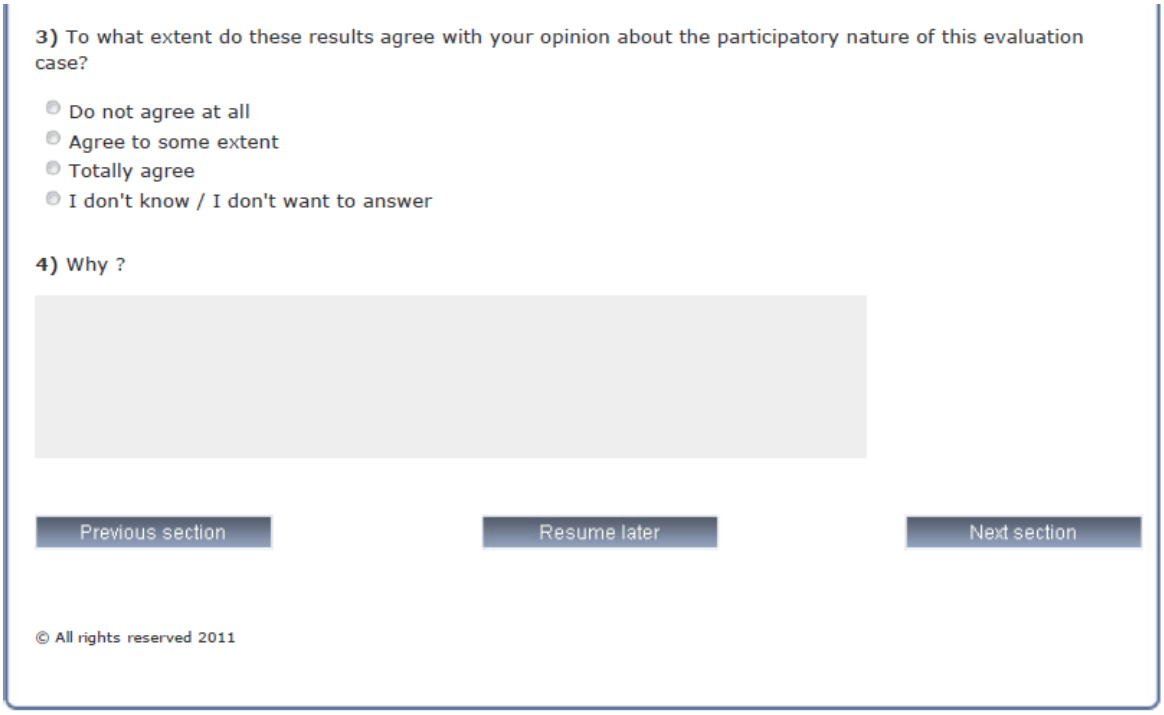
We now turn to the results on stakeholder participation for the evaluation in which you were involved.

	Scores
1- Extent of Involvement	0
2- Diversity of Participants	0
3- Control of the Evaluation Process	0.25
Overall Level of Participation	0

Your evaluation is not participatory; it is *technocratic*.



Page 5 (suite)



3) To what extent do these results agree with your opinion about the participatory nature of this evaluation case?

- Do not agree at all
- Agree to some extent
- Totally agree
- I don't know / I don't want to answer

4) Why ?

Previous section | Resume later | Next section

Page 6

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

SECTION 2 : EVALUATION INVOLVEMENT SCALE

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Status : EVALUATION INVOLVEMENT SCALE

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.

1) Stakeholders were involved in the discussions that focused the evaluation.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

2) Stakeholders were involved in identifying evaluation planning team members.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

3) Stakeholders were involved in developing the evaluation plan.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

4) Stakeholders were involved in developing data collection instruments.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

Page 6 (suite)

5) Stakeholders were involved in developing data collection processes.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

6) Stakeholders were involved in collecting data.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

7) Stakeholders were involved in reviewing collected data for accuracy and/or completeness.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

8) Stakeholders were involved in analyzing data.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

9) Stakeholders were involved in interpreting collected data.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don’t think this activity took place

10) Stakeholders were involved in writing evaluation reports.

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

Page 6 (suite)

11) Stakeholders were involved in presenting evaluation findings (e.g., to staff, to stakeholders, to an external audience).

- No
- Yes, a little
- Yes, some
- Yes, extensively
- I don't think this activity took place

Previous section | Resume later | Submit

Page 7

Validation of a Measurement Instrument of Stakeholder Participation in Evaluation

Thanks

For any problem regarding this questionnaire, please contact us at: apti@fss.ulaval.ca or 1 418-656-2131, ext. 6781.

Thank you for having participated in this study..

Annexe F : Résultats non agrégés – accord intercodeur

Échantillon	*n*	κ_{DoP1}	κ_{DoP2}	κ_{DoP3}	κ_{DoP4}	κ_{EoI1}	κ_{EoI2}	κ_{EoI3}	κ_{EoI4}
Échantillon final (échantillon principal + prétest 2)	40	0,59(.000)***	0,50(.003)**	0,75(.000)***	0,38(.029)*	0,72(.001)***	1,00(.000)***	0,75(.000)***	0,76(.000)***
Échantillon commun aux assistants et aux auteurs	25	0,50(.057)	-0,07(1.00)	0,29(.202)	0,19(.562)	-0,04(1.00)	0,32(.166)	-0,07(1.00)	0,11(1.00)

NOTE: DoP = Depth of participation; EoI = Extent of involvement; les indices 1, 2, 3 et 4 désignent les indicateurs

*p = significatif au niveau de 0,05; **p = significatif au niveau de .01; ***p = significatif au niveau de .001

Annexe G : Cadre conceptuel détaillant la nature, les conditions contextuelles et les conséquences de l'évaluation participative