L'examen de la qualité des évaluations fédérales : une méta-évaluation réussie ?

Pierre-Marc Daigneault

Université Laval

Résumé: La qualité des évaluations représente un enjeu fondamental au sein du mouvement de la gestion et des politiques fondées sur des données probantes. Le Secrétariat du Conseil du Trésor du Canada (SCT) a ainsi réalisé en 2004 un examen de la qualité des évaluations fédérales d'une envergure considérable. Au regard de sa pertinence, de sa méthodologie, et de la crédibilité de ses conclusions, cette méta-évaluation est-elle réussie? Cet article offre une réponse à cette question en présentant une évaluation de l'examen du SCT. Malgré des lacunes importantes au niveau de la théorie et des critères de qualité, du devis de recherche, du processus de codage, et de l'analyse des données, cette méta-évaluation présente des conclusions qui sont dans l'ensemble pertinentes et d'un niveau de crédibilité acceptable. Les enseignements que l'on peut tirer de cet examen sont proposés à titre de recommandations à l'intention des évaluateurs et des responsables des services d'évaluation désirant réaliser un examen similaire mais qui ne souffrirait pas des mêmes lacunes.

Abstract: Evaluation quality is a fundamental issue within the evidence- based management and policy movement. The Treasury Board Secretariat of Canada (TBS) conducted a wide-ranging review of the quality of federal evaluations in 2004. In terms of its relevance and methodology and the credibility of its conclusions, is the meta-evaluation a success? This article answers the question by presenting an evaluation of the TBS quality review. Despite serious shortcomings with respect to the quality theory and criteria, design, coding process, and data analysis, the meta-evaluation conclusions are relevant and credible overall. Lessons learned from this quality review are proposed as recommendations for evaluators and officials responsible for evaluation units who wish to conduct a similar review that does not suffer from the same shortcomings.

L'enjeu de la qualité

Originaire du domaine de la santé et s'étendant progressivement à d'autres domaines (éducation, politiques et management publics, etc.), le mouvement « *evidence-based* » est caractérisé par son exigence à l'effet que la pratique professionnelle et les décisions

soient systématiquement fondées sur les meilleures connaissances disponibles (Oakley, 2002). Les prescriptions spécifiques de ce courant pour l'évaluation sont sans équivoque : les décisions relatives aux politiques et programmes publics doivent être guidées par des conclusions et des recommandations provenant d'évaluations de qualité, c'est-à-dire rigoureuses, valides, fiables, et crédibles. Dans le cas contraire, de « mauvaises » décisions risquent d'être prises. Certains auteurs se sont en effet inquiétés des conséquences négatives de l'utilisation, intentionnelle ou non, d'évaluations de mauvaise qualité (Muir, 1999, Smith, 1999, cités dans Schwartz & Mayne, 2005a). De telles évaluations peuvent induire en erreur les décideurs en leur suggérant par exemple d'augmenter le financement de programmes à l'efficacité douteuse. Si la non-utilisation d'évaluations de mauvaise qualité est justifiable sur le plan normatif (Christie & Alkin, 1999; Cousins & Shulha, 2006; Patton, 2005), cela représente par ailleurs un gaspillage de ressources et une opportunité manquée d'améliorer la pertinence, l'efficacité, l'efficience, et l'équité des interventions publiques. Plusieurs études empiriques ont ainsi établi que les évaluations de mauvaise qualité sont moins susceptibles d'être utilisées par les titulaires de charge publique (Alkin, Daillak, & White, 1979; Caplan, 1977; Weiss & Bucuvalas, 1980). Weiss et Bucuvalas ont par exemple démontré que les décideurs appliquent des « tests de vérité » aux rapports qui portent notamment sur leur qualité méthodologique. De toutes les variables examinées dans cette étude, la qualité de l'évaluation est celle qui présente l'association statistique la plus forte avec l'utilisation. L'importance de ce facteur a d'ailleurs été corroborée par plusieurs revues de littérature et synthèses de connaissances (Cousins & Leithwood, 1986; Huie Hofstetter & Alkin, 2003; Leviton & Hughes, 1981). Le constat à l'effet que la qualité méthodologique d'une évaluation augmente les probabilités qu'elle exerce une influence sur le processus décisionnel trouve également écho chez plusieurs auteurs (Henry, 2003; Schwartz & Mayne, 2005a).

Certes, le modèle de décision rationaliste qui est proposé par le mouvement des données probantes est davantage un idéal qu'une description juste de la réalité (Albæk, 1995). Christie (2007) a par exemple démontré que les informations de nature anecdotique exercent aussi à l'occasion une influence réelle sur la prise de décision. D'autres considérations que la raison instrumentale, notamment le savoir tacite difficilement formalisable, peuvent en outre influer sur la prise de décision (Schwandt, 2008). D'autres études ont établi que la congruence entre les conclusions d'une évaluation et les intérêts d'un acteur exerce dans certains cas une influence plus importante sur son utilisation que sa qualité méthodologique (Albæk, 1996; Rich & Oh, 2000). En somme, si la qualité de l'évaluation ne garantit pas son impact sur la prise de décision relative aux programmes publics, elle accroît ses probabilités d'utilisation et contribue à une prise de décision mieux informée et donc plus rationnelle. Dans ce contexte, il n'est pas surprenant que les autorités publiques de plusieurs juridictions adoptent des dispositifs visant à assurer la qualité des évaluations.

L'un des dispositifs d'assurance qualité les plus discutés au sein de la profession de l'évaluation est certainement la **méta-évaluation** (voir Birch & Jacob, 2005; Cook & Gruder, 1978; Cooksy & Caracelli, 2005; Mark, 2001; Scriven, 2005b; Stufflebeam, 2001b; Worthen, 2001). Certains standards de qualité de la profession évaluative lui sont

d'ailleurs directement consacrés (voir Joint Committee on Stan- dards for Educational Evaluation, 1994; Widmer, Landert, & Bach- man, 2000). Scriven (2005b) définit simplement la méta-évaluation comme « l'évaluation d'évaluations » (p. 53, traduction de l'auteur). Contrairement à l'évaluation dite « primaire » qui a pour objet un pro- gramme public, la méta-évaluation est une « évaluation secondaire » qui applique la logique et les méthodes évaluatives à des processus et rapports d'évaluation. Bien que la méta-évaluation puisse servir à mieux comprendre les processus de politiques publiques dans leur phase d'évaluation (Bustelo, s.d.), sa fonction principale consiste d'abord à déterminer et à contrôler la qualité de l'évaluation (Cooksy & Caracelli, 2005). Ainsi, la méta-évaluation est considérée par plu- sieurs comme une obligation éthique et un impératif professionnel (Scriven, 2005b; Stufflebeam, 2001b). Scriven (2005b) soutient par exemple que puisque l'évaluation est un domaine réflexif et auto- référent, il n'y a aucune raison de limiter l'appréciation de la qualité aux seules évaluations primaires. Après tout, les conclusions d'une méta-évaluation alimentent tout autant le processus décisionnel que celles d'une évaluation. À l'instar de l'évaluation primaire, la réalisation d'une méta-évaluation ne répond pas à un modèle unique. Elle peut ainsi être réalisée à l'interne ou à l'externe de l'organisation; de manière prospective, concomitante, ou rétrospective; utiliser une ou plusieurs méthodes et sources de données; posséder une visée formative ou sommative (Bustelo, s.d.; Cook & Gruder, 1978; Cooksy & Caracelli, 2005; Scriven, 2005b). Une méta-évaluation peut en outre porter sur une seule ou plusieurs évaluations et ré-analyser ou non les données originales d'évaluation (Cook & Gruder, 1978; Cooksy & Caracelli, 2005).

Or, malgré les appels insistants en faveur d'un recours accru à la méta-évaluation, il existe relativement peu de publications empiriques traitant de cette pratique (voir Addison & Amo, 2005; Worthen, 2001). Si ce nombre est peu élevé, il n'existe à la connaissance de l'auteur aucun article qui présenterait une évaluation systématique d'une méta-évaluation ou ce qu'il conviendrait d'appeler une « méta-méta- évaluation » ou « évaluation tertiaire ». L'évaluation est tertiaire en ce sens qu'elle prend pour objet une méta-évaluation : il y a donc trois degrés de séparation entre celle-ci et le programme public évalué à l'origine. On trouve certes des articles présentant des réflexions générales sur le thème de la méta-évaluation et de ses méthodes (voir Cook & Gruder, 1978; Scriven, 2005a; Scriven, 2005b; Stufflebeam, 2001a, 2001b) ou encore des cas spécifiques de méta-évaluations (e.g., Cooksy & Caracelli, 2005; Uusikylä & Virtanen, 2000), mais aucun exemple d'examen critique et systématique de la qualité d'une méta- évaluation. Une telle évaluation contribuerait pourtant à produire un portrait réflexif sur les pratiques d'excellence pour la réalisation d'une méta-évaluation.

Cet article tente de remédier à cette lacune en présentant l'évaluation d'une méta-évaluation, soit l'*Examen de la qualité des évaluations dans les ministères et les organismes* réalisé par le Secrétariat du Conseil du Trésor du Canada (SCT, 2004). Bien que la littérature théorique soit utilisée afin de décrire et classer l'examen du SCT, la contribution de cette étude est d'abord de nature appliquée. Il s'agit d'une part de procéder à une évaluation de la qualité de la méta- évaluation du SCT et de déterminer si cette dernière est « réussie » du point de vue de sa pertinence, de sa qualité méthodologique, et de la crédibilité de ses conclusions. Il s'agit d'autre part de tirer

les leçons de cet examen afin d'améliorer l'état des pratiques méta-évaluatives. La typologie de Schwartz et Mayne (2005a; 2005b) est tout d'abord utilisée afin de réaliser un bref état des lieux de l'infrastructure d'assurance qualité mise en place au gouvernement fédéral, de situer l'examen de qualité par rapport à cette infrastructure et d'en présenter les faits saillants. La seconde partie consiste en une présentation du cadre d'évaluation (i.e., les objectifs, méthodes, et données) retenu pour évaluer l'examen du SCT. La troisième partie présente les résultats de cette évaluation sous forme de constats et de conclusions. La dernière partie offre une discussion des résultats et des limites de cette évaluation. Des recommandations à l'intention des évaluateurs et responsables des services d'évaluation désirant réaliser un exercice méta-évaluatif similaire mais ne souffrant des mêmes lacunes ainsi que quelques pistes pour de futures recherches sont également proposées.

Le contrôle de la qualité des évaluations au gouvernement fédéral

Une typologie des approches d'assurance qualité

Le développement de la fonction d'évaluation au gouvernement du Canada s'est structuré autour de différents exercices ponctuels d'appréciation de la qualité de l'évaluation dont l'examen du SCT n'est qu'un exemple. Jacob (2006) affirme à cet égard que le dispositif institutionnel d'évaluation du fédéral est caractérisé par une « recherche constante de la qualité » (p. 516). Les exercices méta-évaluatifs ayant été réalisés ont jouit d'une bonne visibilité, certes, mais ils ne représentent que la pointe de l'iceberg. Le gouvernement fédéral a eu recours à travers les années à l'ensemble des approches d'assurance qualité présentées précédemment. Cette partie n'offre pas de présentation exhaustive du dispositif fédéral d'évaluation actuel ni de son évolution historique (sur ce point, voir Jacob; Müller-Clemm & Barnes, 1997; Segsworth, 2005). L'objectif est plutôt d'utiliser la typologie précédente pour brosser à grands traits les contours du dispositif fédéral de contrôle de la qualité des évaluations, ce qui permettra par la suite de comprendre où l'examen du SCT s'insère au sein de ce dispositif.

Si les approches structurelles forment la pierre d'assise sur laquelle les autres approches sont fondées, la *Politique d'évaluation* (SCT, 2001) représentait sans contredit les fondations du dispositif de contrôle de la qualité au gouvernement fédéral au moment de l'examen de qualité (une nouvelle politique d'évaluation a été adoptée en 2009). Ce document clarifie les responsabilités des acteurs gouvernementaux par rapport à l'évaluation et précise le rôle joué par celle-ci au sein de l'administration fédérale. La fonction de l'évaluation y est appréhendée dans une perspective de gestion fondée sur des données probantes : « l'évaluation rigoureuse et objective est un outil important qui aide les gestionnaires à gérer les résultats » (SCT, 2001, p. 1). La politique énonce en outre des normes de qualité qui doivent guider les ministères dans leur recours à l'évaluation. Ces normes portent notamment sur les questions d'évaluation devant être abordées, la compétence des évaluateurs et l'intégrité et l'objectivité des rapports.

Les normes générales énoncées par la politique sont partiellement précisées dans le *Guide pour l'examen des rapports d'évaluation* (SCT, Centre d'excellence en évaluation,

2004). Le Centre d'excellence en évaluation (CEÉ), une entité administrative qui relève du SCT (2005), a été mis en place dans la foulée de la politique de 2001. Le CEÉ a pour mission d'exercer un leadership et de guider la pratique évaluative au gouvernement fédéral. Il organise notamment des activités de formation et de perfectionnement pour les nouveaux évaluateurs, un autre mécanisme issu de l'approche structurelle.

Du côté des approches formatives, un exemple de mécanisme d'assurance qualité est l'offre des services-conseils par le CEÉ aux évaluateurs des ministères lors de la réalisation d'évaluations (SCT, 2005). Ces conseils prodigués en cours de route permettent aux évaluateurs d'ajuster le tir et ainsi améliorer la qualité des évaluations produites. Le CEÉ fournit également des conseils sur les plans d'évaluation et autres documents administratifs liés indirectement à l'évaluation tels que les cadres de responsabilisation et présentations au Conseil du Trésor. Du côté sommatif, la méta-évaluation ponctuelle des rapports fédéraux réalisée par le SCT en 2004 est un exemple très visible de mécanisme issu de cette approche. Le suivi électronique continu de la qualité des rapports d'évaluation constitue un autre mécanisme issu de cette approche (SCT, 2005). Du côté des approches systémiques, notons l'examen des fonctions d'évaluation de 38 ministères réalisé par le CEÉ (SCT, 2005). Ce contrôle de qualité avait pour objectif d'examiner les systèmes de production de l'évaluation au sein de chaque ministère, notamment l'existence d'un engagement à renforcer la capacité, d'un plan d'évaluation axé sur le risque, et d'un comité d'évaluation actif.

Le portrait tracé précédemment permet de situer la méta-évaluation du SCT dans le contexte plus large des efforts du gouvernement canadien pour assurer la qualité des évaluations. La qualité de l'évaluation semble ainsi être une préoccupation constante au sein de l'administration publique fédérale qui ne se limite pas à des examens rétrospectifs et ponctuels des rapports.

Objectifs et faits saillants de l'examen du SCT

L'examen méta-évaluatif du SCT est un mécanisme issu de l'approche sommative d'assurance qualité qui s'inscrit dans le mandat de surveillance de la qualité des évaluations et de renforcement de la capacité évaluative du CEÉ. La réalisation de cet examen a été motivée par deux questions de recherche principales qui s'insèrent dans la foulée des efforts déployés pour assurer la qualité évaluative depuis le début des années 2000 : (a) La qualité des rapports d'évaluation est-elle acceptable ? et (b) Est-ce qu'il y a eu amélioration de la qualité suite à l'adoption de la *Politique d'évaluation* en 2001? (SCT, 2004). L'exercice a été piloté par le CEÉ du SCT, en collaboration avec un groupe de travail formé de huit fonctionnaires provenant de ministères et organismes différents ainsi que de trois évaluateurs externes de la firme Ekos Research Associates (Ekos). Le groupe de travail a fourni des commentaires et des suggestions sur la portée de l'examen, les critères d'évaluation, et l'instrument de mesure de la qualité utilisés. Les évaluateurs d'Ekos ont contribué au développement de l'instrument d'évaluation et ont effectué le codage et l'analyse des résultats.

L'examen a porté sur la qualité des rapports d'évaluation, par opposition à leurs processus ou à leurs impacts (SCT, 2004). Pour ce faire, un gabarit d'examen (ou grille

évaluative) a été développé à partir des critères de qualité énoncés dans divers documents fédéraux, notamment la *Politique d'évaluation* (SCT, 2001) et le *Guide pour l'examen des rapports d'évaluation* (SCT, CEÉ, 2004). On compte près d'une centaine de critères assez diversifiés qui concernent notamment la description du programme évalué et du contexte de l'évaluation (objectifs, portée, éléments d'évaluation, etc.), la méthodologie (fiabilité des données, triangulation des méthodes, recours à des groupes de comparaison, etc.), la qualité des jugements (i.e., les conclusions de l'évaluation s'appuient sur des constats valides qui sont eux-mêmes fondés sur des données), ainsi que la forme du rapport (concision, clarté, présence d'un résumé, etc.). Les indicateurs sont de nature ordinale ou nominale et ont été réunis au sein d'une même grille que l'on trouve à l'annexe A du rapport de l'examen (SCT, 2004). Plusieurs versions du gabarit d'examen ont été développées et celui-ci a fait l'objet de prétests. La fiabilité interjuge a en outre été appréciée de manière qualitative avant l'examen. Initialement, un échantillon stratifié des rapports produits dans les deux ans suivant l'adoption de la *Politique d'évaluation* devait être constitué, ce qui, selon les estimations du CEÉ, aurait représenté environ 500 rapports pour la période allant d'avril 2001 à avril 2003. Or, le SCT ne disposait que de 122 rapports complets (les études et rapports de vérification ayant été écartés), dont 115 ont finalement été retenus pour l'examen (7 rapports ayant été exclus pour des raisons de surreprésentation). Les rapports retenus provenaient finalement de 35 ministères et organismes différents. Chacun de ces rapports a été examiné par un seul évaluateur pendant deux heures et demie en moyenne. Des analyses descriptives croisées ont été effectuées afin de déterminer si la taille et le type d'entité ayant produit le rapport (petite, moyenne ou grande; ministère ou organisme), de même que l'année de production (avant ou après avril 2002) ont influencé la qualité. Les conclusions de l'examen de qualité sont nombreuses et, compte tenu de la diversité des critères d'évaluation, touchent à plusieurs aspects des rapports. Une présentation même sommaire des forces et faiblesses des évaluations fédérales alourdirait indûment le texte et, surtout, est disponible ailleurs (pour un résumé, voir Jacob, 2006; pour une présentation exhaustive, voir SCT, 2004). Il suffit ici de noter que la qualité globale des rapports examinés était adéquate ou plus qu'adéquate dans la grande majorité (77 %) des cas. Au niveau de l'analyse des déterminants de la qualité, aucune variation claire de la qualité des rapports n'a été constatée par rapport au type ou à la taille des organisations. Une amélioration de la qualité a toutefois été observée après avril 2002 : d'une part, la proportion de rapports inadéquats est passée de 32 à 18 % alors que la proportion de rap- ports plus qu'adéquats est passée de 22 à 37 % ; d'autre part, des améliorations ont été constatées par le SCT sur plusieurs critères de qualité. L'examen contient par ailleurs quatre recommandations relatives à l'amélioration de la qualité des rapports, la révision et la diffusion des attentes du SCT en matière de méthodologie et de qualité des rapports, la surveillance de la qualité à l'aide de fiches de rendement individuelles par ministère ou organisme, et la mise en place de normes et d'incitatifs afin d'assurer la qualité (SCT, 2004).

Cadre d'évaluation

Objectifs

Cette évaluation vise principalement à déterminer dans quelle mesure le devis, les méthodes et les données de l'examen du SCT ont permis de générer des conclusions valides, fiables, et crédibles sur la qualité des évaluations fédérales. Elle vise également à explorer la pertinence et les impacts de l'examen de même que l'accueil qui lui a été réservé par la communauté d'évaluation fédérale.

Théorie et critères de qualité

À l'instar de tout autre type d'évaluation, la réalisation d'une méta- évaluation nécessite le recours à divers critères normatifs. Quatre types de critères sont généralement utilisés dans le cadre de méta- évaluations : les standards d'évaluation de programme du Joint Committee on Standards for Educational Evaluation (1994), les principes directeurs de l'American Evaluation Association (2004), les critères méthodologiques traditionnels des sciences sociales, et les critères dits émergents (Cooksy & Caracelli, 2007). Puisque cette méta-évaluation porte d'abord sur la qualité méthodologique de l'examen du SCT et sur la validité de ses conclusions, les critères traditionnels de sciences sociales et les critères dits de « précision » de la Société suisse d'évaluation (Widmer et al., 2000), calqués sur ceux du Joint Committee, semblent les plus appropriés. Certains des critères appliqués aux évaluations fédérales par le SCT dans le cadre de son examen sont également pertinents. En ce sens, cette évaluation est davantage axée sur la qualité du point de vue « du producteur », c'est-à-dire la qualité méthodologique de l'examen du SCT, que « du consommateur », soit son utilité (Toulemonde, 2007).

Il faut cependant souligner qu'aucun critère spécifique d'évaluation n'a été énoncé de manière explicite avant la méta-évaluation. Seules les catégories générales devant faire l'objet de l'évaluation avaient été identifiées *a priori*, soit la conception et les critères de qualité; la qualité méthodologique et la validité des constats et jugements qui en découlent; pertinence, utilité, et impacts. Par conséquent, aucune grille formelle d'évaluation, à l'instar de celle utilisée par le SCT dans le cadre de l'examen de qualité, n'a été développée pour cette évaluation. Au regard des coûts en termes de développement et de validation, le recours à une grille formelle ne semblait pas justifié pour l'évaluation d'une évaluation unique, comme c'est le cas ici. Des critères un peu plus spécifiques, qui sont en réalité des questions et des éléments d'évaluation servant à guider et informer le jugement plutôt que des critères devant faire l'objet d'un codage mécanique, ont en partie « émergé » à la lecture des documents du SCT et lors des entretiens (voir Tableau 1).

Méthode et données

Tel que mentionné précédemment, il existe plusieurs approches méthodologiques appropriées pour la réalisation d'une méta-évaluation. L'approche retenue ici se concentre sur la **critique du devis d'évaluation**, soit une approche qui demande peu de ressources mais permet néanmoins de générer des constats justes et utiles (Scriven, 2005b; voir aussi Cook & Gruder, 1978). Elle consiste en une appréciation des différentes caractéristiques d'une évaluation, notamment sa méthode, à l'aune de critères de qualité. Cette approche se base généralement sur les données contenues

exclusivement dans le rapport et ne procède pas à une analyse secondaire des données originales de l'évaluation. La présente évaluation diffère de ce modèle à deux égards. Premièrement, certaines analyses secondaires ont été effectuées à partir des données du rapport de l'examen et de son annexe technique.

Deuxièmement, les données de l'examen du SCT ont été complétées par des entretiens réalisés auprès de répondants-clés des administrations publiques fédérale et provinciale. Le recours aux entretiens visait deux objectifs. Il s'agissait tout d'abord de colliger des données portant sur le contexte de réalisation de l'examen du SCT, notamment l'accueil qui lui a été réservé par les évaluateurs. Cette procédure de triangulation des sources de données permet d'accroître la validité et la fiabilité des résultats et (Mathison, 2004; Patton, 1987, cité dans Yin, 2003). Le recours à des entretiens visait ensuite à étayer les conclusions, interprétations, et jugements formulés à partir des documents du SCT en faisant appel à d'autres individus, ce qui correspond au principe de triangulation des investigateurs. Puisque les entretiens répondaient d'abord à une logique de généralisation analytique et théorique plutôt qu'à une logique de généralisation statistique (Beaud, 2003; Yin, 2003), la sélection des répondants a été effectuée « par choix raisonné » (Beaud), c'est-à-dire en fonction de considérations théoriques. Ce type d'échantillon possède en outre l'avantage d'être peu coûteux et facile d'utilisation. En revanche, les opinions exprimées par les répondants ne sont pas nécessairement représentatives de celles des fonctionnaires canadiens et québécois œuvrant dans le domaine de l'évaluation et doivent de ce fait être interprétées avec prudence.

Cet échantillon est composé de trois catégories de répondants. La première est constituée d'évaluateurs et de responsables de servi- ces d'évaluation fédéraux dont l'organisation a vu ses rapports examinés par le SCT. Le nombre de rapports d'une même organisation ayant été évalués par le SCT variant de un à douze, il a été décidé de contacter les services d'évaluation des huit organisations ayant eu plus de cinq rapports examinés sur la base du postulat selon lequel les fonctionnaires de ces organisations étaient davantage susceptibles de connaître l'examen et d'avoir de ce fait des commentaires à formuler à son endroit. Quatre répondants provenant d'organisations différentes ont ainsi accepté de participer à l'étude pour un taux de réponse de 50 % sur une base organisationnelle. Deuxièmement, un entretien avec un fonctionnaire du SCT ayant été directement impliqué dans la réalisation de l'examen de qualité semblait essentiel. Nous avons donc contacté l'un des responsables de l'examen de qualité sur la recommandation d'un évaluateur fédéral ayant été approché mais ayant refusé de participer. Enfin, une organisation rassemblant des fonctionnaires d'une juridiction provinciale dont la mission est d'assurer le développement de la fonction d'évaluation au sein de la fonction publique a été contactée en raison de son expérience en matière de méta-évaluation. Cette organisation organise en effet un concours annuel lors duquel la qualité d'évaluations réalisées par les ministères et organismes de cette juridiction est appréciée par un jury d'experts. Deux membres de cette organisation ont ainsi accepté de participer à cette étude (aucun refus), ce qui était suffisant pour les besoins de cette étude. En somme, sept répondants, dont cinq occupent des fonctions au sein de l'administration

fédérale, ont accepté de participer à des entretiens de type semi-dirigés. Une grille contenant principale- ment des questions ouvertes et laissant une certaine flexibilité aux personnes interviewées dans leurs réponses a été utilisée (voir Annexe A pour les thèmes abordés). Les entretiens impliquant les fonctionnaires fédéraux ont été effectués par téléphone à l'été 2006 (durée moyenne de 25 à 30 minutes) alors que ceux visant les fonctionnaires provinciaux ont été réalisés en personne au printemps 2007 (durée moyenne de 60 minutes). Dans au moins un cas, des questions de suivi ont été acheminées par courriel.

Résultats

Les résultats de cette évaluation se situent à deux niveaux. Le premier niveau est formé de **constats** qui sont davantage de nature factuelle et descriptive. Leur niveau de certitude est par conséquent élevé. Le second niveau est formé de **conclusions** qui constituent des interprétations et des jugements portés sur ces mêmes constats. Si ce second type de résultat est davantage subjectif, il représente par ailleurs un intérêt beaucoup plus grand pour les évaluateurs et responsables des services d'évaluation des administrations publiques. La distinction entre constats et conclusions donne une plus grande transparence au processus d'évaluation, ce qui permet au lecteur d'apprécier plus facilement la crédibilité des résultats présentés ici. L'ordre de présentation des résultats suit généralement celui des questions et éléments d'évaluation (voir Tableau 1).

Tableau 1. Liste des questions et des éléments d'évaluation

La conception et les critères de qualité sont-ils adéquats?

- Description et justification de la conception théorique de la qualité évaluative
- Couverture des dimensions importantes de la qualité
- Cohérence des critères de qualité avec la conception de la qualité

Les conclusions sont-elles valides, fiables, et crédibles?

- Description claire des questions et objectifs de la méta-évaluation
- Description de l'objet d'évaluation et de son contexte
- Pertinence du devis d'évaluation et des méthodes par rapport aux questions et objectifs de la méta-évaluation
- Qualité des données / représentativité de l'échantillon
- Rigueur du processus de collecte et de codage
- Pertinence et rigueur de l'analyse
- Conclusions fondées sur les constatations de l'évaluation

L'exercice méta-évaluatif est-il pertinent et utile?

- Réponse aux besoins informationnels des parties prenantes
- Effets voulus et non voulus induits par la méta-évaluation

La théorie et les critères de qualité sont-ils adéquats?

Constat : l'examen du SCT est fondé sur des critères mais pas sur une théorie complète et explicite de la qualité évaluative

Plutôt que de sélectionner une conception théorique de la qualité, d'en analyser les différentes dimensions, et d'en inférer des indicateurs de mesure pour développer son gabarit d'examen, le SCT (2004) s'est basé sur une recension des critères de qualité contenus dans différents documents fédéraux traitant de cette question. Le SCT ne fait appel à aucune théorie explicite en matière de qualité telles que l'évaluation axée sur l'utilisation (Patton, 1997; 2001) ou le cadre théorique proposé par certaines associations professionnelles d'évaluation (i.e., Joint Committee on Standards for Educational Evaluation, 1994; Widmer et al., 2000).

Conclusions

Le recours à des critères en l'absence de théorie explicite de la qua- lité évaluative pose des problèmes de validité analogues à ceux d'un scientifique qui utilise des indicateurs pour mesurer un concept qui n'a pas été préalablement théorisé de manière explicite (i.e., problème de validité du construit). D'une part, la pertinence des critères utilisés par le SCT dans son examen est difficile à apprécier dans l'abstrait. En effet, comment déterminer si les critères retenus mesurent véritablement les différentes dimensions de la qualité des évaluations? D'autre part, comment s'assurer qu'aucune dimension fondamentale ou facette de la qualité n'a été négligée dans l'examen?

La reconstruction de la théorie implicite de qualité du SCT à partir des critères qu'il a utilisés dans son examen et sa comparaison avec d'autres théories permet de suggérer quelques pistes de réponses à cette question. Une comparaison avec l'évaluation axée sur l'utilisation de Patton (1997) permet par exemple de constater que la grille d'examen est, à l'exception de quelques critères tels que l'inclusion de la perspective de tous les intéressés, muette sur l'implication des par- ties prenantes. L'implication et la participation des parties prenantes à l'évaluation constitue pourtant une question fondamentale au sein de la profession (Cousins & Whitmore, 1998; King, 2005; Mark, 2001; Patton, 1997; 2001). Un répondant a ainsi suggéré qu'il aurait trouvé pertinent que la question de la participation des parties prenantes et leur appropriation de la démarche et des résultats de l'évaluation soit examinée dans le cadre de la méta-évaluation fédérale. Par ailleurs, le SCT ne semble pas considérer l'éthique et la déontologie comme une dimension de la qualité évaluative, contrairement au Joint Committee (1994) et à la Société suisse d'évaluation (Widmer et al., 2000) notamment. Cette dimension regroupe des questions liées à la protection des droits individuels telles que la confidentialité offerte aux répondants, la protection des renseignements personnels, et la déclaration de conflits d'intérêt relatifs à l'évaluation. Étant donné l'implication du gouvernement fédéral dans la livraison de services à plusieurs groupes vulnérables (autochtones, chômeurs, immigrants, handicapés, etc.) en particulier, il aurait été opportun que le SCT se penche sur ces questions. Dans tous les cas,

l'absence d'une théorie explicite ne permet pas de déterminer si ces omissions sont intentionnelles et, le cas échéant, justifiables. Le fait que le SCT se soit concentré sur la qualité des rapports pour son examen, par opposition aux processus, peut en partie expliquer l'absence de ces critères mais ne la justifie pas complètement. En effet, les rapports peuvent être examinés du point de vue du respect de ces critères.

Constat : les critères de qualité posent des exigences élevées qui n'ont pas été modulées en fonction du type d'évaluation ou de son contexte

Ce constat découle principalement des critères relatifs à la couverture des enjeux (pertinence, réussite, rentabilité et qualité de l'exécution et de la mise en œuvre du programme) par une évaluation. À la lecture des résultats de l'examen, on constate que chaque évaluation examinée par le SCT devait implicitement couvrir tous ces enjeux à la fois. De plus, les critères ont été appliqués de manière non différenciée, c'est-à-dire sans tenir compte de la visée de l'évaluation (formative ou sommative), du type de programme évalué (programme récent ou bien établi), et des fonds engagés dans celui-ci.

Conclusions

L'exigence à l'effet qu'une évaluation doive couvrir tous les enjeux d'évaluation est difficilement justifiable à la fois d'un point de vue théorique et pratique. Certains répondants ont en effet questionné l'application inflexible des critères à l'égard des différents types d'évaluation et de programme. À titre d'exemple, l'évaluation formative est souvent initiée afin de répondre à un besoin interne d'information sur la mise en œuvre et les aspects opérationnels d'un programme. Dans ce cas, il est moins pertinent d'examiner la rentabilité du programme. Même une évaluation sommative peut être de qualité supérieure sans traiter de tous les éléments d'évaluation suggérés par le SCT. En fait, un répondant a affirmé que si la « recette intégrale » en matière de couverture des enjeux était appliquée lors d'une évaluation, la qualité et la quantité des données colligées de même que leur ana- lyse souffriraient à coup sûr de ce choix, entraînant possiblement à la baisse la qualité générale de l'évaluation. En effet, l'évaluation n'est pas réalisée dans l'abstrait : les évaluateurs doivent composer avec plusieurs contraintes (temps, ressources et disponibilité des informations) qui font en sorte qu'il est difficile répondre à toutes les questions parfaitement. En plus d'être discutable, l'exigence à l'effet que tous les enjeux doivent être couverts entre en contradiction ou, à tout le moins, implique un arbitrage, avec le critère de concision du rapport d'évaluation. Il est en outre utopique de satisfaire aux critères de recours à des sources de données multiples et d'équilibre des méthodes qualitatives et quantitatives lors de l'évaluation d'un nouveau programme pour lequel les données sont fragmentaires, voire inexistantes. La taille du programme évalué est aussi un facteur à prendre en compte dans l'application des critères d'évaluation, comme l'a indiqué un répondant. Ainsi, il n'est pas toujours rentable de procéder à une analyse exhaustive de la rentabilité dans le cas de très petits programmes engageant peu de fonds publics, les coûts de l'évaluation dépassant parfois ceux du programme.

En résumé, les exemples précédents témoignent de l'utilisation de critères trop exigeants (*« the gold standard »* selon un répondant) et d'une application insensible au contexte du programme et de l'évaluation. Il semble donc que le portrait de la qualité des évaluations fédérales tracé par le SCT paraisse plus sombre qu'il ne l'est en réalité en ce qui concerne certains critères. Ainsi, s'il est juste de souligner que seulement 29 % des rapports contenaient des conclusions relatives à la rentabilité (SCT, 2004, p. 32), cela ne doit pas nécessairement être interprété comme une lacune des évaluations fédérales. On peut par ailleurs s'interroger sur l'acceptabilité des conclusions de l'examen de qualité auprès des évaluateurs fédéraux étant donné les insatisfactions relatives aux critères et à leur application qui ont été soulevées par certains répondants.

Les conclusions sont-elles fiables, valides, et crédibles?

Constat : le rapport présente clairement les informations relatives à la visée, à l'objet, et au contexte de l'évaluation

Le SCT présente ainsi les objectifs, le contexte, les critères, l'échantillon, l'instrument de mesure, les constatations, les conclusions, de même que plusieurs limites de la méta-évaluation qu'il a réalisée. Le langage utilisé est de plus clair et précis.

Conclusions

L'évaluation doit être présentée et diffusée de manière à permettre à autrui d'interpréter son importance et d'apprécier sa qualité. C'est là un principe fondamental de la démarche scientifique (King, Keohane, & Verba, 1994) mais aussi de la pratique de l'évaluation (American Evaluation Association, 2004). La transparence du SCT doit donc être saluée. Le SCT agit en outre de manière cohérente avec plusieurs critères de qualité appliqués dans le cadre de son examen, notamment ceux ayant trait à la description claire de l'objet d'évaluation et de la méthodologie.

Constat : le devis retenu pour l'examen ne permet pas de déterminer si l'adoption de la Politique d'évaluation de 2001 a eu un impact positif sur la qualité des évaluations fédérales

Étant donné que seuls des rapports produits après l'adoption de la politique ont été examinés, on ne dispose d'aucune mesure de référence antérieure à 2001 avec laquelle on pourrait comparer les résultats de l'examen. Toute analyse de type avant-après de l'impact de la politique sur la qualité des rapports est donc impossible à réaliser à partir des seuls résultats de l'examen.

Conclusions

Une bonne évaluation doit choisir sa méthodologie et ses données en fonction des questions d'évaluation. Le fait que cette limite soit reconnue par le SCT dans son rapport n'enlève rien au caractère inadéquat du devis d'évaluation à cet égard. En ne respectant pas ce principe, le SCT (2004) contrevient à l'un des critères de son propre examen qui exige le recours à une méthodologie adaptée aux éléments d'évaluation et à un devis adapté aux objectifs de l'étude (p. 54). Ce devis permet tout de même de

vérifier l'hypothèse selon laquelle « la qualité des évaluations augmente avec le temps, à mesure que la Politique est mise en œuvre et que les évaluateurs et les responsables du CEE se familiarisent avec elle » (SCT, 2004, p. 10).

Constat : l'échantillon retenu dans le cadre de l'examen a été constitué à partir d'une base de données qui ne contenait que les rapports présentés au SCT par les ministères et organismes fédéraux; cet échantillon contenait 115 rapports, dont 5 n'étaient pas des évaluations

Le SCT (2004) ne disposait pas de la population des rapports produits pendant la période d'étude mais seulement d'un échantillon. Il a donc dû se contenter d'un échantillon de 115 rapports complets contenus dans sa base de données. Alors que le rapport du SCT mentionne que les études spéciales, rapports de vérifications, et autres examens ont été exclus de l'échantillon final parce qu'ils n'étaient pas de véritables évaluations, son annexe méthodologique signale plutôt que cinq des documents examinés (soit 4,3 % des rapports ou 4,7 % des rapports valides) n'étaient **pas** des évaluations mais plutôt des études spéciales ou des documents d'autres types (Ekos, 2004, p. 3).

Conclusions

Le SCT mentionne qu'il est difficile d'établir dans quelle mesure l'échantillon retenu est biaisé. Pour des raisons liées à l'intérêt organisationnel, on peut présumer que les ministères et organismes fédéraux étaient plus susceptibles de faire parvenir au SCT leurs meilleures évaluations. En effet, un rapport d'évaluation de bonne qualité démontre que l'organisation productrice agit conformément aux politiques et règles édictées par le gouvernement. On peut dès lors présumer que la qualité de la **population** des évaluations fédérales est inférieure à celle de l'**échantillon** examiné par le SCT. Ce n'est là qu'une hypothèse, cependant. Étant donné les contraintes importantes auxquelles faisait face le SCT en termes de disponibilité des données, le recours à un échantillon de convenance tel que celui utilisé semble la meilleure option dans les circonstances. En particulier, le nombre et la diversité des évaluations examinées paraît satisfaisant *a priori*. En revanche, l'inclusion de documents qui ne sont pas des évaluations dans l'échantillon constitue un problème. On peut s'interroger sur la pertinence d'apprécier de tels documents dans le cadre d'un examen sur la qualité des évaluations. L'impact de l'inclusion de ces quelques rapports sur les conclusions du SCT est difficile à établir. Une hypothèse plausible serait cependant que ces documents diminuent la qualité moyenne des rapports fédéraux car ils ne répondent pas à plusieurs critères exclusifs aux évaluations tels que la couverture des éléments d'évaluation (pertinence, réussite, rentabilité, etc.).

Constat : l'examen est exclusivement basé sur les données contenues dans les rapports des évaluations fédérales

Aucune autre source de données n'a été utilisée pour mesurer la qualité des rapports d'évaluation fédéraux.

Conclusions

Le recours exclusif aux rapports d'évaluation et à une grille méta-évaluative est une stratégie de recherche intéressante qui permet de générer des données comparables à

propos d'un grand nombre de cas. Il faut toutefois souligner les lacunes de cette méthode au niveau de la disponibilité des données et de la validation des constats et jugements produits. À titre d'exemple, les examinateurs ont été incapables d'évaluer la pertinence de l'analyse et la qualité de la conception de l'évaluation pour 50 % et 23 % des rapports respectivement (SCT, 2004, pp. 22, 30). Cette proportion importante de données manquantes est particulièrement préoccupante à la lumière des résultats empiriques établissant une relation négative entre la qualité méthodologique d'une étude et la proportion d'informations manquantes qu'elle compte (Huwiler-Müntener et al., 2002, cités dans Petticrew & Roberts, 2006, p. 127). L'applicabilité de ces résultats aux rapports examinés par le SCT demeure évidemment une question empirique. Certains répondants ont néanmoins affirmé que le SCT aurait dû contacter les services d'évaluation des organisations concernées, en leur soumettant par exemple l'évaluation détaillée des rapports afin de compléter les informations manquantes et de valider les jugements portés avec les premiers concernés. Aucune procédure de vérification systématique de ce genre n'a toutefois été mise en place dans le cadre de l'examen. Si tel avait été le cas, le SCT serait peut-être arrivé à la conclusion que la qualité des rapports est supérieure à ce qui a été déterminé à travers le seul examen des rapports. Le SCT (2004, p. 12) a d'ailleurs reconnu que le recours exclusif aux rapports comme source de données a pu tirer les résultats de certains rapports vers le bas.

Constat : chaque rapport a été évalué par une seule personne; aucun commentaire n'a été consigné lors de l'évaluation; la fiabilité interjuge a seulement été évaluée de manière qualitative avant la réalisation de l'examen

En raison des contraintes au niveau des ressources, les informations qualitatives relatives au codage de chaque rapport n'ont pas été colligées. De plus, aucune mesure quantitative de fiabilité entre les juges (taux de concordance, kappa de Cohen, corrélation interjuge) n'a été calculée et aucune évaluation concomitante ou rétrospective de la fiabilité n'a été effectuée.

Conclusions

Les individus peuvent faire des erreurs, en particulier, lorsqu'ils doivent coder une multitude d'éléments sur plusieurs rapports. Ils peuvent en outre introduire des biais dans l'évaluation des rapports, notamment en étant trop sévère ou en codant un élément différemment des autres juges. Plusieurs procédures d'évaluation visant à s'assurer de la cohérence du codage à travers le temps et les juges existent. La plus courante consiste à recourir à au moins deux juges pour chaque rapport et à calculer une mesure quantitative de fiabilité. Si l'évaluation qualitative de la fiabilité utilisée dans le cadre de l'examen n'est pas une faiblesse en soi, elle est beaucoup moins précise et transparente que sa contrepartie quantitative. Il est ainsi difficile de déterminer la nature du seuil de fiabilité fixé pour l'examen (e.g., est-il restrictif ou libéral?) et de juger de sa pertinence. En outre, on ne peut présumer que la fiabilité se maintienne durant toute la période de codage, en particulier lorsque le nombre de rapports à coder est élevé (Orwin, 1994). Il aurait été par conséquent souhaitable que la fiabilité interjuge soit également appréciée, même de manière qualitative, **pendant** l'examen afin de s'assurer de sa stabilité, par exemple à travers une contrevérification

de quelques rapports (voir Forss & Carlsson, 1997). Une alternative intéressante aurait consisté à consigner l'information qualitative relative au codage afin de per- mettre la vérification de ce dernier par les fonctionnaires fédéraux concernés. Le SCT n'a toutefois pas retenu cette option étant donné les contraintes de temps pour l'examen de qualité. Dans tous les cas, les conclusions du SCT présentent un niveau de certitude moindre en raison des risques d'erreurs et de biais découlant du codage.

Constat : plusieurs dimensions pertinentes à l'analyse des déterminants de la qualité n'ont pas été considérées dans l'examen du SCT

La méta-évaluation du SCT ne visait pas explicitement à déterminer quels sont les facteurs qui affectent la qualité des rapports. Outre l'année de production du rapport qui était nécessaire à la mesure de l'évolution de la qualité, la taille des organisations (petite, moyenne, ou grande) et le type d'organisation (ministère ou organisme) ayant produit le rapport ont servi à l'analyse croisée des données. Si le type d'évaluation (formatif ou sommatif) a également été considéré lors de la collecte des données (voir Ekos, 2004), cette variable est absente de l'analyse contenue dans le rapport de l'examen.

Conclusions

Certains répondants ont déploré le fait que plusieurs facteurs susceptibles d'exercer une influence sur la qualité des rapports n'aient pas été considérés dans l'examen du SCT : évaluation réalisée à l'interne ou à l'externe, évaluation obligatoire en vertu d'une clause légale ou initiée librement par l'organisation, quantité des ressources investies dans l'évaluation, compétences des évaluateurs, et ainsi de suite. À titre d'exemple, les évaluations externes sont souvent présumées plus rigoureuses et crédibles que les évaluations réalisées à l'interne. La prise de décision au gouvernement fédéral serait-elle renforcée si une plus grande proportion des évaluations était confiée à des consultants du secteur privé? Impossible de répondre à cette question sans une analyse pertinente traitant de cette variable. Selon le répondant du SCT, il semble que certains facteurs supplémentaires aient été considérés mais que la disponibilité pour le moins problématique des données ait coupé court à ce projet, faisant ainsi écho à la conclusion précédente sur les risques de recourir aux rapports comme source exclusive de données. Par ailleurs, ce problème de données manquantes ne s'applique pas à la visée de l'évaluation car avec près de 75 évaluations formatives et sommatives identifiées comme telles dans l'annexe technique (Ekos, 2004, p. 3), il aurait été possible—et très pertinent—d'analyser la relation entre cette variable et la qualité du rapport. Pour ce faire, l'emploi d'un test statistique de différence de moyennes ou de la régression linéaire aurait été approprié. Bref, si les lacunes précédentes n'ont pas empêché le SCT de réaliser un état des lieux acceptable de la qualité des évaluations, les facteurs affectant cette dernière demeurent sous-étudiés. Or, le SCT aurait gagné à non seulement établir un portrait juste de la qualité des rapports mais également à améliorer cette dernière par diverses mesures ciblant les variables identifiées à travers l'examen.

Constat : les conclusions de l'examen sont fondées sur une analyse des données se

résumant à des statistiques descriptives simples et croisées

L'analyse se limite généralement à une présentation de la proportion globale de rapports qui satisfont à chaque élément d'évaluation, d'une part, et à la ventilation des résultats en termes d'année de production, de la taille et du type de l'organisation, d'autre part. L'analyse contenue dans l'annexe technique de l'examen n'est pas davantage approfondie. Les analyses réalisées dans le cadre de l'examen semblent donc se limiter à la présentation de résultats bruts.

Conclusions

Le manque de profondeur des analyses croisées effectuées nuit considérablement à la crédibilité des conclusions de l'examen. Lorsqu'une différence est constatée sur l'une des dimensions retenues pour l'ana- lyse, par exemple que les rapports des grandes organisations sont moins susceptibles (63 %) de contenir des recommandations formelles que ceux des petites et moyennes organisations (89 % et 86 % respectivement; SCT, 2004, p. 41), aucune précision n'est donnée quant à la magnitude de cette différence (i.e., est-ce une différence majeure ou marginale?) ou encore au niveau de certitude associé à ce résultat (i.e., la relation observée est-elle due à la chance?). L'utilisation d'outils statistiques, la régression de type linéaire ou logistique par exemple, aurait permis d'aller au-delà de ces résultats bruts. En négligeant l'analyse des données, le SCT (2004) contrevient à l'un de ses propres critères exigeant que « les données soutiennent l'analyse (selon, par exemple, les **tests de signification** et les taux de réponse) » (p. 60, caractères gras ajoutés).

L'analyse qui sous-tend la conclusion suivante, fondamentale au regard des objectifs de l'examen, soulève par ailleurs des doutes importants quant à sa crédibilité :

> Une comparaison des rapports élaborés avant avril 2002 et de ceux élaborés par la suite démontre toutefois une amélioration de la qualité concernant un certain nombre de critères dans les évaluations les plus récentes. […] Cette amélioration de la qualité avec le temps laisse croire que les efforts du SCT pour améliorer la qualité des évaluations ont peut-être un effet positif, en ayant accordé une année, jusqu'en avril 2002, aux ministères et aux organismes pour comprendre entièrement la Poli- tique et pour donner le temps au Centre d'excellence en évaluation de commencer à fonctionner. (SCT, 2004, p. 2)

Le SCT (2004) soutient ainsi que les rapports produits après avril 2002 ont en général des notes plus élevées que ceux produits avant cette date. La présentation d'une douzaine de critères où une amélioration a été constatée vient appuyer cette conclusion. Or, cette présentation est incomplète et biaisée. En effet, seule une fraction des résultats est présentée lors de la discussion de la dimension temporelle—tous positifs de surcroît—ce qui laisse croire que la qualité s'est davantage améliorée que ce n'est le cas en réalité (SCT, 2004, pp. 41–43). Cela ne signifie pas nécessairement que la qualité des rapports ne s'est pas améliorée d'une période à l'autre mais seulement que la démonstration sur laquelle cette conclusion est fondée est loin d'être convaincante.

Afin de vérifier la conclusion à l'effet qu'il y a eu amélioration de la qualité des évaluations, les notes de qualité globale attribuées aux rapports produits avant et après avril 2002 ont été comparées. Selon ce critère, la proportion de rapports adéquats ou plus qu'adéquats est passée de 67,5 % (n = 37) à 82,1 % (n = 78) (Ekos, 2004, p. 206). Puisque nous avons des raisons de croire que la qualité s'est améliorée suite aux efforts du SCT, un test d'hypothèse unilatéral portant sur la différence entre ces proportions a été effectué. Le résultat du test (z = 1,73; p < 0,05) suggère que la qualité des évaluations fédérales s'est effectivement améliorée après avril 2002. Bien qu'on compte un peu plus de 90 autres critères d'évaluation qui permettent le recours à ce genre d'outils statistiques, seule la qualité globale a été examinée en raison de son importance fondamentale comme critère d'évaluation (en comparaison, la présence d'un résumé n'est qu'un indicateur secondaire de la qualité). À la lumière du résultat précédent, l'hypothèse selon laquelle la qualité des rapports s'est améliorée depuis l'adoption de la politique d'évaluation en 2001 semble très plausible. Le rapport de l'examen du SCT ne contient cependant aucune démonstration satisfaisante à l'appui de cette conclusion. Cela constitue sans contredit une lacune majeure.

L'exercice méta-évaluatif est-il pertinent et utile?

Constat : avant l'examen, aucun portrait exhaustif et à jour de la qualité des rapports d'évaluation fédéraux n'était disponible

Les derniers examens à grande échelle des rapports d'évaluation remontaient au début des années 90 et avaient été réalisés par le Bureau du Vérificateur général et le Bureau du Contrôleur général du Canada (Jacob, 2006). Avec sa politique d'évaluation de 2001, le SCT a renouvelé son engagement à l'égard de l'outil de gestion important que constitue l'évaluation (cet engagement a été renouvelé une fois de plus en 2009 avec l'adoption d'une nouvelle politique d'évaluation). L'examen méta-évaluatif a permis d'une part d'envoyer un signal clair à cet égard et d'autre part de réaliser le suivi de la mise en œuvre de la politique en établissant un portrait de la qualité des rapports.

Conclusions

L'exercice méta-évaluatif du SCT semble répondre à des besoins informationnels réels de surveillance de la qualité des évaluations au gouvernement fédéral, surtout en l'absence d'un portrait complet et à jour de la qualité des évaluations. Un tel exercice établit une mesure de référence permettant d'une part de juger l'état actuel de la qualité des rapports et d'autre part de suivre son évolution. Cette fonction de surveillance, utile en soi, constitue en outre la pierre d'assise de toute démarche d'amélioration de la qualité des évaluations. Comme l'a mentionné un répondant : « À partir du moment où l'on se mesure, on s'améliore ». L'envergure de l'examen aurait notamment sensibilisé évaluateurs et gestionnaires à l'importance de l'évaluation en général et à la qualité des rapports d'évaluation en particulier. Un répondant a par exemple constaté que le statut de l'évaluation par rapport à la vérification s'était trouvé renforcé au sein de son organisation suite à l'examen.

En mettant au jour les faiblesses précises des évaluations, l'examen du SCT a par

ailleurs fourni des pistes pour améliorer la qualité des pratiques évaluatives. Les entretiens réalisés indiquent que ce diagnostic a été jugé utile par plusieurs répondants. Outre le portrait général de la qualité tracé par l'examen du SCT, certains répondants ont d'ailleurs mentionné avoir obtenu des informations relatives aux forces et faiblesses particulières à leur organisation ce qui leur a permis de mieux cibler leurs efforts d'amélioration de la qualité évaluative. Il a été par ailleurs souligné que l'examen a contribué à clarifier les attentes du SCT en matière de qualité des évaluations. Si la *Politique d'évaluation* (SCT, 2001) énonce des normes d'évaluation dont les évaluateurs pouvaient s'inspirer avant l'examen, celles-ci étaient incomplètes et formulées en termes trop généraux au goût de certains répondants. Les critères utilisés dans le cadre de l'examen ont donc permis de préciser les attentes spécifiques du SCT concernant la qualité des rapports. Des répondants ont également mentionné avoir recours à la grille d'examen du SCT dans une perspective d'autoévaluation ce qui constitue une procédure simple mais efficace de contrôle de la qualité des évaluations à l'interne (Scriven, 2005b). Un consensus semble se dégager chez les personnes interviewées à l'effet que l'examen de qualité a, de manière générale, été relativement bien reçu par les évaluateurs fédéraux. Certes, certains répondants ont exprimé des réserves quant aux critères utilisés, notamment le fait que plusieurs d'entre eux étaient inconnus des évaluateurs avant l'examen, alors qu'un autre a observé une résistance au changement chez certains collègues en poste depuis longtemps. L'adhésion des parties prenantes à l'examen de qualité semble néanmoins satisfaisante de manière générale. En somme, malgré leur caractère fragmentaire, les données précédentes sur l'adhésion laissent croire que l'examen du SCT est pertinent du point de vue de certains fonctionnaires fédéraux impliqués en évaluation.

Discussion des résultats

Avant de revenir sur les enseignements que l'on peut tirer de l'examen, il convient tout d'abord de nuancer les conclusions présentées jusqu'ici. D'une part, certaines conclusions, particulièrement celles ayant trait à la pertinence et à l'utilité de l'examen, reposent sur des données d'entretiens effectués auprès de quelques répondants qui ne sont pas nécessairement généralisables à l'ensemble des fonctionnaires fédéraux œuvrant en évaluation. Ces conclusions jouent donc d'abord un rôle exploratoire : elles suggèrent des pistes d'analyse qui devront être corroborées lors de recherches ultérieures. D'autre part, la qualité de toute évaluation devrait être appréciée à l'aune des contraintes humaines, légales, organisationnelles, et matérielles qui la caractérisent. L'examen du SCT ne fait pas exception à cette règle : les ressources mobilisées dans le cadre de cette initiative sont limitées. Même imparfaite, une méta-évaluation faisable et abordable qui offre en temps opportun des conclusions d'un niveau de crédibilité correct vaut souvent mieux qu'une méta-évaluation « parfaite » qui mobilise une part indue des ressources organisationnelles ou dont les résultats arrivent trop tard. Cela étant dit, même si la qualité absolue de l'évaluation ne garantit pas l'utilisation, elle y contribue certainement.

En définitive, l'examen de la qualité des évaluations fédérales est-il une méta-évaluation réussie? Cette question appelle une réponse nuancée. D'un côté, l'examen

du SCT est transparent quant à sa dé- marche d'évaluation, dresse un portrait acceptable quoiqu'imparfait de la qualité des évaluations fédérales qui semble répondre à des besoins informationnels réels. À terme, les résultats de cette étude laissent croire que l'examen pourrait contribuer à améliorer les pratiques d'évaluation au gouvernement fédéral et ainsi exercer un impact positif sur la qualité des rapports. Quoiqu'il en soit, l'ampleur et la portée de l'examen en font un exemple à suivre pour les évaluateurs et responsables des services d'évaluation des administrations publiques du monde entier. Il faut d'un autre côté rappeler les principales lacunes dont souffre cet exercice : les critères de qualité retenus pour l'examen sont dans certains cas trop exigeants et appliqués de manière trop rigide, aucune procédure systématique de vérification des erreurs et des biais n'a été appliquée pendant ni après l'examen, le devis d'évaluation retenu ne permet pas de déterminer s'il y a eu amélioration de la qualité suite à l'adoption de la Politique, et les analyses effectuées manquent de profondeur. Ces faiblesses sont sérieuses et entachent la crédibilité des conclusions d'un rapport qui est par ailleurs de qualité acceptable.

Recommandations à l'intention des évaluateurs et des responsables des services d'évaluation

La mise au jour des lacunes de l'examen du SCT n'est pas une fin en soi : il est souhaitable d'en tirer des enseignements. Ceux-ci sont présentés sous la forme de recommandations s'adressant aux évaluateurs et responsables des services d'évaluation du gouvernement fédéral ainsi qu'aux « gardiens de la qualité » des administrations publiques du Canada et d'ailleurs. Ces recommandations visent la réalisation d'une méta-évaluation d'une envergure similaire à celle de l'examen du SCT mais qui ne souffrirait pas des mêmes lacunes. Les quatre recommandations proposées reviennent sur plusieurs des thèmes abordés au dans cet article :

1. *Bien définir la conception théorique de la qualité de l'évaluation retenue et en déduire des indicateurs de mesure. Appliquer ces critères de manière appropriée et flexible, c'est-à-dire en prenant en compte le type, l'objet, et le contexte d'évaluation.* D'une part, la conception de la qualité retenue doit être complète, cohérente, et appropriée à l'objet de la méta-évaluation. D'autre part, les critères d'évaluation doivent découler de cette conception et en couvrir toutes les dimensions. En ce qui a trait à leur application, les critères d'évaluation doivent être adaptés au type d'évaluation (for- matif ou sommatif), à l'objet d'évaluation (projet pilote de portée limitée ou politique transversale à haute visibilité), et au contexte de l'évaluation (capacité organisationnelle, disponibilité des données, etc.). Lorsque des évaluations diverses sont évaluées au cours d'un même exercice, il peut être opportun d'avoir recours à plus d'une grille d'évaluation ou encore à une grille à « géométrie variable », c'est-à-dire un instrument qui applique des critères différenciés selon le type d'évaluation, son objet, et son contexte.

2. *Mettre en œuvre une stratégie de maximisation de la validité et de la fiabilité de la mesure de la qualité des évaluations; éviter le recours aux rapports d'évaluation comme source exclusive de données*. Afin que les résultats d'une méta-évaluation soient crédibles,

la mesure de la qualité doit être stable à travers les juges et le temps. Cela implique d'apprécier la fiabilité interjuge à différents moments de la méta-évaluation et de corriger celle-ci au besoin. En outre, les juges devraient, dans la mesure du possible, consigner les informations et justifications relatives au codage de manière à permettre une appréciation du codage par autrui. Ensuite, si le recours aux rapports d'évaluation comme source de données n'est pas un problème en soi, il faut cependant veiller à ce que les informations qu'ils contiennent soient complétées par des données, notamment sur le contexte d'évaluation. Pour ce faire, les documents administratifs peuvent être utiles mais des échanges avec les principaux intéressés, soit les évaluateurs et gestionnaires des organisations fédérales ayant produit ces rapports, s'avèrent précieux. Une procédure relativement simple et économique de triangulation consiste à soumettre les conclusions de la méta-évaluation aux principaux intéressés afin qu'ils vérifient celles-ci et corrigent les erreurs de fait et d'interprétation. La logique sous-tendant les conclusions doit par conséquent être explicite, ce qui nécessite un codage transparent tel que recommandé précédemment.

3. *Le devis doit être approprié aux objectifs de la méta-évaluation; les analyses doivent être suffisamment approfondies pour supporter les conclusions de la méta-évaluation.* La méta- évaluation est une évaluation d'un type particulier où l'objet est une ou plusieurs évaluations. Elle vise par conséquent à produire un jugement de valeur, certes, mais repose par ailleurs sur une « démarche méthodologique, transparente et reproductible » (Jacob, 2004, p. 203). Cela signifie que la conception de la méta-évaluation doit être adaptée à ses questions; par exemple, pour déterminer si la qualité des rapports d'évaluations s'est améliorée en raison de l'adoption d'une politique, il faut évaluer des rapports produits avant et après cette politique. Ensuite, l'analyse descriptive et explicative des données doit permettre d'étayer les constats formulés. Lorsque le nombre de cas est suffisamment élevé, le recours aux outils statistiques peut s'avérer utile. Ces outils, divers types de régression statistique notamment, permettent en effet de déterminer si la relation observée (e.g., l'amélioration de la qualité) est due à la chance ou non, et si cette association est marginale, faible, modérée, ou élevée. Les faits parlant rarement pour eux-mêmes, il est nécessaire de les analyser et les interpréter afin d'en tirer le maximum d'informations.

4. *L'implication des différentes parties prenantes durant l'en- semble de la démarche peut contribuer à améliorer la crédibilité et l'acceptabilité des conclusions de la méta-évaluation.* L'implication des parties prenantes via un groupe de travail aux premières étapes de la méta-évaluation telles que la définition de la portée de l'examen, la sélection des critères d'évaluation, et le développement de l'instrument de mesure représente à cet égard une pratique intéressante mais d'une portée limitée. Une consultation, même limitée, des évaluateurs et responsables des services d'évaluation sur les critères et la méthodologie de la méta-évaluation permet d'apporter des correctifs avant et pendant la réalisation de l'examen. La vérification du codage, en totalité ou en partie, par des parties prenantes

permet également de corriger les erreurs et d'accroître le niveau de crédibilité des conclusions. Les principaux intéressés à une méta-évaluation sont d'ailleurs plus susceptibles de s'approprier ses conclusions lorsqu'ils sont associés à sa planification et à sa réalisation (Cousins, 2003). Plusieurs lacunes de l'examen du SCT ont d'ailleurs été identifiées suite à des entretiens effectués auprès d'évaluateurs fédéraux. Par exemple, certains répondants ont dénoncé l'application inflexible aux rapports de critères d'évaluation trop exigeants. Il aurait été possible de corriger cette lacune si le SCT avait consulté davantage les évaluateurs fédéraux préalablement à et pendant son examen.

Quelques pistes pour des recherches futures

Il apparaît maintenant opportun de suggérer quelques pistes pour la réalisation de recherches ultérieures sur les impacts de cet examen et des efforts généraux du SCT afin de promouvoir la qualité des évaluations fédérales. Cette évaluation n'a pas permis de déterminer si la qualité des rapports s'est améliorée suite à l'examen du SCT, et ce même si ses résultats (portrait des forces et faiblesses des rapports fédéraux, accroissement de la crédibilité de l'évaluation, spécification des attentes du SCT en matière de qualité, etc.) laissent croire que cela pourrait être le cas à terme. Une première piste de recherche consisterait donc à étudier de manière systématique l'utilisation de l'examen de qualité par les décideurs et évaluateurs du gouvernement fédéral ainsi que son impact sur la qualité des rapports. Quelles décisions prises par les gestionnaires du SCT découlent directement de cet examen? Quant aux responsables des services d'évaluation et évaluateurs, ont-ils adaptés leurs pratiques afin de répondre aux exigences du SCT? À cet égard, quelles sont les principales améliorations observées? Doit-on attribuer ces améliorations exclusivement à l'examen, à d'autres mesures gouvernementales mises en place pour assurer la qualité des rapports, ou à des facteurs exogènes?

Un second chantier de recherches a trait aux effets de l'examen sur l'institutionnalisation de l'évaluation au gouvernement fédéral. Varone et Jacob (2004) définissent l'institutionnalisation comme la « routinisation » du recours attendu ou obligé à l'évaluation et donc de sa pratique effective au sein des organisations. Les données colligées lors des entretiens suggèrent que l'examen aurait eu un impact positif à cet égard. L'attention portée à la qualité des évaluations semble avoir renforcé la crédibilité de l'évaluation comme outil de pilotage de l'action publique, en particulier auprès des gestionnaires des ministères et organismes fédéraux. En effet, la « menace » d'un contrôle de qualité par une agence centrale incite les gestionnaires à s'intéresser davantage à la qualité des évaluations produites par leur organisation. L'envergure de l'examen pourrait de plus être interprétée comme un signal à l'effet que la qualité des évaluations est importante dans l'administration fédérale.

Une dernière piste de recherche concerne les impacts négatifs ou dysfonctionnements que l'examen pourrait avoir engendrés, soit le gaspillage des ressources, l'adhésion purement rituelle et symbolique aux règles et procédures visant à contrôler la qualité (i.e., la dissociation), et l'incrustation excessive des valeurs et pratiques d'assurance qualité au cœur des pratiques organisationnelles qui nuit à l'innovation et amène l'organisation à privilégier ce qui est mesurable à ce qui l'est moins

(i.e., la colonisation) (Schwartz & Mayne, 2005a; 2005b).

En définitive, une évaluation peut, malgré ses lacunes, exercer un impact sur la gestion publique. L'examen du SCT pourrait potentiellement représenter une illustration de cette affirmation. Mais, même si elle demeure un enjeu fondamental, la qualité constitue un moyen qui ne doit pas faire perdre de vue la fin, soit l'amélioration des programmes publics et, ultimement, des conditions sociales.

Remerciements

Cet article est en partie basé sur un essai rédigé dans le cadre d'une maîtrise en analyse des politiques complétée à Université Laval. Une communication portant sur une version antérieure de cette étude a été présentée le 3 novembre, 2006, à Québec, lors du 15ième colloque annuel de la Société québécoise d'évaluation de programme (SQEP). L'auteur tient sincèrement à remercier les sept répondants ayant accepté de participer à cette étude de même que Lisa Birch, Benoît Collette, Nouhoun Diallo, Mbaïrewaye Mbaï-Hadji, et Steve Jacob pour leurs commentaires sur une version antérieure de ce texte. L'auteur remercie également Jérôme Couture, François Piette, et François Gélineau pour leurs conseils en statistique. Toute lacune demeure évidemment de l'entière responsabilité de l'auteur.

Références

Addison, E., & Amo, C.F. (2005). Two decades of the *Canadian Journal of Program Evaluation*: A content analysis. *Revue canadienne d'évaluation de programme, 20*(3), 17–40.

Albæk, E. (1995). Between knowledge and power: Utilization of social science in public policy making. *Policy Sciences, 28*(1), 79–100.

Albæk, E. (1996). Why all this evaluation? Theoretical notes and empirical observations on the functions and growth of evaluation, with Denmark as an illustrative case. *Revue canadienne d'évaluation de programme, 11*(2), 1–34.

Alkin, M.C., Daillak, R., & White, P. (1979). *Using evaluations: Does evaluation make a difference?* Beverly Hills, CA: Sage.

American Evaluation Association. (2004). *American Evaluation Association guiding principles for evaluators*. Repéré le 12 avril, 2008, de <http:// www.eval.org/GPTraining/GP%20Training%20Final/gp.principles. pdf>

Beaud, J.-P. (2003). L'échantillonnage. Dans B. Gauthier (Éd.), *Recherche sociale : de la problématique à la collecte des données* (quatriéme éd., pp. 211–242). Québec : Presses de l'Université du Québec.

Birch, L.M., & Jacob, S. (2005). Program evaluation in Canada seen through the articles published in CJPE. *Revue canadienne d'évaluation de programme, 20*(3), 1–16.

Bustelo, M. (s.d.). *Metaevaluation as a tool for the improvement and development of the evaluation function in public administrations*. Communication présentée à la conférence de 2002 de l'European Evaluation Society. Repéré le 4 décembre, 2005, de <http://www.europeaneva- luation.org/docs/BusteloSevillaEES.pdf>

Caplan, N. (1977). A minimal set of conditions necessary for the utilization of social science knowledge in policy formulation at the national level. Dans C.H. Weiss (Éd.), *Using social research in public policy making* (pp. 183–197). Lexington, MA: Lexington Books.

Christie, C. A. (2007). Reported influence of evaluation data on decision makers' actions: An empirical examination. *American Journal of Evaluation, 28*(1), 8–25.

Christie, C.A., & Alkin, M.C. (1999). Further reflections on evaluation mis- utilization. *Studies in Educational Evaluation, 25*, 1–10.

Cook, T.D., & Gruder, C.L. (1978). Metaevaluation research. *Evaluation Review, 2*(1), 5–51.

Cooksy, L.J., & Caracelli, V.J. (2005). Quality, context, and use: Issues in achieving the goals of metaevaluation. *American Journal of Evalua- tion, 26*(1), 31–42.

Cooksy, L.J., & Caracelli, V.J. (2007, novembre). The practice of metaevaluation : Does evaluation practice measure up? Communication présentée à la 21ième conférence annuelle de l'American Evaluation Association sous le thème « Evaluation 2007: Evaluation and Learning ». Baltimore, MD.

Cousins, J.B. (2003). Utilization effects of participatory evaluation. Dans

T. Kellaghan & D.L. Stufflebeam (Éds.), *International handbook of educational evaluation* (pp. 245–265). Dordrecht : Kluwer Academic.

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

Cousins, J.B., & Shulha, L.M. (2006). A comparative analysis of evaluation utilization and its cognate field of inquiry: Current issues and trends. Dans I.F. Shaw, J.C. Greene, & M.M. Mark (Éds.), *Handbook of evaluation: Policies, programs and practices* (pp. 266–291). Thousand Oaks, CA : Sage.

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

Ekos Research Associates. (2004). *Review of the quality of evaluations across departments and agencies: Technical appendix supporting tables of results*. Ottawa : Auteur (préparé pour le Secrétariat du Conseil du Trésor, Centre d'excellence en évaluation).

Forss, K., & Carlsson, J. (1997). The quest for quality—Or can evaluation findings be trusted? *Evaluation, 3*(4), 481–501.

Henry, G. T. (2003). Influential evaluations. *American Journal of Evaluation, 24*(4), 515–524.

Huie Hofstetter, C., & Alkin, M.C. (2003). Evaluation use revisited. Dans

T. Kellaghan & D.L. Stufflebeam (Éds.), *International handbook of educational evaluation* (pp. 197–222). Dordrecht : Kluwer Academic.

Jacob, S. (2004). Évaluation. Dans L. Boussaguet, S. Jacquot, & P. Ravinet (Éds.), *Dictionnaire des politiques publiques* (pp. 201–208). Paris : Presses de la Fondation Nationale des Sciences Politiques.

Jacob, S. (2006). Trente ans d'évaluation de programme au Canada : l'institutionnalisation interne en quête de qualité. *Revue française d'administration publique, 119*, 515–532.

Joint Committee on Standards for Educational Evaluation. (1994). *The Pro- gram Evaluation Standards: How to assess evaluations of educational programs* (deuxième éd.). Thousand Oaks, CA : Sage.

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

King, J.A. (2005). Participatory evaluation. Dans S. Mathison (Éd.), *Encyclopedia of evaluation* (pp. 291–294). Thousand Oaks, CA : Sage.

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

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

Mathison, S. (Éd.). (2004). Triangulation. Dans *Encyclopedia of evaluation*. Repéré le 17 mars,

2010, de <http://www.sage-ereference.com/evalua- tion/Article_n555.html>
Müller-Clemm, W.J., & Barnes, M.P. (1997). A historical perspective on federal program evaluation in Canada. *Revue canadienne d'évaluation de programme, 12*(1), 47–70.
Oakley, A. (2002). Social science and evidence-based everything: The case of education. *Educational Review, 54*(3), 277–286.
Orwin, R.G. (1994). Evaluating coding decisions. Dans H. Cooper & L.V. Hedges (Éds.), *Handbook of research synthesis* (pp. 139–162). New York : Russell Sage Foundation.
Patton, M.Q. (1997). *Utilization-focused evaluation: The new century text*
(troisième éd.). Thousand Oaks, CA: Sage.
Patton, M.Q. (2001). Use as a criterion of quality in evaluation. Dans
A.P. Benson, D.M. Hinn & C. Lloyd (Éds.), *Visions of quality: How evaluators define, understand and represent program quality* (*7*, pp. 155–180). New-York : Elsevier.
Patton, M.Q. (2004). Misuse of evaluations. Dans S. Mathison (Éd.), *Encyclopedia of evaluation*. Repéré le 17 mars, 2010, de <http://www. sage-ereference.com/evaluation/Article_n344.html>.
Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Malden, MA : Blackwell.
Rich, R.F., & Oh, C.H. (2000). Rationality and use of information in policy decisions: A search for alternatives. *Science Communication, 22*(2), 173–211.
Schwandt, T.A. (2008). The relevance of practical knowledge traditions to evaluation practice. Dans N.L. Smith & P.R. Brandon (Éds.), *Fundamental issues in evaluation* (pp. 29–40). New York : Guilford.
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. (Éds.). (2005b). *Quality matters: Seeking confidence in evaluating, auditing, and performance reporting*. New Brunswick, NJ: Transaction.
Scriven, M. (2005a). Checklists. Dans S. Mathison (Éd.), *Encyclopedia of evaluation* (pp. 53–59). Thousand Oaks, CA: Sage.
Scriven, M. (2005b). Metaevaluation. Dans S. Mathison (Éd.), *Encyclopedia of evaluation* (pp. 53–59). Thousand Oaks, CA : Sage.
Secrétariat du Conseil du Trésor. (2001). *Politique d'évaluation*. Ottawa : Auteur. Repéré le 15 novembre, 2008, de <http://www.tbs-sct.gc.ca/pol/doc-fra.aspx?id=12309&section=HTML>
Secrétariat du Conseil du Trésor. (2004). *Examen de la qualité des évalua- tions dans les ministères et les organismes*. Ottawa : Auteur. Repéré le 15 novembre, 2008, de < http://www.tbs-sct.gc.ca/eval/pubs/rqeda- eqemo_f.pdf>
Secrétariat du Conseil du Trésor. (2004). *Guide pour l'examen des rap- ports d'évaluation*. Ottawa : Centre d'excellence en évaluation. Repéré le 15 novembre, 2008, de <http://www.tbs-sct.gc.ca/eval/ tools_outils/4001752_f.pdf>
Secrétariat du Conseil du Trésor. (2005). *La santé de la fonction d'évaluation au gouvernement du Canada : Rapport pour l'exercice 2004-2005*. Ottawa : Auteur. Repéré le 15 novembre, 2008, de <http://www.tbs-sct. gc.ca/eval/dev/health-santé/pdf/hefgc-sfegc_f.pdf>
Segsworth, R.V. (2005). Program evaluation in the government of Canada: Plus ça change ... *Revue canadienne d'évaluation de programme, 20*(3), 175–197.
Stufflebeam, D.L. (2001a). Evaluation checklists: Practical tools for guiding and judging evaluations. *American Journal of Evaluation, 22*(1), 71–79.
Stufflebeam, D.L. (2001b). The metaevaluation imperative. *American Journal of Evaluation, 22*(2), 183–209.

Toulemonde, J. (2007). L'appropriation des résultats de l'évaluation. Leçons de la pratique en Région Limousin. Dans S. Jacob, F. Varob & J.-L. Genard (Éds.), *L'évaluation des politiques au niveau régional* (pp. 131–142). Bruxelles : P.I.E. Peter Lang.

Uusikylä, P., & Virtanen, P. (2000). Meta-evaluation as a tool for learning: A case study of the European structural fund evaluations in Finland. *Evaluation, 6*(1), 50–65.

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.

Weiss, C.H., & Bucuvalas, M.J. (1980). Truth tests and utility tests : Decision-makers' frame of reference for social science research. *American Sociological Review, 45*(2), 302–313.

Widmer, T., Landert, C., & Bachman, N. (2000). Standards d'évaluation de la Société suisse d'évaluation (Standards SEVAL). Repéré le 15 avril, 2007, de <http://www.seval.ch/fr/documents/SEVAL_ Standards_2001_fr.pdf>

Worthen, B.R. (2001). Whither evaluation? That all depends. *American Journal of Evaluation, 22*(3), 409–418.

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

Annexe A. Dimensions abordées lors des entretiens

- Familiarité avec l'examen du SCT
- Perceptions générales face à l'examen
- Réception et accueil par les collègues
- Pertinence de l'examen
- Motifs de réalisation de l'examen
- Valeur des critères d'évaluation de l'examen
- Valeur de la méthodologie de l'examen (sources de données, codage, analyse, etc.)
- Impacts de l'examen