ÉCOLE NATIONALE D'ADMINISTRATION PUBLIQUE

LES EFFETS DE LA STANDARDISATION, DE LA NORMALISATION ET DE LA PONDÉRATION DES INDICATEURS SUR LA ROBUSTESSE D'UNE COTE GLOBALE: LE CAS DE L'ÉVALUATION SOMMATIVE DE LA PERFORMANCE DES ÉCOLES

MÉMOIRE PRÉSENTÉ COMME EXIGENCE PARTIELLE DE LA MAÎTRISE EN MESURE ET ÉVALUATION DE L'INTERVENTION PUBLIQUE

PAR
SYLVAIN BERNIER

SEPTEMBRE 2002

REMERCIEMENTS

Je remercie sincèrement Richard Marceau, mon directeur de mémoire, pour son implication sans réserves et pour m'avoir donné l'opportunité de l'assister dans ses recherches. Je tiens également à remercier Michel Boucher pour nos nombreuses discussions de couloir, Natalie Rinfret pour sa rigueur et la pertinence de ces commentaires ainsi que l'ÉNAP pour son soutien financier. Finalement, je ne pourrais passer sous silence les encouragements et le support moral de ma compagne de vie, Karine, et de ma famille.

RÉSUMÉ

Ce mémoire compare les fondements épistémologiques et méthodologiques des différentes évaluations de la performance des écoles secondaires au Québec et présente des résultats empiriques de l'effet de l'adoption de diverses pratiques méthodologiques. Les principales divergences méthodologiques concernent le nombre d'indicateurs utilisés pour évaluer la performance des écoles, la méthode d'agrégation des indicateurs, le recours à une échelle de mesure commune ainsi qu'à la pondération des indicateurs de performance. Les effets de différentes pratiques évaluatives sur la cote globale des écoles, leur classement ainsi que sur la robustesse de ceux-ci ont été vérifiés. Les résultats montrent qu'il est préférable qu'une évaluation de la mesure de la performance des écoles : 1) comprenne plus d'un indicateur, 2)utilise une échelle de mesure standardisée et 3) pondère les indicateurs utilisés par la composite de manière à donner plus d'importance aux indicateurs fortement corrélés.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

Au Québec, comme dans bon nombre de pays industrialisés, la performance du système national d'éducation fait l'objet d'une attention croissante (OCDE, 1995). Les écoles sont désormais sous les feux de la rampe et doivent relever un défi important : être davantage responsables de leur propre performance. Ainsi, au Québec, la performance des écoles est maintenant évaluée annuellement par le Ministère de l'Éducation du Québec (MEQ) ainsi que par *Le Bulletin des écoles secondaires du Québec* (*Bulletin*). S'il y a une tendance forte sur l'importance de mesurer la performance des écoles, il en va tout autrement du choix des indicateurs, de leur traitement statistique et de la manière dont ils doivent être présentés aux diverses parties concernées.

Les divergences méthodologiques propre à l'évaluation de la performance des écoles secondaires du Québec sont les suivantes : le nombre d'indicateurs, l'échelle utilisée pour mesurer et comparer la performance des écoles et la pondération des indicateurs de performance. Le MEQ présente les résultats moyens des écoles aux examens de fin de secondaire et utilise le taux de réussite moyen à ces examens pour classer les écoles secondaires du Québec. Le *Bulletin* utilise les résultats moyens des écoles aux examens de fin de secondaire, les standardise, les agrège et les pondère pour produire une cote globale pour chaque établissement. Ensuite, un classement des écoles secondaires du Québec est constitué à partir de ces cotes globales.

Ce portrait met en évidence le fossé qui sépare les méthodes utilisées pour évaluer la performance des écoles. Nous sommes en présence d'un différent méthodologique qui risque de fausser les résultats obtenus par les écoles et de miner l'importance des efforts et ressources investis dans l'élaboration de mesure de la performance des écoles.

Dans un rapport présenté à la conférence annuelle de l'American Educational Research Association, Stevens, Estrada et Parkes (2000) énumèrent les principaux enjeux liés à la mesure lors de l'établissement d'un système d'imputabilité pour les écoles. Ils identifient cinq champs importants : les instruments de mesure et les indicateurs choisis, le design et l'imputabilité, l'utilisation de scores composites et l'agrégation de résultats, le niveau et l'unité d'analyse et le recours aux comparaisons longitudinales et transversales.

Dans la première section sur les instruments de mesure et les indicateurs choisis, les auteurs citent les travaux de Fitz-Gibbon et Kochan (2000) qui portent sur le type d'indicateurs qui sont utilisés dans les différents systèmes d'imputabilité à travers le monde. On y retrouve également des critères de stabilité, de validité et de fiabilité qui doivent être considérés lors de l'adoption d'un nouvel indicateur tel que décrit par Mandeville et Anderson (1987).

Dans la section portant sur le design et l'imputabilité, les auteurs discutent de l'utilisation de tests standardisés, de la raison d'être des évaluations de la performance des écoles (CCSSO, 1999) ainsi que de l'importance d'évaluer les effets de l'implantation d'un système d'imputabilité (Messick, 1989, 1994).

Dans la troisième section sur l'utilisation des scores composites d'agrégation de résultats, ils traitent de méthodes d'analyse de données multiples (Schmidt et Kaplan, 1971), et de l'importance d'utiliser une échelle de mesure commune (Stevens et Aleamoni, 1986).

La quatrième section porte sur le niveau d'analyse et discute de l'état actuel de cette question. On y parle non seulement du problème inhérent à l'utilisation de mesures de la performance des étudiants dans l'évaluation de la performance des écoles, mais aussi des solutions proposées par divers chercheurs dont Gray, Jesson, Goldstein, Hedger et Rashbash (1995). Pour terminer, les auteurs abordent le

problème de l'estimation de la variabilité à divers niveaux hiérarchiques (Goldstein, 1995).

Finalement, la cinquième section discute des comparaisons longitudinales et transversales et recense les nombreux écrits qui discutent de l'efficacité, des avantages et des inconvénients liés à l'utilisation de cohortes réelles et fictives dans l'évaluation de la performance des écoles. En plus de cette discussion théorique, ils s'attardent aussi aux différentes méthodes statistiques utilisées dans le cadre d'évaluations longitudinales.

De ces cinq thèmes, la composite et l'agrégation de résultats est celui qui est le plus susceptible de nous aider à comprendre l'enjeu évaluatif puisqu'il devrait nous aider à éclaircir deux des trois grandes divergences méthodologiques entourant l'évaluation de la performance des écoles québécoises. Bien que l'étude des autres thèmes soit souhaitable, ils ne feront pas l'objet d'une étude approfondie pour le moment puisqu'ils ne sont pas à l'origine des différences que l'on peut observer entre les deux types d'évaluations utilisées au Québec.

Dans la section concernant la composite et l'agrégation de résultats, le rapport de Stevens, Estrada et Parkes (2000) cite des auteurs dont les recherches portent sur des domaines d'études autres que celui de la mesure de la performance scolaire (Stevens et Aleamoni, 1986 et Schmidt et Kaplan, 1971), alors que les autres sections réfèrent à des études qui lui sont directement liées. En poussant la recherche, on découvre non seulement que la théorie entourant la mesure, l'agrégation et la pondération des indicateurs servant à l'évaluation de la performance des écoles est quasi inexistante au Québec, mais que ça semble aussi être le cas dans les autres

pays[1] qui se sont penchés sur la question de l'évaluation de la performance des écoles jusqu'à présent.

Les écrits concernant la mesure de la performance des écoles sont rares. Cette discipline regroupe néanmoins des aspects empruntés à d'autres disciplines connues comme l'évaluation de programmes publics ou celle de la performance académique au sens large. Il faudra donc s'inspirer, comme le font Stevens, Estrada et Parkes (2000), des travaux effectués dans ces disciplines connexes pour tenter de mettre fin à la controverse entourant l'échelle de mesure, l'agrégation de résultats et la pondération des indicateurs de performance des écoles secondaires du Québec. Ainsi, l'objectif du présent mémoire consiste à apporter des pistes de solution théorique au problème de standardisation des pratiques d'évaluation de la performance des écoles du Québec et de démontrer l'effet des changements proposés sur le résultat des ces évaluations.

Pour ce faire, le présent mémoire se divise en deux sections. La première section, la recension des écrits, se compose de trois chapitres. Le premier chapitre rapporte les écrits importants concernant les fondements théoriques de l'évaluation et de l'utilisation de plusieurs indicateurs, le second porte sur l'utilisation d'échelles de mesure communes et le dernier traite de l'agrégation et de la pondération des résultats. La seconde section comporte deux chapitres. Le premier décrit la méthodologie d'évaluation des solutions répertoriées lors de la recension des écrits et le second présente et discute des résultats des analyses statistiques effectuées.

---

[1] La France, le Royaume-Uni et les États-Unis sont des pays qui procèdent à l'évaluation de la performance des écoles depuis plus de 10 ans.

CHAPITRE I

L'ÉVALUATION SOMMATIVE DE LA PERFORMANCE DES ÉCOLES : CRITÈRES MULTIPLES OU COMPOSITE?

Le chapitre qui suit montre qu'il est préférable d'utiliser une composite aux lieu de critères multiples lors de l'évaluation de la performance des écoles. Cet exercice requiert l'attribution d'un jugement de valeur à partir de plus d'un indicateur et convient bien à la problématique étudiée. Nous étudions les fondements de la pensée évaluative, les différentes pratiques et propositions de pratiques évaluatives en éducation, ainsi que les méthodes d'analyse de données qui s'offrent au chercheur ayant plus d'un indicateur à sa disposition.

1.1 Les fondements de la pensée évaluative

Au cour des vingt-cinq dernières années, les sciences sociales ont vu naître de nouveaux paradigmes épistémologiques (Lincoln et Guba, 2000). L'évaluation de programme a aussi été influencé par ces approches nouvelles. Le positivisme est le paradigme dit dominant en sciences sociales. Le postpositivisme, la théorie critique et le constructivisme sont des approches nouvelles qui se distinguent du positivisme à plusieurs égards. Comme les récents développements en évaluation de programmes sont issus du courant constructiviste et que les évaluations de la performance des écoles au Québec sont de type positiviste, nous concentrons ici nos efforts d'analyse sur les approches positivistes et constructivistes en évaluation de programmes.

Selon Lincoln et Guba (2000), le positivisme et le constructivisme diffèrent principalement au niveau ontologique, épistémologique et méthodologique. Premièrement, au niveau ontologique, le paradigme positiviste veut qu'une réalité

unique existe indépendamment de l'intérêt de l'observateur et qu'elle agisse selon des lois immuables qui prennent souvent la forme de relations de cause à effet. Le paradigme constructiviste s'oppose à cette perspective ontologique réaliste et adopte une approche relativiste. Selon les constructivistes, il existe de multiples réalités qui sont le fruit de construits sociaux et qui ne sont gouvernées par aucune lois naturelles.

Deuxièmement, au niveau épistémologique, les positivistes adoptent une position objectiviste et affirment qu'il est possible pour un observateur de se détacher du phénomène étudié et de poser un jugement clair et impartial à partir d'observations. À l'opposé, les tenants du constructivisme optent pour une vision épistémologique subjective. Ils affirment qu'observateur et phénomène observé sont indissociables, tant et si bien que les observations sont des créations du processus de recherche.

Troisièmement, au niveau méthodologique, les positivistes croient qu'en contrôlant les facteurs externes lors d'expérimentations, on peut expliquer la nature et le fonctionnement des phénomènes étudiés, ce qui donne la capacité de les prédire et de les contrôler. L'approche méthodologique constructiviste repose sur un cycle continu d'itération, d'analyse, de critique, de réitération, de nouvelle analyse, etc... conduisant à la « construction » d'une réalité commune.

House et Howe (1999) ont analysé les différentes interprétations des concepts de « valeur » et de « fait » en évaluation et concluent qu'elles sont reliées aux paradigmes positivistes et constructivistes. Ils illustrent ce qu'ils entendent par fait et valeur à l'aide d'un continuum où l'on retrouve d'un côté les énoncés qui ne portent que sur des faits (faits bruts) comme « l'école X compte plus d'étudiants que l'école Y ». De l'autre, les énoncés qui reposent en totalité sur les préférences personnelles (valeurs pures) comme « les cours de mathématique sont plus intéressants que les cours de géographie ».

Selon House et Howe, les évaluateurs positivistes croient généralement que fait et valeur sont des concepts bien distincts. Non seulement les évaluateurs peuvent légitimement déterminer les valeurs à étudier de même que les faits qui s'y rattachent, ils en ont le devoir. Ainsi, la définition des concepts de valeur et de fait adoptée par les positivistes place l'évaluateur au cœur même de l'évaluation d'un programme puisque le choix des valeurs et des faits utilisés pour fin d'évaluation oriente généralement le déroulement de celle-ci.

Contrairement aux évaluateurs positivistes, les évaluateurs constructivistes refusent de séparer les faits bruts des valeurs pures :

> The positivists associated facts with science, means, cognition, objectivity, truth and rationality. On the value side were politics, ends, interests, subjectivity, power, and irrationality. By contrast, "radical constructivists" deny a sharp fact-value distinction by applying the radical undecidability thesis (reserved for values in the received view) to *both* sides of the fact value distinction. There is no truth or objectivity to be found anywhere (House et Howe, 1999, p. 56).

L'approche constructiviste de Guba et Lincoln (1989) repose sur l'intégration de ces concepts. Elle confère à l'évaluateur constructiviste un rôle différent de celui de l'évaluateur positiviste.

L'évaluateur positiviste se place au cœur de l'évaluation en tant que juge des valeurs à évaluer et des faits qui doivent être retenus. Guba et Lincoln (1989) suggèrent quant à eux à l'évaluateur de n'assumer qu'un rôle de médiateur dans la sélection des faits et des valeurs à utiliser pour les fins de l'évaluation. Il devra tirer des conclusions basées sur les faits et les valeurs identifiées par l'ensemble des parties prenantes du programme et il devra s'acquitter de cette tâche sans jamais remettre en doute ou influencer leurs choix. Eux seuls possèdent une connaissance approfondie de l'objet d'évaluation et c'est pourquoi l'évaluation est, selon Guba et Lincoln, un construit social.

Les approches constructivistes et positivistes en évaluation de programme diffèrent au niveau des postulats ontologiques, épistémologiques et méthodologiques et leur représentation divergente des concepts de valeur et de fait donnent naissance à des processus d'évaluation propres à chacune de ces approches.

Scriven (1991) définit la discipline de l'évaluation comme « the process of determining the merit, worth and value of things » (p. 1). Selon cette définition, l'évaluation ne consiste pas seulement à compiler des données pertinentes au processus de prise de décision. Aussi laborieuse puisse-t-elle être, la collecte et l'analyse des données recueillies n'est qu'une des deux composantes essentielles à l'évaluation. À elles seules, les données ne peuvent porter un jugement de valeur sur un programme. Une composante additionnelle doit conférer un sens précis aux données rassemblées par l'évaluateur : l'objectif du programme.

Scriven (1991) décrit la pratique de l'évaluation en affirmant que « A more straightforward approach is just to say that evaluation has two arms, only one of which is engaged in data-gathering. The other arm collects, clarifies, and verifies relevant values and standard » (Scriven, 1991, p. 5). La pratique de l'évaluation à laquelle il souscrit consiste à : (a) déterminer une cible et les différents indicateurs nécessaires à son évaluation; (b) trouver les objectifs de performance pour chacune des cibles; (c) rassembler les données nécessaires, et (d) pondérer les indicateurs et agréger les résultats afin de juger du succès ou de l'atteinte des objectifs d'une politique ou d'un programme public.

Scriven (1993) distingue également deux types d'évaluations : l'évaluation formative et l'évaluation sommative. Quand l'objectif d'une évaluation est de fournir de l'information visant à contribuer à l'amélioration d'un programme, on la dit formative. L'évaluation formative amène l'évaluateur à travailler de concert avec plusieurs intervenants, dont l'administrateur du programme, qui désireront en savoir davantage sur l'implantation, la conceptualisation, les impacts et l'efficience du

programme dont ils ont la responsabilité. Quand l'objectif d'une évaluation est d'amener le preneur de décision à juger la pertinence de l'existence d'un programme, on la dit sommative. L'évaluation sommative doit s'appuyer sur des principes et standards scientifiques crédibles afin de fournir une assise solide au client principal.

Le client principal d'une évaluation, toujours selon Scriven (1993), est le consommateur. Selon lui, l'approche « consumériste »[2] en évaluation de programme ne devrait pas être différente de celle utilisée pour évaluer un bien ou un produit quelconque. Il suggère que l'évaluation d'un programme public soit calquée sur les évaluations de biens de consommation que proposent les magazines comme *Protégez-vous*. Les ministères et agences sont des instruments du gouvernement qui, en démocratie, sont en quelque sorte une agence du peuple. Puisque les consommateurs s'intéressent au produit final et non à son processus de fabrication, l'évaluation d'un programme devrait s'appuyer sur l'atteinte de résultats – non pas les résultats que le gouvernement estime être importants – mais bien ceux qui sont importants aux yeux des citoyens.

Guba et Lincoln (1989) s'opposent à la méthodologie évaluative positiviste. Selon eux, l'évaluation de type positiviste :

1. ne constitue pas une description de ce que « sont réellement les choses », mais représente plutôt un construit que des acteurs individuels ou des groupes d'acteurs forment pour « donner un sens » aux différentes situations dans lesquelles ils se trouvent;
2. que ces construits sont fortement influencés par les valeurs des gens qui les construisent;
3. qu'ils dépendent des contextes physiques, psychologiques, sociaux et culturels dans lesquels ils sont construits ou auxquels ils réfèrent et;

---

[2] Traduction libre de *consumer oriented.*

4. que les évaluations positives *peuvent* être édifiées de manière à affranchir ou à désaffranchir certaines des parties prenantes lors de l'évaluation.

Pour remédier à ces problèmes, ils proposent une alternative à ce modèle évaluatif : l'approche évaluative de quatrième génération. Cette approche reprend les postulats de base du constructivisme et les applique à l'évaluation de programme. Ainsi, avec l'approche évaluative de quatrième génération :

1. les « faits » et les « valeurs » sont inextricablement liés par le processus de reconstruction qui donne naissance l'évaluation subjective;
2. la responsabilité d'un programme appartient à l'ensemble des parties prenantes et aucune ne peut être individuellement tenue responsable de son échec ou de sa réussite et;
3. le rôle de l'évaluateur est d'orchestrer le processus de négociation dont l'objectif est de cheminer vers un construit plus sophistiqué et mieux informé.

Pour y arriver, le processus proposé par Guba et Lincoln (1989) consiste à : (a) identifier toutes les parties prenantes concernées par l'évaluation, (b) consulter chacune d'elles pour connaître quelles devraient être les cibles à évaluer, (c) fournir aux parties prenantes un contexte méthodologique qui rencontre les exigences de l'approche constructiviste mentionnées précédemment, (d) générer un consensus autour du plus grand nombre de constructions possibles, (e) préparer un agenda des négociations concernant les items qui ne font pas l'objet d'un consensus, (f) ramasser et fournir les informations nécessaires à la négociation, (g) établir un forum de médiation où les diverses parties prenantes peuvent négocier, (h) obtenir un consensus par la négociation et (i) recommencer le processus à nouveau.

En somme, les paradigmes positivistes et constructivistes ont influencé la discipline de l'évaluation. Chaque approche repose sur une conception ontologique, épistémologique et méthodologique différente, et ont une conception des faits et valeurs qui leur est propre ainsi que des processus d'évaluation dans lesquels l'évaluateur tient un rôle plus ou moins important. Nous pouvons maintenant montrer comment ces paradigmes ont influencé, implicitement ou explicitement, l'évaluation de la performance des écoles secondaires au Québec.

1.2 L'évaluation de la performance des écoles secondaires au Québec

Au Québec, il existe deux évaluations de la performance des écoles secondaires et une proposition d'évaluation. Le MEQ et le *Bulletin* évaluent annuellement les performances des écoles et le Conseil supérieur de l'éducation propose une approche évaluative dans son *Rapport annuel sur l'état et les besoins de l'éducation 1998-1999.*

Il est important de souligner que les travaux du MEQ et du *Bulletin* ne constituent pas des évaluations des écoles. Ce sont des mesures de la performance des écoles secondaires. À elle seule, la mesure de la performance ne peut être considérée comme une évaluation exhaustive des écoles. L'approche évaluative du Conseil constitue une proposition et aucune évaluation ou mesure de la performance d'envergure nationale ne s'en est inspirée, jusqu'à présent, dans le contexte québécois de l'évaluation des politiques en matière d'éducation. Malgré cela, les objectifs poursuivis par chacune de ces institutions illustrent bien la différence entre les paradigmes identifiés et décrits à la section précédante.

Les objectifs poursuivis par les deux évaluations de la performance des écoles produites au Québec sont quasi identiques. Dans la publication intitulée *Résultats aux épreuves uniques de juin 2000 par commission scolaire et par école pour les secteurs public et privé et diplomation par commission scolaire*, le MEQ établit clairement les objectifs de son instrument d'information. Ils devraient permettre :

- aux organismes des secteurs public et privé d'y puiser des renseignements pour évaluer leur action éducative, d'apprécier leur rendement dans le temps, de diagnostiquer leurs points forts et leurs points faibles ainsi que de comparer leur situation à celle des autres organismes scolaires,

- à la population d'avoir accès à de l'information officielle, conforme aux registres du Ministère,

et ce, en tenant compte du contexte et des différentes caractéristiques de chacun des organismes scolaires lors de l'interprétation des résultats (Québec, 2001, p. 1).

Quant au *Bulletin*, il vise à fournir une information de qualité à tous ceux qui ont à prendre des décisions à l'école secondaire. On parle ici des parents, étudiants, enseignants, cadres et professionnels, directeurs d'écoles et membres des conseils d'établissement (Marceau, 2000).

Les évaluations du MEQ et du *Bulletin* sont similaires. Premièrement, elles visent les mêmes clientèles : la population et divers organismes privé et public associés de près ou de loin au secteur de l'éducation. Deuxièmement, toutes deux se donnent pour mission d'informer ces derniers à l'aide d'indicateurs de performance et de classement des écoles. De plus, chacune classe les écoles selon leur performance. Le *Bulletin* le fait en fonction de cinq indicateurs standardisés qu'il pondère pour obtenir une cote globale pour chacune des écoles. Ce score composite sert ensuite de base au classement des écoles secondaires. Le classement du MEQ est basé sur le taux de réussite moyen des écoles aux examens de fin de secondaire.

Le *Bulletin* et l'outil du MEQ s'inspirent de l'évaluation consumériste de Scriven (1993). L'objectif poursuivit par ces deux évaluations favorise l'approche client en mettant à sa disposition plusieurs informations ainsi qu'un classement des écoles. Ces classements sont basés sur les résultats aux examens finaux de fin de secondaire et facilitent la comparaison des écoles, au même titre qu'un classement de

biens de consommation produit par un magazine comme *Protégez-vous* aide son lecteur à choisir le bien qui répond le mieux à ses besoins et à ses attentes.

Les évaluations du *Bulletin* et du MEQ partagent l'approche ontologique, épistémologique et méthodologique du paradigme positiviste. En utilisant des résultats aux examens et en ayant recours à un classement, ils appuient l'idée voulant qu'une réalité unique existe indépendamment de l'intérêt de l'observateur et qu'il est possible pour un évaluateur de se détacher du phénomène étudié et de poser un jugement clair et impartial à partir de ces observations. On espère aussi expliquer la nature et le fonctionnement des phénomènes étudiés pour éventuellement avoir la capacité de les prédire et de les contrôler.

Le processus d'évaluation adopté par ces deux institutions repose sur la distinction des concepts de valeur et de fait, et accorde une place importante à l'organisme qui agit à titre d'évaluateur dans le choix des objectifs à évaluer. Cette latitude se traduit par l'adoption du taux de réussite aux examens par le MEQ et de la cote globale par le *Bulletin* pour classer les écoles.

Dans son rapport annuel de 1998-1999 intitulé *L'évaluation institutionnelle en éducation : une dynamique propice au développement*, le *Conseil* propose une approche évaluative qui reprend les grandes lignes du paradigme constructiviste : l'évaluation institutionnelle.

> Il (le Conseil) définit donc ainsi l'évaluation institutionnelle : une démarche continue et concertée des acteurs concernés qui conviennent formellement d'analyser et d'apprécier, en tout ou en partie, la réalisation de la mission éducative de leur établissement ou de leur secteur d'intervention afin de s'ajuster de façon continue à un environnement en constante évolution, fondé sur la prise de décision et pouvoir ensuite en rendre compte à la société.(Québec, 1999, p.18).

On constate que l'évaluation proposée par le *Conseil* embrasse implicitement les grands principes ontologique, épistémologique et méthodologique du constructivisme. Premièrement, le *Conseil* propose que l'évaluation des écoles soit

une « démarche continue et concertée des acteurs concernés » qui varie d'une école à l'autre. Cela suppose qu'il croit en l'existence de multiples réalités qui sont le fruit de construits sociaux. Deuxièmement, l'emphase mise sur les acteurs tend à démontrer qu'observateur et phénomène observé sont indissociables pour le *Conseil.* Troisièmement, en affirmant que l'évaluation des écoles doit « s'ajuster de façon continue à un environnement en constante évolution », le *Conseil* adopte l'approche méthodologique constructiviste qui repose sur un cycle continu d'itérations, d'analyses et de critiques.

Quant à l'interprétation des concepts de valeur et de fait, le *Conseil* semble n'y voir aucune distinction puisqu'il suggère aux évaluateurs, tout comme le font Guba et Lincoln (1989), de diriger les discussions entre les différents acteurs, discussions qui, rappelons-le, servent à déterminer les objectifs à évaluer.

Puisque seuls le MEQ et le *Bulletin* procèdent à une évaluation de la performance des écoles secondaires et que toutes deux partagent un même paradigme, il importe de comprendre pourquoi ces deux évaluations ont recours à des méthodes d'évaluation semblables mais non identiques. La différence entre les évaluations de la performance des écoles du MEQ et du *Bulletin* vient des méthodes d'analyse de données multiples auxquelles elles ont recours et non de leurs assises ontologique, épistémologique et méthodologique ou de leurs définitions des concepts de fait et valeur et du rôle de l'évaluateur.

1.3 Méthodes d'analyse de données multiples.

Deux méthodes d'analyse des données s'offrent à l'évaluateur pour établir le niveau de performance atteint lorsque plusieurs indicateurs sont disponibles: la composite et les critères multiples. Une composite est le résultat de l'agrégation et de la pondération des scores obtenus par un individu, un objet ou un programme pour chacun des critères utilisés dans une évaluation. Les scores composites serviront par la suite à classer ou à comparer les objets de l'évaluation.

La cote globale sur 10 du *Bulletin* est un exemple de composite. Le score composite des écoles est le fruit de l'addition et de la pondération de leurs résultats qui se fondent sur cinq indicateurs distincts : (1) 40 % pour les résultats aux épreuves, (2) 20 % pour le taux d'échec, (3) 20 % pour le taux de promotion, (4) 10 % pour la surestimation des résultats par les écoles et finalement (5), 10 % pour l'écart entre les garçons et les filles.

La méthode des critères multiples mesure la performance d'individus, d'objets ou de programmes face à des critères particuliers. Par contre, les tenants de cette méthode s'opposent à la mise en commun de ces résultats en scores composites. Ils laissent ainsi au client de l'évaluation la tâche de rendre un verdict basé sur une interprétation personnelle des résultats mis à sa disposition.

L'instrument d'information du MEQ utilise un seul indicateur issu des résultats des écoles du Québec aux épreuves du ministère de l'Éducation. Bien qu'il classe les établissements d'enseignement québécois selon leur taux de réussite, on ne peut affirmer que ce classement représente une valeur composite puisqu'il ne tient compte que d'un seul indicateur alors qu'une composite, comme la définition le suggère, est le résultat de l'agrégation et de la pondération d'une série d'indicateurs. Le ministère laisse aux clients la tâche d'interpréter l'ensemble des résultats publiés par le MEQ. Par contre, il fournit un outil supplémentaire aux consommateurs en proposant un classement selon le taux de réussite des écoles. Il est ainsi un bon exemple de l'usage de critères multiples comme méthode d'analyse des données multiples.

Les adeptes de l'utilisation de la valeur composite soutiennent qu'une mesure de la performance globale est essentielle à la prise de décision lorsque l'on considère deux ou plusieurs aspects différents relatifs à la performance d'un individu. Selon Ghiselli (1956), les preneurs de décision imaginent les performances globales des individus comme des points dans un espace multidimensionnel. Dans cet espace, chaque dimension représente un aspect particulier de la performance globale d'un

individu. À moins qu'il ne ramène tous ces résultats individuels à une seule dimension, le décideur ne dispose d'aucune base solide pour justifier le choix d'un individu plutôt que d'un autre. Son choix repose sur l'importance accordée de manière subjective à chacun des critères de performance mesurés.

L'attribution d'un poids relatif à chacun des critères qui forment la composite explique l'importance accordée à chacun d'eux, élimine le biais relatif au jugement du décideur et assure une stabilité dans le temps. Par conséquent, Ghiselli (1956) et d'autres utilisateurs de la valeur composite voient dans l'utilisation de critères multiples une façon d'éviter le problème de la pondération des indicateurs et non pas une façon de le résoudre.

Quant aux chercheurs en faveur de l'utilisation de critères multiples, ils estiment que les résultats d'indicateurs ne devraient pas être combinés puisque divers arrangements de résultats par rapport aux critères décisionnels peuvent s'additionner pour donner des valeurs composites égales. Cattell (1957) et Dunnette (1963), soutiennent quant à eux que l'agrégation de critères noncorrélés devient très difficile à interpréter et que cette lacune suffit à invalider les avantages inhérents à l'utilisation de la valeur composite.

Brogden et Taylor (1950) s'objectent à cette idée de Dunnette (1963) et insistent sur le fait que lorsque les critères s'avèrent tous être des éléments de mesure provenant d'un même standard, ils peuvent être combinés pour former un index composite indépendamment de leurs indices de corrélation. Même si cette corrélation est nulle ou négative, les critères sont, par définition, positivement reliés au construit sous-jacent et donc des mesures d'une même cible ou d'un même objectif. Il ajoute que l'addition d'indicateurs servant à mesurer un même objectif augmente la précision des résultats de la composite.

Au-delà des arguments théoriques, des démonstrations pratiques confirment la supériorité de la valeur composite sur les critères multiples. Les recherches de Meehl

(1954) et Dawes et Corrigan (1974) concluent que les prédictions statistiques s'avèrent plus justes que les prédictions cliniques[3]. Selon Nisbett et Ross (1980), même une pondération statistique arbitraire des variables qui forment un score composite donne des prédictions plus fiables que celles d'une prédiction clinique. De façon générale, non seulement les gens appliquent des pondérations invalides, mais ils les appliquent de manière inégale.

Plus récemment, Schneider (1999) a démontré que les parents ont généralement accès à peu d'informations concernant la performance des écoles. Il note aussi que les consommateurs ne consacrent que très peu de temps à la recherche d'informations et qu'ils en consacrent encore moins à analyser et à comparer les informations recueillies. Il suggère de circonscrire et de rassembler l'information donnée aux parents de façon à maximiser son utilisation.

1.4 Conclusion

L'étude des fondements de la pensée évaluative montre que les évaluations de la performance des écoles secondaires au Québec, c'est-à-dire celle du MEQ et celle du *Bulletin*, sont toutes deux de type consumériste et adoptent une approche évaluative positiviste. En effet, toutes deux s'inspirent des principes fondamentaux des évaluations positivistes, soit l'utilisation des résultats aux épreuves de fin de secondaire (faits) dans le but d'évaluer la performance des écoles secondaires québécoises (valeur). Quant à l'étude des méthodes d'analyse des données multiples, elle démontre que la différence majeure entre l'évaluation produite par le MEQ et celle du *Bulletin* vient du nombre d'indicateurs publiés, et surtout du traitement des indicateurs lors de l'évaluation de la performance des écoles et du classement de celles-ci.

---

[3] Combiner des données de manière clinique consiste à laisser un expert jeter un coup d'œil à l'ensemble des variables critères afin qu'il en arrive à une prédiction finale qui s'appuie sur son jugement et son expérience.

Le MEQ a recours aux critères multiples comme méthode d'analyse de données multiples et propose un classement basé exclusivement sur le taux de réussite. Ce classement ne constitue pas un exemple de composite. Une valeur composite doit agréger et pondérer plus d'un indicateur d'un même standard. Le classement du MEQ n'est en fait qu'une façon différente de présenter l'information déjà contenue dans l'indicateur du taux de réussite. La cote globale du *Bulletin* est le seul exemple d'évaluation de la performance des écoles secondaires par la valeur composite au Québec. Elle est le résultat de l'agrégation et de la pondération des résultats de chacune des écoles évaluées par cinq indicateurs.

L'étude des méthodes d'analyse de données multiples révèle que l'utilisation de la valeur composite lors de l'attribution d'un jugement de valeur en présence de plus d'un indicateur est recommandée. L'utilisation de la valeur composite dans un cadre évaluatif a pour effet d'augmenter la précision de l'évaluation comme le prouvent les travaux de Meehl (1954), Dawes et Corrigan (1974) et Nisbett et Ross (1980) qui ont comparé les résultats des prédictions cliniques à ceux des prédictions statistiques. Brogden et Taylor (1950) ont pour leur part montré que l'ajout d'indicateurs servant à mesurer un même phénomène augmente généralement la précision des résultats de la composite, peu importe leur corrélation.

Quant à Schneider (1999), il suggère aussi de circonscrire et de rassembler l'information donnée aux parents de façon à maximiser son utilisation. Selon lui, une évaluation sommative de la performance des écoles doit être concise en raison du peu de temps que les parents sont généralement prêt à consacrer à l'analyse de l'ensemble des données.

Nous savons que l'utilisation d'une composite est conseillée lors de l'attribution d'un jugement de valeurs à partir de multiples indicateurs, comme c'est le cas lors de l'évaluation de la performance des écoles. Reste à déterminer l'effet engendré par l'ajout de quatre indicateurs par le *Bulletin* sur le classement des écoles

secondaires du Québec. Pour ce faire, nous devrons comparer le classement des écoles secondaires du *Bulletin* à celui produit annuellement par le MEQ qui repose uniquement sur le taux de promotion.

CHAPITRE II

LA NEUTRALISATION DES INSTRUMENTS DE MESURE PAR L'UTILISATION D'ÉCHELLES COMMUNES

Au chapitre précédent, nous avons établi la supériorité de la valeur composite lorsqu'un évaluateur doit attribuer un jugement de valeur à partir de plus d'un indicateur. Dans le présent chapitre, nous montrons pourquoi la standardisation et la normalisation des données sont les meilleures méthodes pour neutraliser les instruments qui servent à mesurer la performance relative de plusieurs individus à plus d'une épreuve.

Premièrement, nous passons en revue les différentes échelles de mesure commune. Ensuite, nous expliquons pourquoi les scores bruts en éducation doivent être modifiés si l'on désire utiliser une valeur composite. Deuxièmement, nous analysons les effets de trois transformations possibles, soit le rang centile, la standardisation et la normalisation en prenant soin de contraster les avantages et les inconvénients découlant de l'utilisation de chacune d'elles.

2.1 L'échelle de mesure commune

La neutralisation des instruments utilisés pour mesurer la performance des écoles secondaires du Québec passe par l'utilisation d'une échelle de mesure commune. Une échelle commune permet de comparer les résultats de plusieurs indicateurs et d'utiliser ces données pour former une composite. Or, toutes les échelles de mesure ne sont pas des échelles de mesures communes. En fait, peu d'échelles utilisées en éducation méritent d'être considérées comme de véritables échelles de mesures communes. Comme nous le verrons, les échelles de données brutes doivent être transformées si l'on présume les utiliser dans le but de comparer divers indicateurs ou de procéder à l'attribution de scores composites.

2.1.1 Qu'est ce qu'une échelle de mesure?

> « A score scale refers to numbers, assigned to individuals on the basis of test performance, that are intended to reflect increasing levels of achievement or ability » (Petersen, Kolen et Hoover, 1989, p. 221).

Une échelle de mesure sert à attribuer à un individu, sur la base de sa performance à une épreuve, un nombre qui reflète une amélioration des connaissances ou des habiletés. Par exemple, on attribue souvent aux étudiants une note sur une échelle de 100 basée sur sa performance à une ou à plusieurs épreuves. Plus la note se rapproche du score parfait, plus les compétences de ce dernier sont élevées. Ainsi, l'étudiant ayant une note de 80 % devrait, théoriquement, faire preuve d'une plus grande compétence qu'un étudiant ayant obtenu une note de 70 %, toutes choses étant égales par ailleurs.

À elle seule, cette définition d'une échelle de mesure soulève plusieurs questions intéressantes. On peut analyser mot à mot cette définition et y trouver bon nombre de questions toutes aussi pertinentes les unes que les autres. Premièrement, la définition suggère qu'une échelle de mesure est un ensemble de nombres. Mais quels nombres? Certains utilisent une échelle de mesure qui attribue des scores situés entre 0 et 100, comme par exemple les résultats aux examens du MEQ, alors que d'autres utilisent des scores situés entre 0 et 4.3, comme c'est souvent le cas des moyennes cumulatives octroyées par les universités aux étudiants. Comment peut-on comparer ces scores? Sont-ils comparables?

Deuxièmement, selon cette définition, ces nombres sont attribués à un individu sur la base de performance à un ou plusieurs tests. Peuvent-il être attribués à plus d'un individu, un groupe d'individus par exemple? Doit-on administrer seulement un test à tous ces individus ou bien plutôt une batterie de tests mesurant divers aspects de la performance? Et cette performance, comment l'évalue-t-on?

Troisièmement, les nombres doivent refléter un niveau croissant de performance ou d'accomplissement. On peut se demander si l'augmentation du

niveau de performance sur une échelle de mesure est constante entre chacun des nombres. Par exemple, est-ce que le fait de passer d'un score de 70 % à un score de 80 % représente le même progrès que le passage d'un score de 80 % à 90 %? Est-ce qu'une augmentation et une diminution des habiletés mentales se traduit par une augmentation ou une diminution comparable des scores obtenus? Ce sont là autant de questions auxquelles nous essaierons de trouver réponses tout au long de ce chapitre.

2.1.2 Le problème des données brutes.

Le *Bulletin* et l'instrument d'information du MEQ mesurent la performance des écoles secondaires du Québec à l'aide des résultats obtenus par les étudiants de celles-ci aux épreuves du Ministère. L'échelle de mesure utilisée par le Ministère pour évaluer les étudiants est une échelle de 0 à 100 pour laquelle 100 est un score parfait et 60, une note de passage. En plus des résultats aux examens, le *Bulletin* utilise également d'autres données pour mesurer la performance des écoles secondaires de la province : le pourcentage de réussite, le taux de transition, la surestimation des résultats par l'école et la différence moyenne entre garçons et filles à l'égard de la surestimation des résultats par l'école.

Toutes ces variables ont des échelles qui leur sont propres. Les résultats des écoles québécoises sur ces différentes échelles de mesure sont appelés scores bruts. Le tableau 2.1 confirme que les échelles de scores bruts ne peuvent pas être considérées comme des échelles de mesure communes.

Prenons par exemple l'étudiant I. En observant ses résultats aux différents examens, est-il possible d'affirmer qu'il a enregistré son meilleur score à l'épreuve de langue maternelle (195) et son plus faible en langue seconde (20)? Et en additionnant les scores bruts des deux étudiants, est-il juste de dire que l'étudiant I est plus performant que l'étudiant II? Après tout, l'étudiant I cumule 434 points, soit 37 de plus que son rival. Vous conviendrez que pour tirer des conclusions justes quant à la performance de chacun, il devient évident que l'on doive avoir accès à davantage de

renseignements. En servant de point de référence, la moyenne du groupe aux diverses épreuves peut nous aider à y voir plus clair. Cette information aide à fournir une réponse juste et définitive aux questions précédentes.

En comparant les résultats des deux candidats aux diverses épreuves à la moyenne des autres étudiants, on constate que l'étudiant I est au-dessus de la moyenne en langue maternelle, en histoire et en mathématiques alors que les résultats de l'étudiant II sont supérieurs à la moyenne en langue maternelle, en langue seconde, en sciences physiques et en mathématiques. L'étudiant II pourrait alors être considéré supérieur à l'étudiant I puisque ses résultats se situent au-dessus de la moyenne pour 4 des 5 épreuves comparativement à 3 pour l'étudiant I. Néanmoins, si on y regarde de plus près, on remarque que les résultats du premier étudiant sont supérieurs à ceux du deuxième dans 3 des 5 épreuves, donnant cette fois un avantage à l'étudiant I. De plus, les résultats de l'étudiant I sont de beaucoup supérieurs à la moyenne en langue maternelle, en histoire et en mathématiques et également inférieurs à la moyenne en sciences et en langue seconde. Quant au second, il est fortement au-dessus de la moyenne en sciences et en langue seconde, légèrement au-dessus de la moyenne en mathématiques et en langue maternelle et légèrement sous la moyenne en histoire.

Tableau 2.1

Comparaison des scores bruts, des écarts-types et des scores standardisés de deux étudiants à cinq épreuves distinctes

Matières	Moyenne	Écart-type	Scores bruts		Déviations		Scores standardisés	
			I	II	I	II	I	II
Lang. Mat.	155,7	26,4	195	162	+ 39,3	+ 6,3	+ 1,49	+ 0,24
Lang. Sec.	33,7	8,2	20	54	- 13,7	+ 20,3	- 1,67	+ 2,48
Sc. Phys.	54,5	9,3	39	72	- 15.5	+ 17,5	- 1,67	+ 1,88
Histoire	87,1	25,8	139	84	+ 51,9	- 3,1	+ 2,01	- 0,12
Math.	24,8	6,8	41	25	+ 16,2	+ 0,2	+ 2,38	+ 0,03
Sommes			434	397	+ 78,2	+ 41,2	+ 2,54	+ 4,51
Moyennes					+ 15,64	+ 8,24	+ 0,51	+ 0,90

Dans pareil cas, lequel des critères ci-dessus devrait servir de base au choix du meilleur étudiant? L'indice de déviation des résultats de chacun par rapport à la moyenne peut s'avérer d'une grande utilité. Ainsi, si l'on ne tient compte que de ce facteur, l'étudiant I sort vainqueur de l'affrontement. En effet, ces résultats sont en moyenne 15.64 points supérieurs à la moyenne contre seulement 8.24 points en moyenne pour l'étudiant II. Cependant, en utilisant cette méthode, on attribue un poids égal à chaque point d'écart à la moyenne bien que de toute évidence, les épreuves n'utilisent pas la même échelle de mesure. Si un étudiant est au-dessus de la moyenne dans deux disciplines, on peut se demander dans laquelle il est le plus performant? Par exemple, est-ce que l'écart de 39 points en langue maternelle de l'étudiant I est supérieur à son écart de 16,2 points en mathématiques considérant le fait que les moyennes ne sont pas les mêmes? En plus de tenir compte de l'écart entre la moyenne du groupe et celle de l'étudiant I, on doit également se préoccuper de la différence entre l'écart de son score par rapport à la moyenne et à l'écart moyen des autres individus à cette même moyenne.

Comme nous le verrons plus loin, le score standardisé est construit de façon à refléter cette particularité. Si l'on compare les scores standardisés de l'étudiant I en langue maternelle et en mathématiques, les résultats sont tout autres. L'écart de 39 points en langue maternelle est 1.49 écart-type au-dessus de la moyenne, et l'écart de 16.2 points en mathématiques devient, une fois transformé, 2.38 écart-type au-dessus de la moyenne. C'est donc dire que comparativement à l'écart-type de son groupe de référence, l'étudiant I est plus fort en mathématiques qu'en langue maternelle et ce, en dépit du fait que son écart avec la moyenne soit, d'un point de vue strictement numérique, plus important en langue maternelle. Quant au choix du meilleur étudiant, l'utilisation de l'échelle standardisée donne avantage à l'étudiant II dont les résultats sont en moyenne 0.90 écart-type au-dessus de la moyenne comparativement à 0.51 pour l'étudiant I.

Comme le prouve cet exemple, le problème des scores bruts est qu'ils ne peuvent, à eux seuls, fournir une image fidèle et complète de la performance qu'ils sont sensés mesurer chez un individu. La justesse de l'interprétation des résultats nécessite l'étude de renseignements supplémentaires. Ceux-ci peuvent revêtir un caractère fonctionnel ou normatif. Une information fonctionnelle peut prendre la forme d'une note du professeur qui indique le niveau de performance acceptable pour l'examen de langue maternelle et l'examen de mathématiques ou d'autres renseignements de nature qualitative qui nous aident à interpréter les résultats obtenus. Les informations peuvent aussi être de nature normative si elles décrivent les performances d'un groupe d'individus dont les caractéristiques (moyenne, écart-type et rang centile) sont connues des utilisateurs du test et communiquées aux parents (Angoff, 1971).

Dans un autre ordre d'idées, Angoff (1971) maintient qu'il est important de reconnaître que les scores bruts affichent peu ou pas de généralités puisqu'ils sont le produit des diverses questions contenues dans le test. Par contre, il concède que certains y verront un avantage, puisque l'utilisation des résultats bruts comme échelle de mesure permet de découvrir les forces et les faiblesses de construction d'un examen. Par exemple, une distribution des scores bruts souffrant d'une asymétrie positive peut indiquer que l'examen soumis aux participants était trop difficile pour une grande partie d'entre eux ou encore qu'il évalue des compétences qui font défaut à plusieurs.

En dépit de cet avantage, le fait que les scores bruts n'affichent que peu ou pas de généralités occasionne un problème que l'on ne peut ignorer : l'utilisation d'échelles de mesure basées sur les scores bruts peut s'avérer problématique car on risque de confondre les scores obtenus à différentes versions d'un même examen. Comme les examens du MEQ peuvent changer d'années en années, l'échelle brute ne peut nous indiquer si l'amélioration des résultats d'une école est attribuable aux efforts combinés des professeurs et de la direction de l'établissement dans la

préparation des étudiants ou au fait que les changements apportés à l'examen l'ont rendu plus facile.

2.1.3 Présentation des résultats des évaluations de la performance des écoles au Québec.

Nous avons vu précédemment que le MEQ se donne comme mandat d'informer la population et les divers intervenants des secteurs public et privé des résultats obtenus par les élèves des écoles québécoises à leurs épreuves. Pour ce faire, ils publient annuellement les résultats moyens de chaque école à ces épreuves ainsi que leur taux de réussite aux examens du MEQ. Aussi, il prend soin d'inclure des informations normatives à ces résultats en publiant la moyenne nationale à chacune des épreuves et en offrant un classement des écoles basé sur leurs taux de réussite aux épreuves.

Pour sa part, le *Bulletin des écoles secondaires du Québec* publie les résultats bruts des écoles et la moyenne provinciale pour chacun des cinq indicateurs qui forment la cote globale. Il en est de même pour les données contextuelles, tels le pourcentage d'élèves handicapés ou en difficulté d'adaptation et d'apprentissage (EHDAA), le pourcentage d'élèves en retard, et le revenu des parents. Il classe ensuite les écoles à l'aide de la cote globale. Or, malgré ces différences, chacun d'eux a recours aux informations normatives pour préciser les résultats obtenus aux différentes épreuves par les élèves de chaque établissement scolaire.

Le MEQ est conscient du fait que les examens administrés aux étudiants ne sont pas les mêmes d'une année à l'autre. Dans le document *Résultats aux épreuves uniques de juin 2000 par commission scolaire et par école pour les secteurs public et privé et diplomation par commission scolaire*, on peut lire que « Pour chaque matière et chaque session, une nouvelle épreuve est rédigée et que d'une session à une autre, le Ministère s'efforce de construire des épreuves comparables » (Québec, 2001, p. 6). On y mentionne également que « Par souci de justice, le Ministère veille à ce que les

épreuves qu'il prépare comprennent d'années en années un degré de difficulté équivalent » (Québec, 2001, p. 7).

En dépit de la bonne volonté et des efforts déployés par le MEQ pour s'assurer de la comparabilité des examens, on peut supposer que les examens ne sont pas parfaitement identiques d'une version à l'autre. Étant donné cet inconvénient, l'utilisation d'une échelle de mesure commune s'avère nécessaire, car elle seule nous permet de comparer les résultats de divers examens et ce, peu importe comment et par qui les épreuves ont été créées ou quel était leur degré de difficulté.

2.2 L'utilisation d'échelles communes en éducation.

En éducation et en psychologie, la comparaison et l'agrégation fréquente de diverses variables obligent le chercheur à utiliser une échelle de mesure commune. S'il existe des échelles de mesure bien définies assurant la comparabilité des mesures *physiques*, il en va tout autrement pour la mesure des *habiletés mentales*. Pour comprendre comment les mesures physiques diffèrent des mesures d'habiletés mentales, nous allons reprendre l'exemple de Angoff :

> « The notion that one bar of steel is twice as long as a second bar is a meaningful one, easy to transmit and understand, even without the definition or the original derivation of the system of units for measuring them. The fact that this notion is implied when one says that the first bar measures six feet and the second only three derives from a willingness to accept the concept of zero length and the willingness to agree on an operation that defines the distance denoted as one inch, for example, at one part of the yardstick as equal to the distance denoted as one inch at any other part of the yardstick. » (1971, p. 509)

En comparaison, il est difficile de s'imaginer qu'une note de zéro à un examen d'aptitude représente une absence totale d'habileté mentale. De plus, on s'imagine mal que la différence entre le niveau d'habileté de Pierre qui peut taper 20 mots à la minute et celui de Jean qui peut en taper 40, soit équivalente à celle qui sépare Jean de Jacques qui en tape 60. Les efforts et le temps requis pour passer de 20

à 40 mots minute ne sont certainement pas les mêmes que ceux nécessaires au passage de 40 à 60 mots minute. Pourtant, dans chacun des cas, l'amélioration représente 20 points sur l'échelle de mesure. C'est aussi le cas des échelles de mesure utilisées en éducation. Le *Bulletin des écoles secondaires du Québec* utilise les résultats des élèves aux examens du MEQ comme base de comparaison. Or, un score de zéro à un examen de français ne veut pas dire que l'étudiant ne possède aucune connaissance en français et on ne peut affirmer qu'un étudiant ayant obtenu 100 en français possède deux fois plus de connaissances qu'un individu ayant obtenu un score de 50.

Les échelles de mesure physiques sont généralement des échelles de type *interval* puisqu'elles utilisent :

- une unité de mesure fixe comme le mètre ou le kilogramme,
- une valeur zéro qui représente une réelle absence de la caractéristique mesurée.

En l'absence d'une valeur zéro, une échelle utilisant quand même une unité de mesure fixe, comme c'est le cas de l'échelle de mesure thermique « Celsius », devient une échelle de *ratio*[4]. De toute évidence, les échelles utilisées pour mesurer les résultats des élèves québécois aux examens du MEQ ne sont donc pas des échelles *intervalles*.

Contrairement aux échelles *intervalles*, les échelles *ordinales* ont la particularité de pouvoir être ordonnées puisqu'un ordre de grandeur intrinsèque leur est sous-jacent. Pour reprendre l'exemple précédent, bien qu'on ne puisse déterminer de façon précise la différence réelle en terme d'habileté qui sépare Pierre de Jean et Jean de Jacques, on est quand même en mesure d'affirmer que Pierre est moins habile

---

[4] Les statisticiens utilisent souvent le terme « intervalle » pour nommer une échelle de type « ratio ». Comme elles se prêtent aux divers test statistiques aussi bien l'une que l'autre, cette légère confusion n'affecte en rien le contenu du présent travail.

que Jean qui, lui, est moins habile que Jacques. Les résultats aux examens exhibent ainsi les caractéristiques d'une échelle de type *ordinal*.

La présence des caractéristiques propres à l'échelle est d'une importance capitale lors de la mise sur pied d'une composite. Plusieurs chercheurs ont tenté d'expliquer pourquoi la valeur composite nécessite une transformation des données brutes. Stevens (1946) et Suppes et Zinnes (1963) furent les premiers à s'intéresser à cette problématique. Plus récemment, Michell (1986) a recensé trois différentes théories de la mesure. Une d'entre elles, proposée par Stevens (1946) et soutenue par d'autres chercheurs du domaine de la psychologie et de l'éducation, nous aide à comprendre la nécessité de la transformation des scores bruts quand vient le temps d'agréger et de comparer différentes variables. Cette théorie, c'est la *théorie représentationelle.*

La théorie représentationelle repose sur l'idée que les nombres sont utilisés dans la pratique de la mesure pour représenter des relations empiriques entre les objets. Pour ce faire, seules les échelles d'intervalles ou de ratio peuvent être utilisées puisqu'elles sont les seules à pouvoir faire le pont entre d'une part, les nombres, et d'autre part, la relation empirique qui les caractérise. Comme nous l'avons vu précédemment, les échelles d'intervalles et de ratio possèdent une unité de mesure fixe. Sans cette unité de mesure fixe, on ne peut utiliser les nombres à des fins mathématiques. Et comme le dit Michell (1986) :

> Mathematical analysis is powerful because it contains a storehouse of valid argument forms or theorems that may be applied to empirical propositions once numerical assignments are made. This enables us to derive empirical conclusions from data via mathematical arguments (p. 401)

Cette divergence entre échelle *ordinale* et *intervalle* mène des chercheurs comme Stevens (1951) à conclure que les résultats d'analyses statistiques qui sont appropriées lorsque faites à partir des échelles de ratio et d'intervalle, sont peu ou pas pertinents lorsqu'ils sont appliqués aux échelles de mesure d'habileté mentale. Les

échelles d'habileté mentale sont des échelles ordinales, alors que plusieurs opérations statistiques nécessitent l'utilisation d'une échelle de mesure d'intervalle ou de ratio. Or, le calcul de la cote globale du *Bulletin* et des résultats moyens aux épreuves du MEQ nécessite plusieurs opérations statistiques qui commandent l'utilisation d'échelles de mesure communes d'intervalle ou de ratio.

Guilford et Fruchter (1978) abondent dans le même sens et pensent qu'il est essentiel de transformer les scores bruts en valeurs d'une autre échelle de mesure :

> If modern psychology and education have taught anything about measurement, they have amply demonstrated the fact that there are few, if any, *absolute* measures of human behavior. The search for absolute measures has given way to an emphasis upon the concept of individual differences. The mean of the population has become the reference point, and out of the differences between individuals has come the basis for scale units (p. 472-473).

Pour palier ce problème, Petersen, Kolen et Hoover (1989) proposent que les chercheurs utilisent non pas une, mais bien deux échelles de mesure : *l'échelle primaire* et *l'échelle auxiliaire*. L'objectif de l'échelle primaire consiste à présenter les résultats des différents tests utilisés, alors que l'objectif de l'échelle auxiliaire est d'augmenter l'interprétabilité de la première.

> Auxiliary score scales are used because, in many situations, it is desirable to convey more information about test performance than can be incorporated into a single primary score scale. Auxiliary score scales are used to convey additional normative information, test-content information, and information that is jointly normative and content based. For many test uses, an auxiliary scale conveys information that is more crucial than the information conveyed by the primary score scale. In such instances, the auxiliary scale is the one that is focused on, and the primary scale can be viewed more as a vehicle for maintaining interpretability over time. (p. 222)

C'est une position qu'ils partagent d'ailleurs avec Angoff (1971) qui croit, lui aussi, que la création d'une échelle de mesure qui incorpore un sens normatif à partir

de données brutes peut s'avérer une alternative de choix lorsque les données brutes ne procurent que peu d'informations:

> It usually is maintained that the raw score scale yields little or no immediate meaning of its own. For that reason, derived scores scales are established in which normative meaning is directly incorporated (p. 527).

Nous le savons, les échelles ordinales utilisées en éducation ne comportent, à l'état pur, que trop peu d'informations. La création et l'utilisation d'une échelle auxiliaire, proposé par Angoff et Petersen et Kolen et Hoover, offre l'opportunité d'échapper à cette lacune en transformant les données brutes en échelles d'intervalle ou de ratio. Afin de s'assurer que les scores demeurent comparables à travers le temps et d'un examen à l'autre, les données transformées doivent avoir comme point de référence la moyenne de la population ainsi que l'écart de chaque individu par rapport à cette moyenne, tel que proposé Guilfor et Fruchter (1978). C'est l'utilisation de la moyenne et de l'écart-type qui procure à l'échelle auxiliaire les caractéristiques qui font défaut à l'échelle primaire. Autrement dit, c'est en se basant sur la moyenne et l'écart-type des résultats bruts que l'on peut créer une échelle d'intervalle ou de ratio à partir d'une échelle ordinale.

2.3 Les transformation possibles.

Plusieurs modifications statistiques transforment les valeurs brutes en échelles d'intervalle et de ratios. Dans le cadre de la recherche, nous analyserons trois types de transformations régulièrement utilisées à cette fin : le rang centile, la standardisation et la normalisation.

Avant de poursuivre, il importe de comprendre que ces transformations impliquent des changements conceptuels importants. Gulliksen (1962, p. 267) souligne que « In using standard, linear derived, percentile, or normalized scores, we should bear in mind that such scores indicate only the relationship of the individual to a given group. They indicate nothing about the general level of knowledge or

attainment of the group or its members ». En contrepartie, c'est un compromis nécessaire à la transformation de données ordinales et il est toujours possible d'étudier les résultats de l'échelle primaire si ce sont ces données qui nous intéressent. C'est précisément pour cette raison qu'il est recommandé d'utiliser deux échelles, l'échelle primaire et l'échelle secondaire, afin qu'il soit toujours possible de comparer les scores transformés aux scores d'origine.

2.3.1 Le rang centile.

Le rang centile est une échelle de mesure commune de type ordinal. Elle fait tout de même l'objet d'une description plus approfondie puisqu'elle possède des qualités intéressantes. L'échelle centile possède une valeur zéro tangible et son unité de mesure semble équivalente à première vue. Une fois transformés en rang centile, les résultats de divers tests peuvent être comparés entre eux. Cette échelle est divisée en 100 unités (centiles). Si une école obtient la valeur 95, cela signifie que par rapport à un échantillon de 100 établissements d'enseignement, son résultat serait supérieur à celui de 95 autres établissements. Si, par contre, elle obtient une score de 1, c'est que des 100 écoles, elle est la moins performante.

La figure 2.1 montre que l'échelle centile a une distribution rectangulaire. La conversion de scores bruts en scores centiles a pour effet de séparer les personnes près du milieu de la distribution initiale. Ainsi, au milieu de cette distribution rectangulaire, une grande différence en score centile peut être le résultat d'une petite différence en performance réelle. À l'inverse, aux extrémités de la distribution, une grande différence en terme d'habileté peut ne représenter qu'une petite différence en score centile. Des personnes ayant des scores bruts très différents se retrouvent souvent coincées à l'intérieur du même rang centile et c'est pourquoi l'échelle centile se veut une échelle de type ordinal et non pas d'interval, bien qu'elle possède une véritable valeur zéro (Cronbach, 1990).

L'unité de mesure de l'échelle centile est inégale puisqu'elle représente une proportion égale d'un groupe et non un interval égal sur une échelle d'habileté quelconque. Bien que les transformations nécessaires à l'obtention d'une échelle d'interval fassent en sorte que l'on utilise non pas les résultats des individus en tant que tel mais plutôt leurs différences par rapport à la moyenne du groupe de comparaison comme le souligne Gulliksen (1962), il n'en demeure pas moins que cette différence doit s'appuyer sur les résultats obtenus aux épreuves et non sur l'ordre dans lequel ils sont classés.

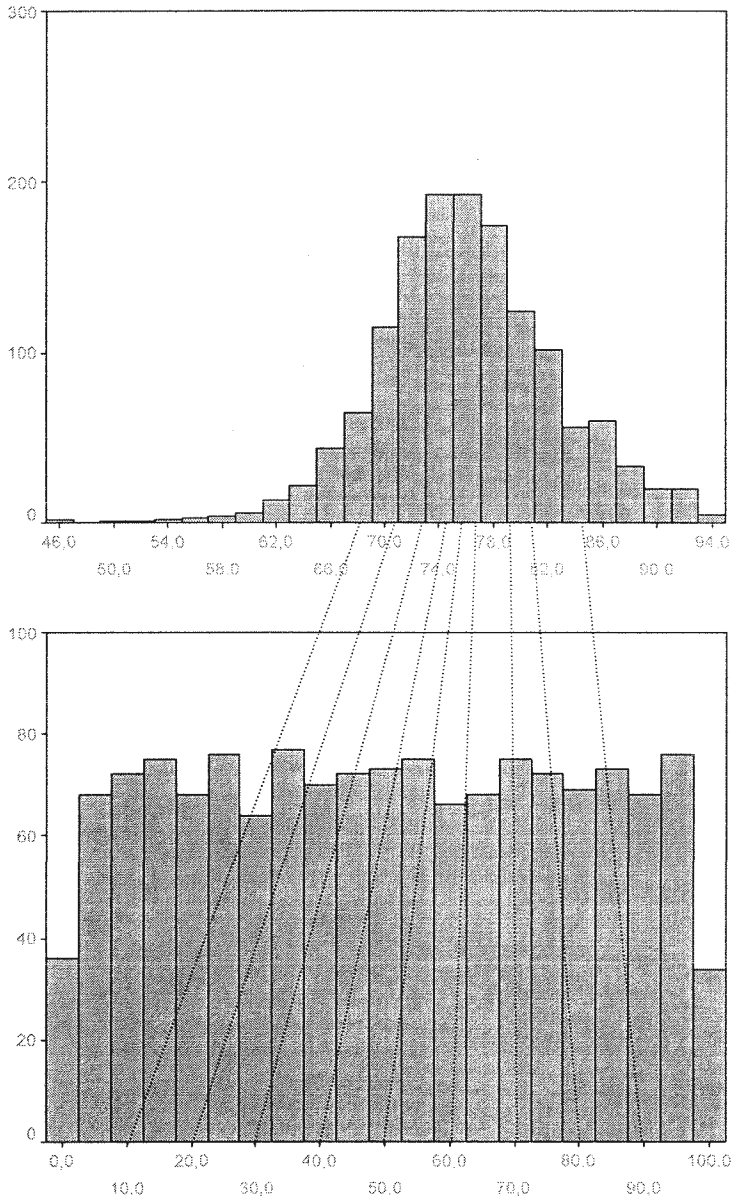

Figure 2.1 Effet du passage de l'échelle brute à centile sur la distribution des données

2.3.2 La standardisation.

Si la transformation des scores bruts en scores centiles permet la comparaison de différents tests entre eux, elle s'applique par contre mal au domaine de l'éducation car son unité de mesure représente une proportion égale d'un groupe et non un interval égal sur une échelle d'habileté. Pour remédier à cet inconvénient, l'échelle commune doit être construite à partir de la moyenne de population et l'écart de chacun par rapport à cette moyenne, ou écart-type. Selon Cronbach (1990, p. 116), « A standard score scale serves the same purpose as the percentile scale. A standard score reports how many standard deviations above or below the mean a person is. Changing from raw scores to standard scores of this kind does not alter the form of the distribution ».

Un score standardisé (souvent appelé score Z[5]) est obtenu à partir d'un score brut en utilisant la formule suivante :

$$z_i = (X_i - \mu) / \sigma \qquad (1)$$

Où z_i est le score standardisé, X_i le score brut, μ la moyenne et σ l'écart-type. Cette échelle de mesure a la particularité de toujours avoir une moyenne de 0 et un écart-type de 1. Fait à noter : la standardisation d'une échelle de mesure ne change pas la position relative des scores et n'affecte pas la distribution d'origine (voir figure 2.2). Tout comme l'échelle brute, elle permet de déterminer les forces et les faiblesses de la construction d'un examen en analysant la forme de la distribution.

---

[5] Plus précisément, les scores Z sont des transformations de scores normaux, alors que les scores standardisés sont le résultat de n'importe quelle variable, même celles qui ne sont pas distribuées de façon normale. Ces deux termes sont utilisés de manière interchangeable dans la littérature, ce qui sera également le cas du présent travail.

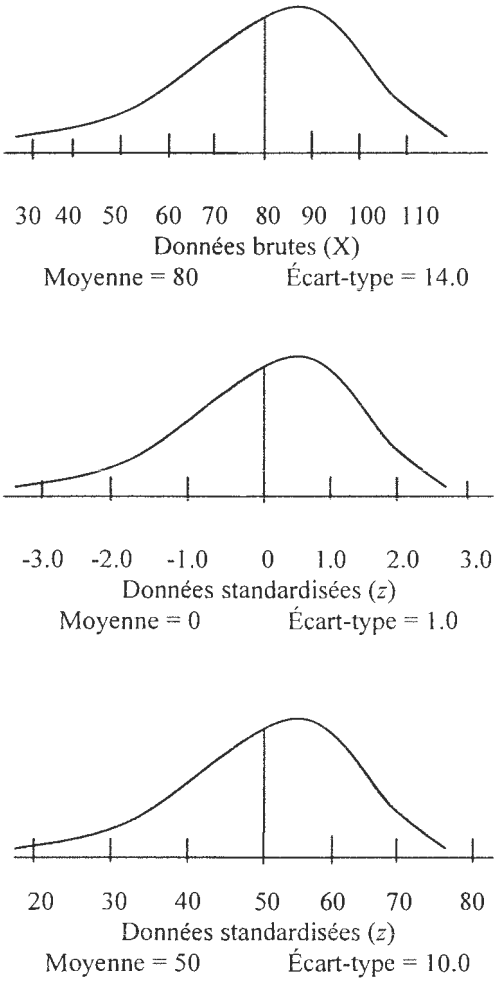

Figure 2.2 Distribution avant et après la conversion de l'échelle brute à l'échelle standardisée avec une moyenne et un écart-type prédéterminé

L'échelle standardisée est l'échelle utilisée par le *Bulletin* dans le but de neutraliser ses instruments de mesure. Lors de la formation des scores composites, les scores bruts sont standardisés chaque fois que des données sont comparées ou agrégées. Il est donc important de comprendre quels sont les avantages, les inconvénients et les contraintes liés à l'utilisation de cette échelle de mesure.

L'inconvénient de l'échelle standardisée est qu'elle est difficile à interpréter. Premièrement, la moitié des scores obtenus par cette transformation sont négatifs, ce qui complique bien des calculs (par exemple leur addition dans une valeur composite). Deuxièmement, son unité de mesure, l'écart-type, est somme toute relativement obscure pour le non initié. On peut aisément remédier au premier inconvénient en additionnant une constante à tous les scores. Les scores ainsi standardisés seront tous positifs. Quant à l'unité de mesure en soi, on peut multiplier chaque score par une autre constante pour rendre l'unité de mesure plus petite et l'étendue des scores plus grande. C'est pourquoi il est possible de transformer l'échelle de mesure standardisée de façon à ce qu'elle adopte la moyenne et l'écart-type souhaité, et ce, sans altérer sa courbe de distribution comme le montre la figure 2.2.

On utilise la formule suivante pour passer directement des données brutes à une échelle standardisée dont la moyenne et l'écart-types sont prédéterminés :

$$X_s = (S_s / S_o) X_o - [(S_s / S_o) X_o - X_s] \qquad (2)$$

où X_s est le score de l'échelle standardisée correspondant à X_o, X_o le score brut, X_o et X_s les moyennes de X_o et X_s respectivement, et S_s et S_o les écarts-types de X_s et X_o respectivement.

À la section précédente, on reprochait à l'échelle centile d'utiliser une unité de mesure inégale puisqu'elle est basée sur la différence de positionnement vis-à-vis un groupe d'appartenance. Sur ce point, l'échelle standardisée, bien qu'elle repose sur la différence d'habileté entre les individus, ne répond que partiellement au problème d'égalité des unités de mesure. Comme le souligne Angoff (1971) :

> In the linear transformation the separation between successive raw score units, or between scaled score units corresponding to successive raw scores, is considered equal only in the operational sense that each score represents one more item answered correctly than the preceding score (p.513)

Puisque l'unité de mesure de l'échelle standardisée est construite à partir d'une différence d'habileté entre les individus, elle constitue une alternative supérieure à l'échelle centile car cette dernière mesure la différence entre individus par rapport à leur positionnement dans un groupe. Par contre, le fait que son unité de mesure ne soit égale qu'au sens opérationnel restreint son utilisation. Angoff (1971) précise que pour que les résultats standardisés de deux tests différents soient comparables, deux conditions doivent être satisfaites : (a) on doit pouvoir s'assurer que leurs moyennes soient identiques et que la distribution de leurs résultats autour de la moyenne ait une dispersion identique et (b), que la forme de la distribution, en terme d'asymétrie et d'aplatissement, soit très similaire d'une habileté à une autre.

Guilford et Fruchter (1978) concluent, en parlant de ces contraintes liées à l'utilisation de l'échelle standardisée, que :

> If we want to achieve communality of scales at all, we often have to proceed on the assumption that actual means, standard deviations, and form of distribution are uniform for all abilities measured. In spite of these limitations, it is almost certain that derived scales, such as the standard-score scale, provide us with more nearly comparable values than raw-score scales do. The recognition of these limitations, however, should be admitted, and interpretations based on the use of standard scores should be made with reservations in line with those limitations. (p. 465-466)

La distribution des scores autour de la moyenne pose un autre problème. Que faire si une variable compte quelques scores qui se détachent nettement de la moyenne? Ces scores sont appelés scores extrêmes par le statisticien. Tabachnick et Fidell (2001) décrivent ces exceptions comme suit :

> Among continuous variables, univariate outliers are cases with very large standardized scores, z scores, on one or more variables, that are disconnected from other z scores. Cases with standardized scores in excess of 3.29 ($p<.001$, two-tailed test) are potential outliers. However, the extremeness of standardized scores depends on the size of the sample; with a very large *N*, a few standardized scores in excess of 3.29 are expected. (p. 67-68)

Après avoir localisé les scores extrêmes, Tabachnick et Fidell proposent deux alternatives. Premièrement, elles suggèrent de réduire l'influence des données extrêmes en changeant leurs scores de sorte qu'elles demeurent déviantes mais qu'elles le soient moins qu'auparavant. Les données extrêmes sont ainsi ramenées à une valeur plus conservatrice de 3.29 écarts-types. L'inconvénient lié à l'emploi de cette procédure est qu'elle n'élimine pas complètement le problème posé par la présence de données extrêmes. Les extrémités de la distribution ainsi transformée sont toujours victimes d'un regroupement des données, mais leur effet sur les résultats des opérations statistiques ultérieures en sera amoindri.

Deuxièmement, Tabachnick et Fidell suggèrent de tout simplement éliminer les cas extrêmes. Si notre échantillon est limité, la perte d'informations engendrée par l'élimination des cas extrêmes peut s'avérer importante. Par contre, l'effet de cette perte d'informations est moindre si l'on dispose d'un échantillon important. L'élimination des données extrêmes a aussi l'avantage de ne pas alourdir la queue de la distribution comme le fait la transformation des scores.

Un inconvénient lié à l'élimination des cas extrêmes vient du fait que cette modification n'élimine pas systématiquement toutes les données extrêmes d'une distribution lorsqu'on utilise l'échelle standardisée comme base de travail. En effet, une fois les données extrêmes d'une échelle standardisée retirées, on doit de nouveau standardiser la distribution afin d'obtenir une échelle commune. Cette opération fait en sorte que des cas qui n'étaient pas considérés comme extrêmes avant l'élimination des cas déviants sont désormais des cas extrêmes. La nouvelle distribution, bien qu'elle comporte encore des données extrêmes, devra alors être laissée telle quelle.

En résumé, l'échelle standardisée utilisée par le *Bulletin*, parce qu'elle est construite à partir de la moyenne de la population et l'écart de chacun par rapport à cette moyenne, permet la comparaison de divers indicateurs entre eux. De plus, comme elle n'altère pas la forme de la distribution, cette transformation peut également servir à évaluer la construction d'un test ou d'une épreuve. Par contre, le

fait que son unité de mesure ne soit égale qu'au sens opérationnel restreint son utilisation comme échelle de mesure commune. L'échelle standardisée peut néanmoins représenter une échelle de mesure commune viable si l'on prend soin de respecter les diverses contraintes que son utilisation impose ou de poser l'hypothèse que celles-ci sont respectées.

2.3.3 La normalisation.

Si, pour comparer différents tests entre eux, le chercheur doit poser l'hypothèse que la moyenne, la dispersion des résultats autour de la moyenne ainsi que la forme de la distribution de ces tests sont identiques, alors une question se pose : Pourquoi ne pas transformer les données brutes en une forme d'échelle dont l'unité de mesure serait à l'abri de ces inconvénients? Une échelle qui serait dispersée également et, par définition, opérationnelle. En émettant l'hypothèse que la forme d'une distribution devrait être normale peu importe l'habileté mesurée, on obtient une échelle que les chercheurs croient capable de mieux représenter l'écart réel entre les divers individus qui composent la population à l'étude. Cette échelle de mesure commune, c'est l'échelle normalisée.

Dans notre quête d'une unité de mesure idéale, donc parfaitement distribuée, on ne peut passer sous silence la distribution normale. Normaliser une échelle de données brutes, c'est en quelque sorte l'étirer pour que sa forme devienne normale. Une distribution normale est une distribution qui respecte la formule suivante :

$$Y = \frac{e^{-(x-\mu)^2 / 2\sigma^2}}{\sigma\sqrt{2\pi}} \qquad (3)$$

Ce que l'on doit retenir de cette formule, c'est qu'elle repose sur la moyenne (μ) et l'écart-type (σ). Les deux autres variables e et π sont des constantes. Il existe donc une infinité de courbes normales, une pour chaque combinaison possible de la moyenne et de l'écart-type. C'est donc dire que deux distributions normales peuvent avoir une allure bien différente.

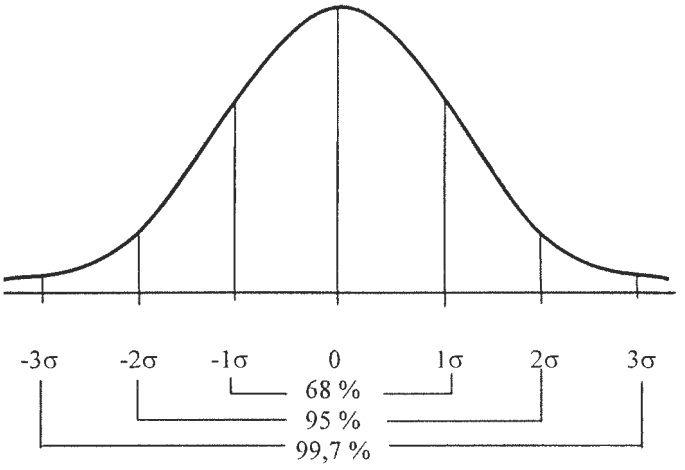

Figure 2.3 Caractéristiques de la distribution normale

Bien qu'elles ne soient pas nécessairement identiques, les courbes de distribution normale ont toutes les mêmes caractéristiques : à un écart-type de la moyenne, une distribution normale inclut 68 % des scores, 95 % à deux-écarts types et 99.7 % à trois écarts-types. Par ailleurs, la moyenne, la médiane et le mode sont identiques.

Pour transformer des scores bruts en scores normalisés, ce sont ces caractéristiques propres à la courbe normale qui servent de point de départ[6]. Une fois le rang centile de chaque score brut défini, on lui attribue, à l'aide d'une table de conversion, un score *Z* équivalent. Transformée de cette façon, l'échelle de mesure normale possède, tout comme l'échelle standardisée, une moyenne de 0 et un écart-type de 1. Et comme dans le cas de l'échelle standardisée, on peut transformer cette

---

[6] Quant on parle de normalisation des données, on réfère généralement à la méthode décrite ci-dessus. Par contre, il existe des transformations dont le but est de modifier la distribution afin qu'elle soit plus près de la normalité. Ces transformations ne feront pas l'objet d'une étude poussée dans le présent mémoire, mais le lecteur peut consulter Tabachnick et Fidell (2001) pour de plus

échelle de sorte que la moyenne et l'écart-type adoptent des valeurs plus conventionnelles. On utilisera aussi l'équation (2) pour y arriver.

Comme le souligne Cronbach (1990, p. 120), « the normal curve does not describe score distributions accurately, but test interpreters keep it in mind because in most circumstances it provides a good approximation ». L'adoption d'une échelle de mesure normale repose sur l'hypothèse que les habiletés mentales ont tendance à être distribuées de manière normale dans une population. Selon Angoff (1971), c'est précisément cette hypothèse qui fait que, contrairement à ceux des scores centiles, les écarts entre scores normalisés, qui pourtant sont déterminés à partir des scores centiles, représentent des changements égaux en terme d'habileté mentale et non par rapport à un groupe de comparaison.

Il nous met cependant en garde quant à la normalisation des données en précisant que :

> The transformation to a normal distribution is *not* considered advantageous when there is reasons to believe that the peculiarities in the shape of the raw score distribution reflect actual peculiarities in the distribution of ability of the group tested. (Angoff, 1971, p. 516)

Cette précision est importante puisqu'elle suggère que la normalisation ne devrait s'appliquer qu'aux distributions de scores bruts dont les déviations de la forme normale ne reflètent pas des particularités réelles de la distribution des habiletés du groupe testé, mais plutôt des erreurs de mesure.

La figure 2.4 montre que si la standardisation nécessitait des variables comparées, donc une distribution quasi identique afin de s'assurer que le nombre de personnes entre deux scores Z soit le même, la normalisation des scores bruts permet quant à elle aux chercheurs d'échapper à cet inconvénient en modifiant les limites des

---

amples informations concernant ces diverses transformations, leurs applications et leurs effets sur la courbe de distribution.

échelles pour s'assurer que l'on y retrouve la même aire pour chaque section sous la courbe de distribution.

Prenons par exemple l'intervalle I de la figure 2.4. L'espace compris entre les scores *Z* qui marquent le commencement et la fin de l'intervalle I est égal à l'espace compris entre les scores qui limitent chacun des autres intervalles. Ce n'est pas le cas de la courbe normale. L'espace compris entre les scores *Z* qui marque le commencement et la fin de l'intervalle I' est beaucoup plus grand que l'espace compris entre les scores qui limitent chacune des autres intervalles de la courbe normale. Le score *Z* qui marque la fin de l'intervalle I' est presque égal au score *Z* qui marque le commencement de l'intervalle I. Par contre, l'aire sous la courbe comprise dans l'intervalle I et l'intervalle I' est la même, tout comme l'est celle de chacun des intervalles réciproques de cette figure. Cet exemple montre que la normalisation donne lieu à un réarrangement important des scores qui marquent les intervalles, mais laisse intacte le nombre de personnes qui se trouvent dans cet intervalle. On peut ainsi comparer plusieurs indicateurs en étant assuré que le nombre de personnes à l'intérieur d'un intervalle et les scores qui le délimitent sont constants d'un indicateur à un autre.

Toutefois, on devra en échange poser l'hypothèse que ces habiletés sont distribuées de façon normale à l'intérieur de la population à l'étude et que si la forme de la distribution des scores bruts diffère de celle de la courbe normale, cette différence est attribuable soit à une mauvaise construction de l'épreuve, soit à une erreur de mesure.

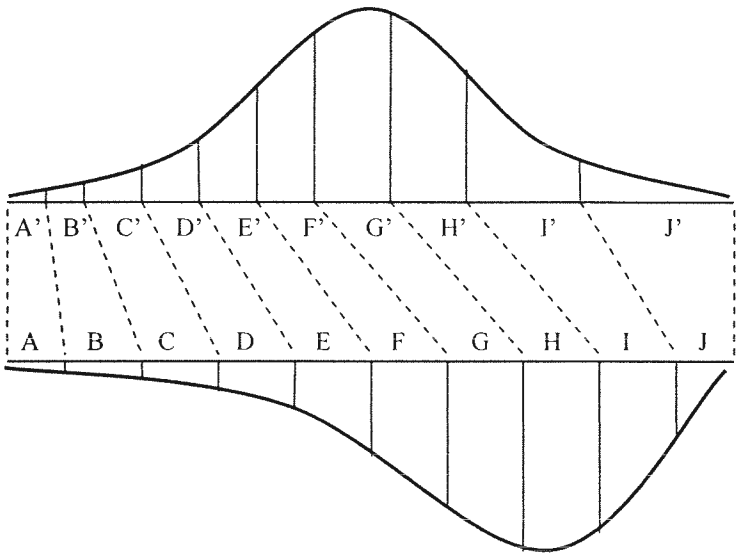

Figure 2.4 Effet du passage de l'échelle standardisée à l'échelle normale sur l'unité de mesure utilisée

Non seulement l'échelle normale permet la comparaison de tests variés et rend compte de l'écart réel entre les habiletés mentales des individus, elle permet aussi aux statisticiens d'utiliser les tests statistiques qu'ils désirent et d'obtenir des résultats plus fiables.

> Although normality of the variables is not always required for analysis, the solution is usually quite a bit better if the variables are all normally distributed. The solution is degraded, if the variables are not normally distributed, and particularly if they are nonnormal in very different ways (Tabachnick et Fidell, 2001, p. 73).

L'échelle normale est aussi particulièrement sensible au nombre de données qu'elle contient. Cette différence est d'autant plus marquée aux extrémités de la distribution. Pour une distribution de 10 cas, les scores centiles limites sont 95 et 5. Le score normalisé correspondant est de ± 1.64. Pour une distribution de 100 cas, les scores centiles limites sont 99.5 et 0.5 ce qui correspond à un score normalisé de ± 2.58. De plus, de petites différences de regroupement aux extrémités de la

distribution, différences qui peuvent être causées par divers degrés de dissymétrie ou d'aplatissement, auront un effet prononcé sur les données normalisées situées au extrémités de la distribution.

Par exemple, dans une distribution de 200 individus, si une personne obtient le plus haut score brut, son rang centile est de 99.75. Une fois la transformation effectuée, cet individu obtient un score normalisé de 2.81. Maintenant, si cinq personnes obtiennent le même score, le score centile de ce groupe de personnes sera de 98.75 et leur score normalisé de 2.24. De ce fait, Gulliksen (1962) nous met en garde quant à l'utilisation de l'échelle normalisée pour fin de comparaison :

> If normalized scores on different tests are to be compared, it is important to be sure that slight differences in grouping in extreme cases do not occur, and also to be certain that the groups are similar in size; otherwise the results reported for normalized scores will be influenced more by the size of the group and by slight differences in grouping in the extremes than by the abilities of the students. (p. 281)

Ces problèmes ne sont pas sans rappeler les problèmes inhérents à la comparaison de variables standardisées. La comparaison de données standardisées nécessitait une étendue et des indices d'asymétrie et d'aplatissement quasi identiques. La courbe normalisée élimine les contraintes liées à l'étendue de la distribution autour de la moyenne mais non celles liées à l'asymétrie et à l'aplatissement. Ainsi, on doit toujours prendre soin de ne comparer que des variables dont le nombre de cas est équivalent et dont la forme de la dispersion ne s'éloigne pas trop de la normalité.

Quant aux données extrêmes, Tabachnick et Fidell (2001) estiment que le fait de normaliser les données a pour conséquence d'amoindrir l'effet des données extrêmes. Comme 99.7 % des cas se trouvent entre $-3\ \sigma$ et $3\ \sigma$, les chances que plusieurs données aient un écart-type supérieur à $\pm$ 3.29 sont assez minces. Si le problème persiste, il est toujours possible de recourir à une modification de la valeur des données extrêmes afin de les ramener à 3.29 comme nous l'avons expliqué précédemment. Même si cette technique implique que les extrémités de la distribution

seront artificiellement gonflées, Tabachnick et Fidell (2001) affirment que leur effet sera tout de même moins important.

En résumé, l'avantage de la normalisation sur la standardisation repose sur sa capacité à comparer des indicateurs en utilisant une unité de mesure égale très proche de celle qui caractérise les mesures physiques tout en réduisant les contraintes liées à son utilisation. À moins d'avoir de bonnes raisons de croire que l'indice de dissymétrie ou d'aplatissement de la courbe d'origine soit le reflet d'une particularité de la population à l'étude et/ou qu'ils diffèrent fortement de la normale, les données brutes servant au calcul de scores composites ont tout avantage à être normalisées au préalable.

2.4 Niveau d'analyse et erreur écologique.

Les informations de nature normative facilitent la comparaison de résultats entre variables. Les normes fournissent une description statistique des performances d'un groupe à un ou plusieurs tests. Elles peuvent décrire la performance d'un groupe d'étudiants ou d'un regroupement de plusieurs groupes d'étudiants comme une école ou une commission scolaire par exemple. Mais comment peut-on utiliser les résultats des étudiants pour déterminer la performance d'une école et ensuite les comparer entre elles?

Bien qu'ils doivent être interprétés avec prudence, les résultats des étudiants peuvent être utilisés dans le but avoué de juger de la performance moyenne d'une école par rapport à celle d'une autre.

> Norms for school averages are constructed by sampling schools from a population of schools, administering the test to students in the schools, tabulating the average score for each school, and forming percentile ranks for the school averages. In this process, the school is considered the unit of analysis (Petersen, Kolen et Hoover, 1989, p. 238).

L'interprétation des informations normatives des écoles commande une certaine prudence puisqu'elles peuvent différer des normes utilisées pour expliquer

les scores des étudiants. La figure 2.5 révèle que le score moyen de l'école la plus performante est plus bas que le score de l'étudiant le plus performant de cette école. Quant au score moyen de l'école la moins performante, il est plus élevé que le résultat de l'étudiant le plus faible de cette école. Conséquemment, les moyennes d'écoles sont moins volatiles que les résultats des étudiants.

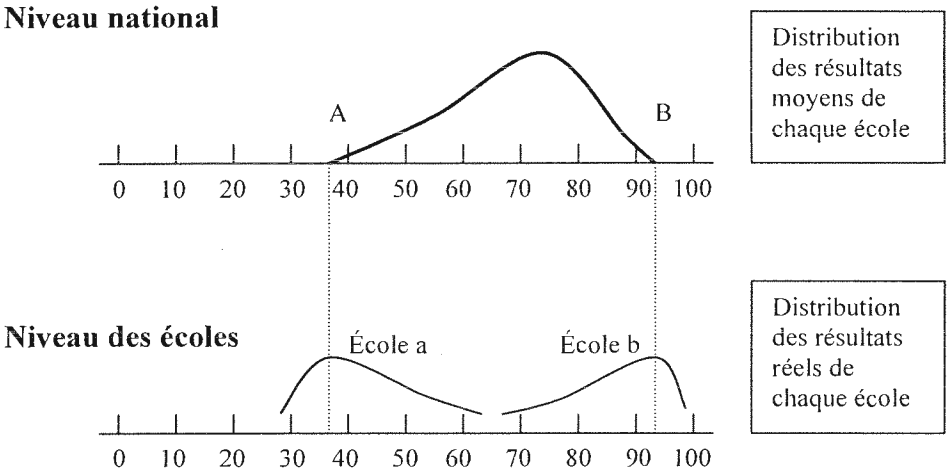

Figure 2.5 Effet du changement de niveau sur la distribution des résultats

Si la comparaison des résultats moyens des élèves à ceux de la moyenne nationale surestime ou sous-estime les résultats de certains étudiants par rapport à la moyenne nationale des écoles, les analyses comparatives entre écoles n'en sont pas affectées. Cette différence s'explique par le fait que la comparaison des scores moyens des écoles entre elles n'implique pas de changement de niveau d'analyse comme c'est le cas pour les comparaisons entre les résultats des étudiants à la moyenne nationale. Le problème que pose l'interprétation des données réside dans la

comparaison entre les deux niveaux et non dans l'utilisation des résultats moyens des étudiants pour comparer les écoles entre elles.

2.5 Conclusion.

Au cours de ce chapitre, nous avons montré pourquoi les données brutes en éducation ne peuvent, à elles seules, donner une image claire et précise de la performance des personnes ou des institutions qu'elles représentent. Pour comparer des indicateurs ensemble, les échelles de mesure utilisées pour chacun d'eux doivent être communes. Une échelle commune a la particularité d'être la même pour chaque indicateur et se doit également d'être une échelle d'intervalle. On suggère d'utiliser une échelle primaire et une échelle secondaire lorsque des données brutes sont transformées en valeurs d'une échelle commune.

Pour être considérée comme une échelle d'intervalle, l'échelle résultante de la transformation des données brutes doit satisfaire deux critères. Elle doit posséder une valeur zéro réelle et une unité de mesure égale. Seule l'échelle normale possède ces deux qualités. Le score centile utilise une unité de mesure basée sur la position d'un sujet par rapport à son groupe d'appartenance et ne reflète pas un changement sur une échelle d'habileté mentale. La normalisation des données brutes repose sur l'hypothèse que les données devraient être distribuées normalement au sein de la population de l'étude, en échange de quoi elle offrira une unité de mesure égale et sera moins contraignante pour l'évaluateur.

Le *Bulletin* utilise le score standardisé pour former sa cote globale. Les scores standardisés sont déterminés à partir de la différence d'habileté mentale entre les sujets et non pas sur leur rang au sein du groupe. Son unité de mesure n'est égale qu'en apparence et l'utilisation de cette transformation à des fins de comparaison entre indicateurs est soumise à diverses contraintes. Par contre, si les postulats qui régissent son utilisation sont respectés, elle pourrait théoriquement être considérée comme une échelle de mesure commune. Quant à l'utilisation des résultats moyens

des étudiants dans le but de comparer les écoles entres elles, cette pratique ne pose pas de problèmes particuliers tant que ces comparaisons se font entre données de même niveau.

Il sera donc important dans cette étude de déterminer quels sont les effets de la standardisation sur les cotes globales attribuées par le *Bulletin*, de déterminer également si les données brutes utilisées par le *Bulletin* respectent les critères d'utilisation proposés par Angoff (1971) – moyenne et distribution autour de la moyenne similaire – ou si le *Bulletin* doit utiliser la standardisation en posant l'hypothèse qu'elles le sont, comme le suggère Guilford et Fruchter (1978). Pour terminer, nous devrons vérifier s'il est possible de recourir à la normalisation pour certains indicateurs en s'assurant qu'ils passent le test de la normalité de Tabachnick et Fidell (2001) ainsi que l'effet de la normalisation de ces indicateurs sur les cotes globales attribuées.

CHAPITRE III

L'AGRÉGATION DES VARIABLES ET L'INTRODUCTION DE LA PONDÉRATION

Dans le chapitre II, nous avons défini ce qu'est une échelle de mesure commune et montré pourquoi les scores bruts en éducation doivent être modifiés si l'on désire utiliser une composite. Ensuite, nous avons analysé les effets de trois transformations : le rang centile, la standardisation et la normalisation et ce, en prenant soin de contraster les avantages et les inconvénients découlant de l'utilisation de chacune d'elles en particulier. Pour sa part, le présent chapitre traite de l'introduction de la pondération dans une évaluation de la performance des écoles.

Nous verrons aussi dans ce chapitre pourquoi l'agrégation de résultats requiert l'utilisation d'une échelle de mesure commune, comment on peut déterminer l'effet d'un changement de pondération sur la robustesse d'une composite, et quelle est la contribution réelle de chacune des variables aux scores composites.

3.1 Agrégation des résultats : l'importance de l'échelle commune.

Lorsqu'on décide d'agréger les résultats de plusieurs tests pour former une composite, il est essentiel que les scores utilisés soient égaux avant que la pondération ne soit introduite. En effet, si les scores utilisés ne sont pas égaux, on voit mal comment la pondération pourrait refléter l'importance qu'accorde un chercheur à chacune des variables utilisées puisqu'elle sera faussée, à la base, par la non équivalence des scores utilisés. C'est pourquoi le recours à une échelle de mesure commune revêt une importance capitale dans le processus de formation d'une composite.

Dans le passé, plusieurs chercheurs se sont penchés sur le problème de l'agrégation de variables (Stevens et Aleamoni, 1986; Terwilliger et Anderson, 1969;

Gardner et Erdle, 1984). L'étude de Stevens et Aleamoni (1986) démontre que la standardisation améliore la capacité du chercheur à comparer et à interpréter les résultats de l'agrégation de plusieurs scores. Pour y parvenir, ils analysent les effets de l'agrégation d'indicateurs avec et sans standardisation a priori. Voici la conclusion de leur étude :

> When aggregates are formed using raw scores, the standard deviations become weighting factors that determine the relative contribution of each component score in the aggregate. If standard deviation of all components are equal, then each component contributes equally to the aggregate score. Most commonly, however, the standard deviations of raw score components differ and the resultant aggregate score represents a weighted sum of the components. (p. 527)

Si la démonstration de Stevens et Aleamoni (1986) prouve empiriquement que l'utilisation de données brutes pondère chacun des tests par la valeur de son écart-type, un simple coup d'œil à l'équation qui permet le calcul du coefficient de corrélation suffit pour arriver au même résultat :

$$r_{xy} = \Sigma xy/N\, S_x S_y \qquad (4)$$

Où r_{xy} est la corrélation entre x et y, x et y sont les déviations des scores de x et y de la moyenne de x et y, Σxy est la somme des produits des déviations de x et y, N le nombre de cas utilisés et S_x et S_y sont les écarts types des scores x et y.

Le fait de standardiser les données utilisées stabilise le dénominateur utilisé dans l'équation 4. Si, par exemple, les données de x et de y sont standardisées de façon à obtenir une moyenne de 0 et un écart-type de 1, le dénominateur de chaque indice de corrélation calculé à partir de ces données sera égal au nombre de données N. C'est de cette façon qu'on arrive à retirer l'effet de l'écart-type, ou de l'étendue des données autour de la moyenne, de chacun des indicateurs lors du calcul de la corrélation entre deux tests. Cette pratique permet d'avoir le même dénominateur, peu importe les indicateurs utilisés pour la corrélation.

Maintenant, revenons à la démonstration empirique de Stevens et Aleamoni (1986). Dans le but de prouver leur hypothèse de départ, ils ont eu recours à l'exemple suivant. Le tableau 3.1 montre les résultats de 10 étudiants à 3 examens et à 1 pré-test. Les six premières colonnes présentent les résultats bruts et standardisés des élèves aux examens, alors que les trois dernières contiennent leurs scores agrégés. Le score agrégé brut (AB) est la somme des résultats obtenus par chaque étudiant à chacun des trois examens. Le score agrégé standardisé (AS) est la somme des résultats standardisés obtenus par ces mêmes étudiants aux mêmes épreuves. Le pré-test est utilisé comme critère extérieur (CE) dans le but de démontrer la relation entre les deux composites et une troisième mesure. Finalement, les scores agrégés standardisés pondérés (ASP) sont obtenus en multipliant chaque score AS d'un individu par l'écart-type des scores AB correspondant et en additionnant ces résultats pour les trois examens.

Tableau 3.1

Scores bruts, standardisés et agrégés de 10 étudiants à trois épreuves distinctes

	Examen 1		Examen 2		Examen 3		Pré-	Scores Agrégés		
Étudiant	Brut	Stand.	Brut	Stand.	Brut	Stand.	Test	AB	AS	ASP
A	39	37,2	46	51,7	14	29,1	25	99	118,0	544,394
B	40	38,9	40	31,6	18	35,8	19	98	106,3	533,954
C	50	56,0	46	51,7	28	52,7	24	124	160,4	794,134
D	45	47,4	47	55,0	29	54,4	27	121	156,8	763,858
E	53	61,1	40	31,6	26	49,3	32	119	142,0	743,657
F	48	52,6	48	58,4	32	59,4	28	128	170,4	834,042
G	53	61,1	46	51,7	30	56,1	32	129	168,9	844,080
H	42	42,3	47	55,0	26	49,3	21	115	146,6	703,831
I	41	40,6	47	55,0	32	59,4	23	120	155,0	753,796
J	54	62,8	48	58,4	29	54,4	32	131	175,6	863,960
Moyenne	46,5	50,0	45,5	50,0	26,4	50,0				
É.T.	5,84	10,0	2,99	10,0	5,93	10,0				

Le tableau 3.2 présente quant à lui les différents indices de corrélation entre les scores AB, AS et ASP et la mesure extérieure CE. La corrélation entre les scores

AB et AS est de +.9829. Comme prévu, le fait que les écarts-types des scores AB ne soient pas égaux fait en sorte que la corrélation entre ceux-ci et les scores AS est inférieure à +1.0. Cette statistique prouve hors de tout doute que les scores AB et AS ne sont pas égaux. On doit maintenant comprendre pourquoi.

Le tableau 3.2 montre également que la corrélation des scores AB et CE est de +.6667 et la corrélation entre les scores AS et CE est elle de +.6093. La différence entre les deux indices de corrélation vient renforcer la preuve à l'effet que les résultats de l'agrégation de données brutes et standardisées ne sont pas équivalents puisqu'une fois comparés à un critère externe, ils affichent des indices de corrélation inégaux. La question est de savoir si cette différence vient du fait que l'utilisation des données standardisées a pour effet de sous-estimer cette corrélation, ou si cette différence est attribuable au fait que l'écart-type des données brutes affecte la pondération lors de l'agrégation des données.

Tableau 3.2

Corrélation entre les 3 mesures agrégées et le critère extérieur

	2 Agrégés Standardisés (AS)	3 Critère Externe (CE)	4 Agrégés Standardisés Pondérés (ASP)
1. Agrégés Bruts (AB)	,9829	,6667	,9999
2. Agrégés Standardisés (AS)	-	,6093	,9830
3. Critère Externe (CE)	-	-	,6668

L'inspection des autres coefficients de corrélation du tableau 3.2 prouve que cette différence est le fruit de la seconde explication. Il est clair que les scores ASP sont équivalents aux scores AB (r = 0,99). Ce fait est aussi démontré par l'équivalence de corrélation successive des scores AS avec les scores AB (r = 0,98) et les scores ASP (r = 0,98). De plus, la relation de CE avec les scores AB et les scores

ASP est aussi équivalente (r = 0,67 et r = 0,67 respectivement). Par conséquent, la pondération des scores standardisés par l'écart-type des scores bruts produit une valeur agrégée équivalente à celle obtenue par la simple agrégation des scores bruts. On peut donc conclure que les scores standardisés assignent un poids équivalent à chaque composante d'un score composite alors que l'agrégation de scores bruts attribue un poids proportionnel à l'ampleur relative de l'écart-type de chaque composante. Ainsi, chaque fois que des scores doivent êtres agrégés, il importe que l'évaluateur transforme ses données afin d'obtenir une échelle de mesure commune s'il veut éviter que ses scores composites ne soient involontairement pondérés par l'écart-type des variables qu'il utilise.

3.2 Introduction de la pondération.

Maintenant que nous connaissons l'importance de l'utilisation d'une échelle commune lors de l'agrégation de données, nous pouvons passer au cœur du sujet de ce chapitre : la pondération. Une fois l'effet de l'écart-type des données retiré par l'entremise de la standardisation ou de la normalisation, l'évaluateur dispose de données égales et comparables. Il peut alors réintroduire des valeurs dans la composite par l'entremise de la pondération.

Il importe de ne pas confondre la pondération effective et la pondération nominale. Bien que l'on refuse d'accorder plus d'importance à une variable sur la base de l'étendue de ses scores autour de la moyenne ou d'autres caractéristiques, cela ne veut pas dire pour autant que chaque variable utilisée dans une valeur composite doit avoir une importance égale. Il est normal de penser que certaines variables soient plus importantes que d'autres quand vient le temps d'évaluer la performance d'un programme quelconque. Le poids que le chercheur décide de donner à chaque variable s'appelle le poids nominal. Quant au poids effectif, il représente le poids réel que chaque variable exerce sur les résultats de la composite.

Il est également important de comprendre en quoi consiste la « robustesse » d'une composite. *On dit d'une composite qu'elle est robuste lorsque les résultats qu'elle procure sont très peu affectés par des changements de pondération.* C'est-à-dire que les scores composites ne changent que très peu lorsque la pondération nominale des indicateurs utilisés est modifiée.

Pour comprendre l'effet de la pondération nominale sur les scores composites, ce chapitre est divisé en deux parties. Dans la première partie, nous analyserons les effets entraînés par un changement de pondération nominale. Nous verrons comment la standardisation et la normalisation des échelles de mesure augmente la robustesse d'une composite. Dans la seconde, nous verrons comment on peut déterminer l'effet réel, la pondération effective, de chacune des variables utilisées sur les scores composites. Nous montrerons que certaines particularités des variables utilisées pour former la composite peuvent avoir des effets indésirables sur la pondération effective. C'est-à-dire que le poids effectif peut différer du poids nominal peu importe la pondération ou l'échelle de mesure choisie. Nous déterminerons quelles sont ces caractéristiques et quelle importance relative elles ont sur le poids effectif.

3.2.1 Effet d'un changement de pondération nominale sur les résultats d'une composite

Selon Gulliksen (1962), le point de départ d'une analyse visant à déterminer la robustesse d'une composite devrait être le coefficient de distribution de Pearson. Le coefficient de distribution de Pearson représente le ratio de l'écart-type de la distribution d'une pondération sur la moyenne de cette distribution. Ainsi, il caractérise chacun des ensembles de poids considérés en prenant une valeur généralement comprise entre 0 et 1. Gulliksen (1962) soutient que si l'on considère seulement les ensembles de pondération dont les poids sont tous positifs – de sorte que le coefficient de Pearson soit plus petit que 1 et plus grand que 0, la corrélation entre les scores obtenus à l'aide de cette composite et ceux d'une composite utilisant des ensembles de poids nominal différents s'approche de 1 si l'intercorrélation

moyenne entre les indicateurs utilisés (r) et le nombre d'indicateurs utilisés pour former la composite (K) augmente.

Prenons par exemple le *Bulletin*. La composite qui donne naissance à la cote globale du *Bulletin* utilise la pondération suivante pour évaluer la performance des écoles secondaires : résultats aux épreuves = .4; taux d'échec = .2; taux de transition = .2; surestimation de l'école = .1; et écart entre les sexes = .1. Ainsi, la pondération utilisée par le *Bulletin* répond aux critères de Gulliksen (1962) mentionnés ci-dessus puisque ses poids sont tous positifs et que son coefficient de distribution de Pearson est entre 0 et 1 à .6125. C'est donc dire que pour que la corrélation entre la pondération présentement utilisée et une pondération alternative soit élevée, assurant du même coup la robustesse de la composite, l'intercorrélation moyenne entre les indicateurs utilisés et le nombre d'indicateurs combinés devra être élevée.

Le *Bulletin* n'est pas seul dans cette situation puisque Gulliksen (1962) confirme que la majorité des ensembles de poids utilisés pour pondérer les composites présentent ces caractéristiques. Dans le cas contraire, si le ratio de l'écart-type de la distribution de la pondération sur la moyenne de cette distribution est plus grand que 1, un changement de pondération amènera des changements importants au niveau des scores composites et ce, peu importe le nombre de variables combinées ou leur degré d'intercorrélation.

Pour illustrer cette situation, prenons l'exemple d'une valeur composite dont certaines variables seraient pondérées positivement et d'autres négativement. La moyenne de tels ensembles de pondération serait très près de 0. Du même coup, le coefficient de distribution de Pearson serait supérieur à 1 et un changement de pondération provoquerait des changements importants au niveau des scores composites, peu importe le nombre d'indicateurs utilisés ou leur degré d'intercorrélation.

Les travaux de Terwilliger et Anderson (1969), Fralick et Raju (1982) et Aamodt et Kimbrough (1985) ont par la suite confirmés les théories de Gulliksen. L'effet du nombre d'indicateurs utilisés pour former la composite (K) et de l'intercorrélation moyenne entre ces indicateurs (r) sur la valeur composite peut se résumer comme suit : si un nombre d'indicateurs importants est combiné (si K est entre 50 et 100) et que les scores sont fortement intercorrélés, l'ensemble de pondérations utilisées n'a que très peu d'influence sur les scores composites. La composite est alors robuste aux changements de pondération nominale. Si par contre un petit nombre d'indicateurs sont combinés (si K se situe entre 3 et 10 par exemple) et que l'intercorrélation moyenne est faible (.5 ou moins), l'ensemble de pondérations choisi aura un effet important sur les résultats obtenus avec la composite. La composite est alors peu robuste aux changements de pondération nominale.

En plus de s'intéresser à l'effet de la standardisation des données sur leur agrégation comme l'ont fait Stevens et Aleamoni (1986), Terwilliger et Anderson (1969) se sont attardés à l'effet de la standardisation des indicateurs sur les scores résultants de la composite. Pour ce faire, ils ont comparé les résultats de composites basés sur des scores bruts et des scores standardisés soumis à des combinaisons diverses d'ensembles de poids et des facteurs identifiés par les études de Gulliksen (1962): le nombre d'indicateurs utilisés (K) et l'indice d'intercorrélation moyen entre les différents indicateurs de la composite (r). Les résultats de cette étude démontrent que :

- Plus la différence entre les écarts-types des indicateurs utilisés est grande, plus l'impact de la standardisation des données est important.
- Plus l'indice de corrélation entre les indicateurs (r) est grand, moins la standardisation des données s'avère effective.
- Si les indicateurs sont positivement corrélés, l'effet de la standardisation s'amoindrit avec l'augmentation du nombre d'indicateurs (K).

Si l'on compare les résultats de Terwilliger et Anderson (1969) à ceux de Gulliksen (1962), on observe que plus les composites sont sujettes à être affectées par le choix d'un ensemble de poids, c'est-à-dire plus r et K sont petits, plus la standardisation des données s'avère un moyen efficace de s'assurer de l'authenticité de la variation engendrée par le changement de pondération. Autrement dit, les changements au niveau des scores composites doivent être causés par une modification de la pondération nominale et non par la disparité des écarts-types des indicateurs de la composite.

Alors que Gulliksen (1962) fait valoir que lorsque seulement entre 3 et 10 indicateurs sont combinés et que l'intercorrélation moyenne est de .5 ou moins, l'ensemble de poids choisi aura un effet important sur les résultats obtenus avec la composite (composite peu robuste), Terwilliger et Anderson (1969) constatent que si le nombre de critères est inférieur à 6 et que leur intercorrélation est inférieure à .4, la standardisation des données aura un effet appréciable sur la diminution de la disparité entre les scores composites obtenus à l'aide de pondérations nominales différentes.

Dans une pareille situation, la standardisation se veut une protection efficace contre les effets d'un changement de la pondération nominale sur les scores composites qui ne seraient pas attribuables à cette modification, mais plutôt au refus d'utiliser une échelle commune. La standardisation agit comme une police d'assurance en garantissant l'authenticité de la mesure de la robustesse d'une composite.

En somme, on peut dire que lorsque le coefficient de distribution de Pearson d'une composite formée de scores standardisés se situe entre 0 et 1, deux caractéristiques des variables utilisées peuvent influencer la robustesse d'une composite : le nombre de variables (K) et leur niveau d'intercorrélation moyen (r). Si, pour une raison quelconque, l'évaluateur ne transforme pas ses échelles brutes en échelles communes (standardisées ou normalisées), on doit ajouter à cette liste l'écart-type des variables utilisées dans le calcul des scores composites comme l'on

démontré Terwilliger et Anderson (1969). Plus la différence entre les écarts-types des indicateurs utilisés est grande, plus la robustesse de la composite est susceptible d'être affectée par l'intercorrélation moyenne entre les indicateurs et le nombre d'indicateurs utilisés.

3.2.2 Effet d'un indicateur sur la composite

Après avoir considéré l'utilisation de différents ensembles de poids et de la standardisation des données sur la corrélation des résultats d'une composite, nous allons maintenant expliquer l'effet d'un indicateur particulier sur les scores obtenus à l'aide d'une composite. La méthode utilisée pour mesurer l'effet précis d'un indicateur sur la composite repose sur l'analyse des informations que procure la régression multivariée. De ces informations, quatre sont nécessaires à la compréhension de ce qu'est la véritable contribution de chacune des variables indépendantes à la variable dépendante. C'est pourquoi nous passerons en revue : le coefficients β, le coefficient de variance entre les variables indépendantes et dépendantes (r^2), le coefficient de variance expliquée (R^2) et le coefficient de corrélation semipartiel au carré (sr^2).

Lorsqu'on effectue une régression, on tente de déterminer la grandeur relative d'un ou de plusieurs phénomènes (variables indépendantes) correspondant à un autre phénomène (variables dépendantes). Le choix de la variable dépendante servant à la régression multiple varie selon sa disponibilité. Quand l'évaluateur dispose d'une variable dépendante, cette dernière doit servir à la régression. Si, au contraire, l'évaluateur ne dispose d'aucune variable dépendante, comme c'est le cas lorsqu'on utilise une composite, le score composite s'avère être la meilleure alternative disponible et devra servir de variable dépendante à l'analyse de régression (Gulliksen, 1962). Dans un cas comme dans l'autre, c'est l'ensemble de variables critères qui servira de variables indépendantes à l'analyse statistique multivariée.

Par exemple, le salaire d'un diplômé à sa sortie de l'université pourrait prendre le rôle de variable dépendante utilisée pour valider la valeur prédictive d'un ou de plusieurs critères, tel la moyenne cumulative, le domaine d'étude ou la rapidité de cheminement des étudiants. Si on ne peut recourir à cet indicateur parce qu'il n'est pas disponible ou pas mesuré, la meilleure alternative restante consiste à utiliser le score composite comme variable indépendante dans le but de déterminer la contribution de chacune des variables indépendantes à la variable dépendante. Dans le cas du *Bulletin*, comme il n'existe pas de mesure connue de la performance des écoles secondaires, nous devrons utiliser les cotes globales comme variable dépendante pour la régression.

Effectuer une régression multivariée produit un coefficient β pour chacune des variables indépendantes utilisées. L'ensemble des coefficients β déterminés par la régression multiple maximise la corrélation entre les variables indépendantes et la variable dépendante en minimisant la somme des erreurs au carré des résidus. Ainsi, les coefficients β peuvent servir de coefficients de pondération lorsqu'une variable dépendante est disponible puisqu'ils maximisent la corrélation entre les variables indépendantes et la variable dépendante. Ces coefficients assurent l'utilisateur de la composite qu'il maximise la capacité explicative de son outil.

Par contre, tenter de déterminer l'impact d'un indicateur en l'absence d'une variable dépendante requiert un ajustement conceptuel important de la part de l'évaluateur. Sans variable dépendante, il devient inutile d'utiliser les coefficients β afin de mesurer l'importance qui devrait être accordée à chacune des variables. En fait, la proportion des coefficients β obtenue à l'aide d'une régression des variables indépendantes sur le score composite devrait être égale au poids nominal accordé par l'évaluateur à chaque variable. Dans ce cas, il est inutile de chercher à utiliser les coefficients β comme mesure de l'importance d'une variable sur la composite puisqu'ils reflètent l'importance qui leur a été attribuée au tout début par l'évaluateur. De plus, même lorsqu'une mesure du standard est utilisée comme variable

dépendante, les coefficients β ne constituent pas, à eux seuls, une mesure complète de l'importance d'une variable prédictrice.

Dans un cas comme dans l'autre, Wilks (1938) croit qu'une façon simple et efficace de déterminer l'importance d'une variable indépendante consiste à utiliser le coefficient de variance (r^2) entre les variables indépendantes et la variable dépendante comme un indice de sa contribution. Selon lui, le coefficient r^2 représente le véritable effet d'une variable sur les scores composites, en d'autres mots, son poids effectif.

La figure 3.2 montre l'importance de deux variables indépendantes (VI_1 et VI_2) par rapport à la variable dépendante (VD) lorsque le coefficient r^2 est utilisé comme mesure de la contribution et que les variables indépendantes ne sont pas intercorrélées. Ainsi, la partie *a* représente la variance de la VD qui ne peut être expliquée par VI_1 et VI_2, la partie *b* la variance de VD qui est expliquée par VI_1 et la partie *c* celle qui est expliquée par la VI_2.

Si une composite est formée de deux variables indépendantes non corrélées, la somme de *a* et *b* est égale à la variance totale expliquée (R^2) par les composantes. Si cette mesure semble, à première vue, être une excellente mesure de la contribution ou de l'importance d'une variable indépendante aux scores composites, la présence de corrélation entre les variables indépendantes fait ressortir la faiblesse du coefficient r^2. La figure 3.3 montre la zone représentée par les coefficients r^2 lorsqu'il y a corrélation entre VI_1 et VI_2.

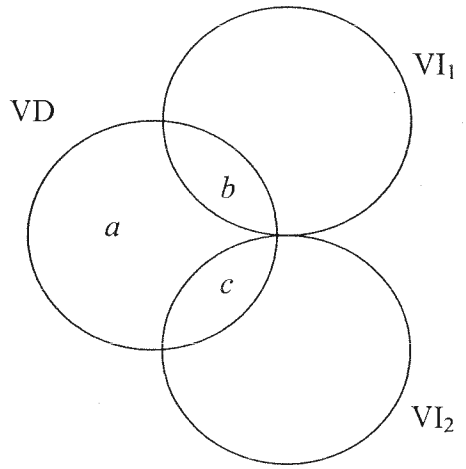

Figure 3.1 Proportion de la variance d'une variable dépendante (VD) expliquée par deux variables indépendantes (VI) en l'absence de corrélation entre les variables indépendantes

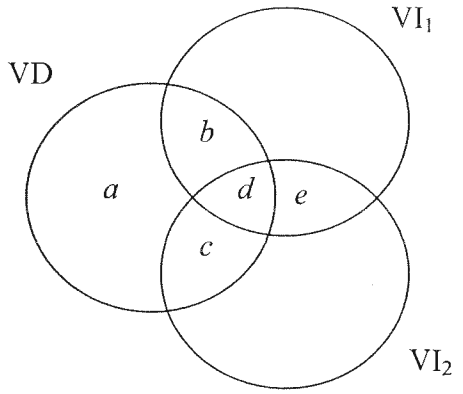

Figure 3.2 Proportion de la variance d'une variable dépendante (VD) expliquée par deux variables indépendantes (VI_1) en présence de corrélation entre les variables indépendantes

On constate qu'il y a présence de corrélation entre VI_1 et VI_2 à la figure 3.3. Cette corrélation est représentée par la zone formée de la partie *d* et *e*. Contrairement à l'exemple précédent, celui de la figure 3.3 montre que la variance expliquée par VI_1

et celle expliquée par VI_2 ne sont pas mutuellement exclusives. Ces deux variables indépendantes se partagent une partie de la variance totale expliquée ($b + c + d$), soit la partie d. Ainsi, la somme des r^2 de chaque variable indépendante n'est plus égale à R^2 lorsqu'il y a corrélation entre ces dernières. Cette juxtaposition des variances fait en sorte qu'il devient impossible pour l'évaluateur de déterminer la contribution unique de chacune des variables indépendantes sur les scores composites à partir du coefficient r^2 seulement.

Gulliksen (1962) montre concrètement les implications de la découverte de Wilks (1938). Il affirme que l'indice de corrélation entre la composite et l'un de ses indicateurs est déterminée par la formule suivante :

$$r_{gC}\, s_C = W_g\, s_g + \Sigma\, W_h\, r_{gh}\, s_h \qquad (5)$$

où r_{gC} est l'indice de corrélation entre le test g et la composite, W_g (ou W_h) est le poids assigné à l'indicateur, s_g (ou s_h) est l'écart-type de cet indicateur, r_{gh} intercorrélation entre les différents indicateurs et s_C l'écart-type de la composite. Comme s_C est identique pour chacun des indicateurs, il peut être ignoré.

Cette équation révèle que la corrélation entre une composite et une de ses parties est entièrement déterminée par son poids[7], son écart-type et l'intercorrélation de l'ensemble des parties. Ainsi :

- Si le poids attribué à un indicateur augmente, la corrélation de cet indicateur avec la composite augmentera elle aussi.
- Si la corrélation moyenne d'un indicateur avec les autres indicateurs augmente, sa corrélation avec la composite augmentera elle aussi.
- Si l'écart-type d'un indicateur augmente, sa corrélation avec la composite s'en verra augmentée elle aussi.

Le tableau 3.3 met en évidence les particularités inhérentes à la combinaison du poids et de l'écart-type sur la corrélation d'une partie avec la composite lorsque le degré d'intercorrélation entre les indicateurs est très élevé et que les indicateurs ne sont pas mesurés à l'aide d'une échelle commune.

Tableau 3.3

Effets de la combinaison du poids et de l'écart-type sur l'indice de corrélation

	Poids	Indice de corrélation	Écart-type
Indicateur 1	1/7	,86	7,7
Indicateur 2	2/7	,74	4,1
Indicateur 3	4/7	,48	2,5

(Tiré de l'exemple donné dans Stuit, 1947, p. 305-306)

Dans le tableau 3.3, on attribue respectivement aux indicateurs 1, 2 et 3 les poids 1/7, 2/7 et 4/7. Le choix de cette pondération montre que le chercheur attribue une grande importance à l'indicateur 3 puisque son poids est plus important que la somme des poids attribuée aux deux autres indicateurs. Pourtant, son indice de corrélation avec la composite n'est que de .48. C'est l'indicateur 1 qui, en dépit du fait qu'il affiche le poids le plus faible, présente la plus forte corrélation avec la composite soit .86. Ce phénomène s'explique par la différence des écarts-types de chacun de ces indicateurs. Ainsi, en dépit de son faible poids, l'importance de son écart-type réussit à elle seule à procurer à l'indicateur 1 le plus fort indice de corrélation. L'inverse est aussi vrai pour l'indicateur 3 dont la faiblesse de l'écart-type suffit à le priver de l'effet procuré par l'importance relative qui lui avait été

[7] Le poids d'une variable indépendante dans une composite est égale au coefficient β que l'on obtient lorsqu'on effectue une régression entre un score composite et ses composantes.

accordée et de sa moyenne élevée. De plus, non seulement la non standardisation des données influence les résultats du coefficient de corrélation, mais il en est de même pour chacun des quatre mesures de l'importance d'une variable. Cet inconvénient majeur contribue, lui aussi, à renforcer la preuve du besoin réel d'utiliser une échelle commune lors de la formation d'une composite.

Nous avons convenu au chapitre II qu'une des caractéristiques des échelles standardisées et normalisées est que leur moyenne est toujours égale à 0 et leur écarts-types à 1. Conséquemment, utiliser une échelle standardisée ou normalisée lors du calcul de la corrélation d'une partie avec la composite annule l'effet de l'écart-type dans l'équation. La corrélation qui en résulte n'est donc plus influencée par trois, mais bien par deux variables : le poids et le niveau d'intercorrélation. Par conséquent, si l'indice d'intercorrélation diminue, il en sera de même de l'influence des autres parties sur la partie qui nous intéresse. Si l'ensemble des indicateurs a une intercorrélation nulle et que l'effet de l'écart-type est retiré, l'effet de chaque indicateur sur la composite sera proportionnel à son poids.

Cette dernière constatation nous amène à démontrer que l'indice d'intercorrélation moyenne r est à l'origine d'un parallèle intéressant entre la corrélation de deux ensembles de poids et l'effet d'un indicateur sur la composite. En effet, alors qu'un fort indice d'intercorrélation entre les indicateurs était un gage de stabilité des résultats lors de la comparaison de résultats obtenus à l'aide de deux ensembles de poids différents, un indice de corrélation moyenne élevé masque la contribution marginale d'un indicateur sur la composite. On peut conclure que la stabilité des scores procurée par un r élevé à des composites pondérées différemment est attribuable au fait qu'il amoindrit l'effet relatif individuel de la pondération de chaque indicateur sur les résultats.

Dans le cas où on cherche à faire en sorte que les différentes combinaisons de poids influencent le moins possible les résultats de la composite, un indice d'intercorrélation élevé est un atout. Si, par contre, notre but est de déterminer avec

précision l'impact d'un seul indicateur sur l'ensemble de la composite, ce même indice s'avère être un inconvénient majeur car il complique la tâche de l'évaluateur qui souhaite que l'attribution d'un poids reflète davantage la contribution unique d'une partie à l'ensemble de la composite.

Afin de déterminer la contribution unique d'une variable, l'évaluateur peut avoir recours au coefficient de variance expliquée (R^2) ou au coefficient de corrélation semipartiel au carré (sr^2). La contribution d'une variable se définit comme la diminution de R^2 provoquée par le retrait de cette dernière de la composite et que l'on recalcule l'équation de régression uniquement avec les variables restantes. Prenons par exemple la figure 3.3. Si on considère que le R^2 de VI_1 et VI_2 est représenté par les régions *b*, *c* et *d*, la diminution de R^2 causée par le retrait de VI_1 serait égale à *b*. Ainsi, la contribution unique de VI_1 est égale à la région *b* et celle de VI_2 à *c*. Par conséquent, la région *d* est la variance de VD qui est partagée par VI_1 et VI_2.

Quant au coefficient de corrélation semipartiel au carré (sr^2), il nous indique précisément quelle est l'utilité d'une variable indépendante. Alors que nous devions effectuer une régression des variables indépendantes sur les scores composites, retirer la variable à l'étude, effectuer une seconde régression afin de trouver le nouveau coefficient R^2 et finalement faire la différence entre le premier coefficient et le second pour déterminer la contribution unique d'une variable, le sr^2 nous donne instantanément cette mesure. Ainsi, à la figure 3.3, le sr^2 de VI_1 est égal à la région *b* et celui de VI_2 à la région *c*.

En somme, lorsqu'une mesure du standard est disponible (variable dépendante), l'évaluateur doit commencer par analyser les β. Ces derniers renseignent l'évaluateur sur le poids à attribuer à chacune des variables pour maximiser l'ampleur de la variable dépendante expliquée par les variables indépendantes. Puisqu'un standard de comparaison est souvent manquant (les β sont

alors équivalents au poids nominaux), l'évaluateur peut avoir recours au r^2 pour avoir une mesure de la contribution d'une variable indépendante sur la composite.

Comme le montre l'équation 5, le coefficient r^2 d'une variable indépendante est influencé par sa corrélation avec les autres variables indépendantes. Ainsi, la contribution unique d'une variable peut être très différente de sa contribution d'ensemble et il est important que l'évaluateur analyse en détail les coefficients R^2 et sr^2 pour comprendre quelle est la proportion de la corrélation de chacune des variables qui lui est unique. Une fois ces coefficients analysés, l'évaluateur est en mesure de comprendre l'importance de chaque variable indépendante, l'effet des autres variables sur cette dernière ainsi que sa contribution unique aux scores composites.

En terminant, il a été souvent question au cours de ce chapitre des effets engendrés par l'adoption d'une échelle de mesure commune sur l'agrégation des données et la formation d'une composite. À chaque fois, l'échelle standardisée était citée en exemple. Ce qui ressort de ces analyses, c'est que premièrement, le fait de recourir à une échelle standardisée annule l'effet de pondération involontaire résultant de la disparité des écarts-types de chaque indicateur lors de l'agrégation des données et, deuxièmement, protège les scores composites de l'influence des écarts-types des indicateurs lorsque ceux-ci sont les plus susceptibles d'être affectés par un changement de pondération. Ces avantages de l'échelle standardisée découlent de sa capacité à uniformiser la moyenne et l'écart-type des indicateurs, particularité qu'elle partage avec l'échelle normale. Ainsi, la normalisation procure, dans le cas qui nous intéresse, les mêmes avantages que la standardisation.

3.3 Conclusion.

Pour conclure, on peut dire que la transformation d'une échelle brute en échelle commune est nécessaire à l'agrégation de données ainsi qu'à la juste pondération des indicateurs de la composite. Les travaux de Stevens et Aleamoni

(1986) ont prouvé que l'échelle commune procure à l'évaluateur la certitude que les indicateurs qu'il utilise ne seront pas involontairement pondérés par l'ampleur de leurs écarts-types en démontrant que les données standardisées et les données brutes ne sont pas corrélées de manière identique avec le critère externe et que cette différence est attribuable à la variation des écart-types des variables agrégées. Il sera intéressant de mesurer l'effet de la standardisation des données utilisée par le *Bulletin* sur la cote globale des écoles québécoises.

De leur côté, Terwilliger et Andersen (1969) ont étudié l'effet conjugué de ces facteurs et de la standardisation des données. En résumé, ils ont découvert que le coefficient de distribution de Pearson, l'intercorrélation moyenne entre les indicateurs utilisés (r) et le nombre d'indicateurs utilisés pour former la composite (K) sont trois facteurs qui affectent la robustesse d'une composite. Aussi, plus le changement de pondération est susceptible d'influencer les scores composites, plus la standardisation des données aide à diminuer cet effet. Reste à savoir si les données du *Bulletin* font en sorte que les cotes globales qui résultent de leur agrégation et de leur pondération sont robustes aux changements de pondération.

Finalement, Wilks (1938) identifie le coefficient bêta (β), le coefficient de variance entre les variables indépendantes et dépendantes (r^2), le coefficient de variance expliquée (R^2) et le coefficient de corrélation semipartiel au carré (sr^2) comme étant les quatre mesures nécessaires à la compréhension de la contribution entière et partielle d'une variable prédictrice sur les scores composites. La présente étude devra donc également tenter de déterminer, à l'aide de ces coefficients, quelle est la contribution de chacun des indicateurs utilisés par le *Bulletin* ainsi que leur effet sur la robustesse des cotes globales.

CHAPITRE IV

MÉTHODOLOGIE

La recension des écrits des trois chapitres précédents a révélé que la valeur composite est susceptible d'être influencée par trois types de changements : le nombre d'informations utilisées (chapitre I), le passage des données brutes à une échelle de mesure commune (chapitre II) et le changement de la pondération nominale par l'évaluateur (chapitre III). Le présent chapitre explique comment nous avons évalué empiriquement l'effet du nombre d'indicateurs utilisés, de la standardisation, de la normalisation et de la pondération sur la composite.

Nous avons utilisé une partie des données utilisées par le *Bulletin des écoles secondaires du Québec 2001* et par le MEQ dans son document intitulé *Résultats aux épreuves uniques de juin 2000 par commission scolaire et par école pour les secteurs public et privé et diplomation par commission scolaire* pour analyser les effets de chacune des transformations statistiques mentionnées ci-dessus sur la cote globale et le rang de 382 écoles québécoises.

4.1 Traitement des données

Le présent mémoire utilise les données brutes du *Bulletin* pour déterminer l'effet de l'ajout d'indicateurs de performance sur la composite, de la pondération nominale et de l'utilisation d'une échelle de mesure commune sur la cote globale des écoles québécoises. Bien que le *Bulletin* présente les résultats de 462 écoles dans son édition 2001, seulement 382 écoles ont été retenues pour cette analyse. Dans le but d'assurer la plus grande validité interne à notre étude, seules les écoles dont les données brutes étaient disponibles pour chacun des cinq indicateurs du *Bulletin* ont été utilisées dans la présente étude.

Les données brutes du *Bulletin* ont servi de point de départ à notre analyse. Les données brutes des indicateurs résultats aux examens (*resexam*), taux d'échec (*echec*), taux de transition (*trans*), surestimation des résultats par l'établissement (*surest*) et écart entre les sexes (*ecart*) proviennent intégralement de cette banque de données. Ces variables ont servi à la formation d'une première cote globale (*cg1*) en attribuant un poids égal de .2 à chacun des cinq indicateurs. Par la suite, une autre cote globale (*cg2*) a été créée à partir des mêmes indicateurs, mais cette fois pondérée à .4, .2, .2, .1 et .1 respectivement. Cette pondération est celle utilisée pour le calcul de la cote globale du *Bulletin*. Pour terminer, un rang a été attribué à chacune des écoles de l'échantillon selon leur résultat obtenu à *cg1* et *cg2* pour former les variables *rcg1* et *rcg2*.

Les données brutes ont ensuite été standardisées avec une moyenne de 50 et un écart-type de 10. Les variables standardisées (*zresexam*, *zechec*, *ztrans*, *zsurest* et *zecart*) ont servi au calcul de deux valeurs composites standardisées ayant des pondérations respectives égales à celles utilisées pour le calcul des cotes globales brutes, les variables *zcg1* et *zcg2*. Ces résultats ont ensuite été restandardisés pour produire les variables *zzcg1* et *zzcg2*. Un rang a alors été attribué aux écoles sur la base de ces cotes globales standardisées, créant du même coup les variables *rzzcg1* et *rzzcg2*.

Finalement, les données brutes ont été normalisées avec une moyenne de 50 et un écart-type de 10. Les variables normalisés *nresexam* et *ntrans* ont servi, avec les variables standardisées *zechec*, *zecart* et *zsurest*, au calcul des deux valeurs composites ayant des pondérations respectives égales à celles utilisées pour le calcul des cotes globales brutes, les variables *ncg1* et *ncg2*. Ces résultats ont aussi été restandardisés pour produire les variables *zncg1* et *zncg2*. Les écoles se sont vues attribuer un rang basé sur ces deux variables, créant ainsi les variables *rzncg1* et *rzncg2*.

4.2 Questions à résoudre

La recension des écrits des trois chapitres précédents nous a permis de comprendre quels étaient les grands enjeux de l'évaluation de la mesure de la performance des écoles secondaires au Québec. Plusieurs aspects de la mesure de la performance des écoles ont été touchés par cette recension et de nombreuses différences ont été notées. Principalement, les différences entre l'évaluation sommative de la performance des écoles du *Bulletin* et du MEQ concernent le nombre de variables utilisées pour classer les écoles. Par contre, la pondération nominale choisie par le *Bulletin* et l'échelle de mesure commune utilisée se sont également révélées être des aspects importants de la mesure de la performance des écoles.

4.2.1 Effet de l'ajout d'indicateurs sur la composite.

Le premier chapitre a fait ressortir la supériorité de la composite sur les critères multiples comme méthode d'analyse de données multiples (Nisbett et Ross, 1980). Nous y avons aussi mentionné que le *Bulletin* et le MEQ classent les écoles selon diverses méthodes. Le MEQ classe les écoles secondaires du Québec selon leur taux de réussite aux examens. Le *Bulletin* les classe à l'aide d'une cote globale qui comprend les cinq indicateurs de performance pondérés. Un de ces indicateurs, le taux d'échec, est, théoriquement, l'équivalent du taux de réussite utilisé par le MEQ pour classer les écoles. Nous avons vérifié l'indice de corrélation entre ces deux indicateurs afin de mesurer leur équivalence.

Si l'on ne peut affirmer qu'un classement basé sur un seul indicateur représente une véritable composite, on peut néanmoins se questionner sur l'effet réel de l'ajout de quatre autres indicateurs par le *Bulletin* et sur la différence de classement qu'il engendre. À ce sujet, Brogden et Taylor (1950) ont montré que l'ajout d'indicateurs servant à mesurer un même phénomène augmente la précision des résultats d'une composite, peu importe la direction ou la force de leur de la corrélation.

Afin de déterminer l'effet de l'ajout de quatre indicateurs pour classer les écoles, nous avons classé les écoles de notre échantillon selon leur taux de réussite du MEQ (*txreumin*). Ensuite, nous avons vérifié la moyenne et l'écart-type de la distribution de la valeur absolue de l'écart entre les rangs octroyés par le MEQ selon le taux de réussite (*rtxreumin*)et le *Bulletin* aux écoles (*rzzcg2*). Cet écart à été mesuré en comparant le classement du MEQ et de la cote globale du *Bulletin*. Finalement, une corrélation bivariée a été effectuée afin de déterminer, à l'aide du rho de Spearman, sa force ainsi que son niveau de signification.

4.2.2 Effet de la standardisation sur la composite

Le deuxième chapitre recense les écrits sur l'échelle de mesure commune et son utilisation dans une composite. Il ressort de cette analyse que seulement deux échelles de mesure procurent les qualités requises pour être considérées comme de véritables échelles d'intervalle : l'échelle standardisée et l'échelle normalisée. L'échelle normalisée est supérieure à l'échelle standardisée puisqu'elle est la seule échelle à posséder les caractéristiques d'une échelle commune : une valeur zéro réelle et une unité de mesure égale. Par contre, l'évaluateur ne peut recourir à l'échelle normalisée qu'à la condition de pouvoir poser l'hypothèse que l'écart de la distribution par rapport à la courbe normale est le résultat d'une erreur de mesure et non d'un phénomène réel (Angoff, 1971). L'évaluateur peut aussi décider d'utiliser l'échelle standardisée comme échelle de mesure commune s'il respecte les conditions qui régissent son utilisation (Guilford et Fruchter, 1978).

Pour vérifier l'hypothèse voulant que les scores standardisés soient plus fiables que les scores bruts, nous avons procédé à une analyse descriptive des données brutes pour déterminer quelles sont les différences d'écarts-types entre les variables utilisées par le *Bulletin*. Ensuite, nous avons analysé l'effet de la standardisation des données sur la pondération effective (r^2), la contribution unique (sr^2) ainsi que sur la distribution des valeurs absolues de la différence de rang

provoquée par le passage de l'échelle brute à l'échelle standardisée. Pour ces analyses, les cotes globales obtenues grâce à une pondération nominale égale des indicateurs (*cg1* et *zzcg1*) ont été utilisées afin de neutraliser les effet provoqués par la pondération des indicateurs. Une corrélation de Spearman a ensuite été effectuée afin de déterminer si les classements obtenus à l'aide de ces différentes échelles de mesure (*rcg1* et *rzzcg1*) sont corrélés.

4.2.3 Effet de la normalisation sur la composite.

Si Stevens et Aleamoni (1986) ont démontré que l'utilisation de l'échelle standardisée était préférable à celle de l'échelle brute, aucune étude ne semble indiquer de façon précise quel serait l'effet de la normalisation des données sur les résultats d'une composite. L'étude de l'effet de la normalisation sur la composite se divise en deux parties. La première vise à déterminer si les indicateurs utilisés dans la mesure de la performance scolaire peuvent être normalisés. Angoff (1971) insiste sur le fait que seul les variables ayant des distributions très proches de celles d'une courbe normale doivent être normalisées. Dans le cas contraire, il vaut mieux avoir recours à une autre échelle de mesure commune. L'indice d'asymétrie et d'aplatissement de chacun des indicateurs a été retenu afin de vérifier s'il passe le test de normalité proposé par Tabachnick et Fidell (2001). Ce test consiste à diviser l'indice d'asymétrie par son erreur standard. Si la valeur obtenue est inférieur à ± 3,29, l'indicateur sera considéré comme normal.

La deuxième consiste à mesurer l'effet de la normalisation des données sur la composite. Nous avons analysé l'effet de la normalisation des indicateurs identifiés lors de l'exercice précédent sur les coefficients r^2, sr^2 ainsi que sur la distribution de la valeur absolue des différences entre les cotes globales et les rangs provoquées par le passage de l'échelle standardisée à l'échelle normalisée. Les cotes globales obtenues grâce à des poids égaux (*zzcg1* et *zncg1*) ont été utilisées afin de neutraliser les effets engendrés par la pondération des indicateurs. Finalement, une corrélation de Spearman a été effectuée entre les cotes globales obtenues à l'aide des données

standardisées (*rzzcg1*) et normalisées (*rzncg1*) afin de déterminer la force de cette corrélation ainsi que son niveau de signification.

4.2.4 Effet d'un changement de pondération sur la composite

Pour mesurer l'effet de la pondération sur la cote globale des écoles, nous avons analysé les coefficients r^2 et sr^2, les cotes globales ainsi que le classement des écoles obtenu à l'aide de la pondération unique (*zzcg1* et *rzzcg1*), pour ensuite comparer ces résultats à ceux obtenus à l'aide de la pondération utilisée pour le *Bulletin* (*zzcg2* et *rzzcg2*). La méthode utilisée pour évaluer l'effet de ce changement de pondération consiste à analyser l'indice de pondération effective (r^2) de chacun des indicateurs et ensuite leurs contributions marginales (sr^2) afin de comprendre l'apport global de chaque variable aux scores composites. Nous avons également comparé ces valeurs aux pondérations nominales de chaque indicateur et nous avons comparé les différences de rangs et de cotes globales. Finalement, comme dans les cas précédents, une corrélation de Spearman entre les cotes globales obtenues à l'aide de la pondération équivalente (*rzzcg1*) et celle du *Bulletin* (*rzzcg2*) a été effectuée afin de déterminer la force de cette corrélation ainsi que son niveau de signification.

CHAPITRE V

RÉSULTATS ET DISCUSSION

Le chapitre précédent explique comment, dans un premier temps, les données de base ont été transformées en données standardisées et normalisées, et comment, à partir de ces données, nous avons calculé la cote globale et le rang des écoles faisant parties de notre échantillon. Dans un second, nous avons décrit les analyses nécessaires à la vérification de l'effet provoqué par les différents changements proposés et discutés au cours de la recension des écrits.

Le présent chapitre présente et commente les résultats des analyses effectuées pour comprendre les effets liés à l'ajout d'indicateurs, à la standardisation, à la normalisation des données et au changement de pondération sur l'évaluation sommative de la performance des écoles secondaires québécoises.

5.1 Effet de l'ajout d'indicateurs sur le classement des écoles.

Le premier chapitre a fait ressortir la supériorité de la composite sur les critères multiples comme méthode d'analyse de données multiples (Nisbett et Ross, 1980) ainsi que l'augmentation de la précision d'une composite qui accompagne l'ajout d'indicateurs servant à mesurer un même phénomène (Brogden et Taylor, 1950). On y note aussi que le taux de réussite, indicateur utilisé par le MEQ pour classer les écoles, et le taux d'échec, indicateur utilisé par le *Bulletin* pour former la cote globale servant au classement des écoles, sont quasi-identiques. Finalement, on s'interroge, dans ce chapitre, sur l'effet engendré par l'utilisation de quatre indicateurs supplémentaires pour former une cote globale et classer les écoles à partir de cette dernière.

Afin de montrer que le *taux de réussite* du MEQ et le *taux d'échec* du *Bulletin* sont des indicateurs quasi-identiques, nous avons effectué une analyse de corrélation entre les résultats des 382 écoles à la variable taux de réussite du ministère (*txreumin*) et la variable taux d'échec du *Bulletin* (*echec*). Le taux de réussite du MEQ qui sert au classement des écoles secondaires du Québec est très fortement corrélé avec le taux d'échec utilisé pour le calcul de la cote globale du *Bulletin.* ($r = .941$, $p < .01$). Cette forte corrélation vient renforcer l'hypothèse selon laquelle ces deux indicateurs évaluent le même aspect de la performance scolaire.

Bien qu'élevée, la corrélation entre *txreumin* et *echec* n'est pas parfaite. Nous croyons que la différence entre les deux indicateurs est attribuable aux opérations statistiques, telles la modération et la majoration des résultats exécutées par le MEQ. Ces opérations statistiques majorent la note des étudiants dont les résultats à l'épreuve du MEQ se sont avérés être moins bons que ceux obtenus pendant l'année scolaire et modèrent celle des étudiants dont les résultats à l'épreuve du MEQ sont supérieurs à ceux qu'ils ont obtenus au cours de l'année scolaire[8]. Cette transformation aurait pour effet de modifier légèrement les données présentées par le MEQ, de sorte que leur corrélation avec les données brutes utilisées par le *Bulletin* n'est pas parfaite.

L'effet de l'ajout d'indicateurs sur le classement des écoles a été évalué à partir des différences entre le classement du MEQ et celui du *Bulletin*. La différence de rang entre le premier classement des 382 écoles de notre échantillon, classement qui est déterminé exclusivement sur la base de la performance des écoles à l'indicateur du taux de réussite du MEQ (*txreumin*), et le second classement, qui est le résultat de la standardisation, de l'agrégation et de la pondération des cinq indicateurs du *Bulletin* (zcg2) est de 29,5 rangs en moyenne (ÉT = 34,88).

La différence de rang s'explique par la précision accrue que procure l'utilisation d'un nombre plus important d'indicateurs de performance comme le suggère Brogden et Taylor (1950). Malgré la différence de rang provoquée par l'ajout de quatre indicateurs, le classement des écoles secondaires du MEQ et celui du *Bulletin* est fortement corrélé, comme en témoigne la corrélation de Spearman de .915 ($p < .01$) entre la variable *txreumin* et la variable *zcg2*.

On peut déduire de ces analyses que les résultats de la composite *zcg2* diffèrent de ceux obtenus à l'aide de l'indicateur *txreumin* et que les indicateurs additionnels fournissent des informations qui mènent à une évaluation plus juste de la performance des écoles québécoises. Néanmoins, l'indice de corrélation trouvé indique que les classements demeurent tout de même fortement corrélés.

5.2 Effet de la standardisation sur la composite.

Dans le second chapitre, les travaux de Stevens et Aleamoni (1986) ont montré que l'agrégation de données brutes pondère chacun des indicateurs utilisés par son écart-type. Nous avons conclu que la standardisation (ou la normalisation) des données est nécessaire lors de la formation de composite. Ces opérations statistiques améliorent l'adéquation entre la pondération nominale et la pondération effective imposant à chaque indicateur une distribution ayant une moyenne et un écart-type fixes.

Les données brutes utilisées pour former la cote globale du *Bulletin* ont des écarts-types qui varient considérablement. Le tableau 5.1 montre que l'indicateur *trans* a le plus important écart-type (14,899), suivi des indicateurs *echec* (8,199), *resexam* (5,597), *surest* (3,793) et *ecart* (1,847).

---

[8] Pour plus de détails concernant ces opérations statistiques, voir le document intitulé *Résultats aux épreuves uniques de juin 2000 par commission scolaire et par école pour les secteurs public et privé et diplomation par commission scolaire* produit par le MEQ

Tableau 5.1

Analyse descriptive des indicateurs de performance de la cote globale *cg1*

	resexam	*echec*	*trans*	*surest*	*ecart*
Moyenne	74,004	13,701	74,982	6,211	3,936
Écart-type	5,597	8,199	14,899	3,793	1,847
Asymétrie	,236	,797	-,399	3,172	,672
Erreur standard d'asymétrie	,125	,125	,125	,125	,125
Aplatissement	,837	1,322	,624	15,045	,911
Erreur standard d'aplatissement	,249	,249	,249	,249	,249
Minimum	56,670	,000	18,994	1,518	,300
Maximum	90,344	46,954	100,00	35,091	11,250

À l'état brut, les pondérations effectives (r^2) des indicateurs ne sont pas égales aux pondérations nominales qui leur ont été attribuées. Le tableau 5.2 montre que même si chaque indicateur a un poids nominal de .20, trois d'entre eux présentent des pondérations effectives largement supérieures à leur pondération nominale. L'indicateur *resexam* a un r^2 de ,789, l'indicateur *echec* un r^2 de ,781 et l'indicateur *trans* un r^2 de ,831. Le tableau 5.2 montre aussi que deux des cinq indicateurs ont des pondérations effectives inférieures à leur pondération nominale. L'indicateur *surest* a un r^2 de seulement ,0004 et l'indicateur *ecart* un r^2 de ,039.

La différence entre les poids effectifs des indicateurs bruts qui forment la cote globale est en partie liée à la différence de leur écart-type. Plus l'écart-type d'un indicateur est important, plus sa pondération effective est importante. Ainsi, le tableau 5.2 montre que les trois indicateurs bruts ayant les pondérations effectives les plus élevées (*resexam*, *echec* et *trans*) affichent aussi les écarts-types les plus importants (5,597, 8,199 et 14,899) des cinq indicateurs qui forment la cote globale. À l'inverse, les deux indicateurs bruts ayant les pondérations effectives les plus faibles (*ecart* et *surest*) affichent les plus petits écarts-types (3,793 et 1,847).

Tableau 5.2

Effets de la standardisation sur l'écart-type (ÉT), la pondération effective (r^2) et la contribution unique (sr^2) des indicateurs de la cote globale

	Bruts			Standardisés		
	ÉT	r^2	sr^2	ÉT	r^2	sr^2
resexam	5,597	,789	,004	1,000	,721	,009
echec	8,199	,781	,010	1,000	,743	,010
trans	14,899	,831	,167	1,000	,582	,055
ecart	3,793	,0004	,019	1,000	,034	,094
surest	1,847	,039	,005	1,000	,181	,102
somme			,205			,270

La standardisation atténue l'effet de la pondération par l'écart-type sans toutefois éliminer complètement les différences de pondération effective entre les indicateurs de la cote globale. Dans le tableau 5.2, la pondération effective des indicateurs qui avaient les plus grands écarts-types (*resexam*, *echec* et *trans*) a diminué, et celle des indicateurs qui avaient les plus petits écarts-types (*ecart* et *surest*) a augmenté suite à la standardisation des données. L'indicateur qui présente le plus grand écart-type, l'indicateur *trans* (ÉT = 14,899) a vu sa pondération effective passer de ,831 avant la standardisation à ,582 après la standardisation. Celle de l'indicateur ayant le plus petit écart-type à l'état brut, l'indicateur *surest* (ÉT = 1,847), est passée de ,0004, avant la standardisation, à ,034 après celle-ci. Enfin, la pondération effective des indicateurs *ecart* (ÉT = 3,793), *resexam* (ÉT = 5,597) et *echec* (ÉT = 8,199), dont les écarts-types étaient plus près de la moyenne, passent respectivement de ,039 à ,181, de ,789 à ,721 et de ,781 à ,743.

Selon Stevens et Aleamoni (1986), la standardisation des indicateurs devrait annuler l'effet de l'écart-type sur la pondération effective. Or, nous avons constaté qu'une fois standardisés, les indicateurs utilisés pour former la cote globale du *Bulletin* présentent toujours des écarts importants entre leurs poids effectifs et nominaux. L'équation de Gulliksen (1962) présentée au chapitre III (voir p. 64)

explique la différence persistante entre les poids nominaux et effectifs des indicateurs de performance après standardisation :

$$r_{gC}\, s_C = W_g\, s_g + \Sigma\, W_h\, r_{gh}\, s_h \qquad (5)$$

Avec l'équation (5), Gulliksen identifie 3 facteurs susceptibles d'influencer l'indice de corrélation entre le résultat d'une composite et une de ces composantes : la pondération nominale (W), l'écart-type de chaque indicateur (s) et l'intercorrélation (r) des indicateurs de performance. Si on peut contrôler la pondération nominale des indicateurs en modifiant le poids nominal qu'on lui attribue et l'effet de l'écart-type en standardisant ou en normalisant les données utilisées, l'intercorrélation des indicateurs demeure hors de notre contrôle.

L'intercorrélation des indicateurs utilisés pour former la cote globale fait en sorte que même avec des poids et des écarts-types égaux, les pondérations effectives des indicateurs standardisés ne sont pas identiques dans le tableau 5.2. Par exemple, selon l'équation (5), le poids effectif de l'indicateur *resexam* est composé de son propre poids nominal (W_g) et de son écart-type (s_g), à quoi s'ajoute la somme du produit du poids nominal, de l'écart-type et de la corrélation entre l'indicateur *resexam* et les quatre autres indicateurs qui forment la cote globale ($\Sigma\, W_h\, r_{gh}\, s_h$). Comme les poids nominaux et les écarts-types sont les mêmes pour tous les indicateurs qui forment la composite, les coefficients de corrélation jouent un rôle déterminant dans le calcul de la pondération effective d'un indicateur. En somme, plus un indicateur est fortement corrélé avec les autres indicateurs de performance, plus son poids effectif est important.

Le tableau 5.3 montre que trois des cinq indicateurs qui forment la composite, les indicateurs *resexam*, *echec* et *trans*, sont fortement corrélés entre eux. Les deux autres indicateurs, *surest* et *ecart*, sont faiblement corrélés entre eux et avec les trois autres indicateurs de la composite. Ainsi, la forte intercorrélation des trois premiers

indicateurs (*resexam*, *echec* et *trans*) leur confère des poids effectifs plus importants, même une fois standardisés et pondérés de manière identique.

Tableau 5.3

Matrice de corrélation des cinq indicateurs de la cote globale

	resexam	*echec*	*trans*	*surest*	*ecart*
resexam	1,000				
echec	-,945*	1,000			
trans	,693*	-,653*	1,000		
surest	,168*	-,072	,106	1,000	
ecart	-,149*	,135*	-,115	-,084	1,000

* Corrélation significative à .01

Au chapitre III, nous avons également montré, à l'aide des diagrammes de Venne, que plus les indicateurs sont corrélés, plus grande sera leur contribution partagée et qu'à l'opposé, moins les indicateurs d'une composite sont corrélés, moins grande sera leur contribution partagée. La contribution unique (sr^2) des indicateurs de performance *resexam*, *echec* est masquée par leur forte corrélation. Dans le tableau 5.2, les changements des contributions uniques observés suite à la standardisation des données sont minimes. Avant la standardisation, la contribution unique de *resexam* est de ,004 et celle de *echec* de ,010. Une fois standardisées, la contribution unique de *zresexam* a augmenté de 0,5 % pour atteindre ,009 et celle de *zechec* demeure inchangée. La contribution unique (sr^2) des indicateurs de performance *ecart* et *surest* bénéficie de l'annulation de la pondération artificielle par l'écart-type et voient leurs contribution unique augmenter. La contribution unique de *ecart* est de ,019 avant la standardisation et celle de *surest* de ,005. Une fois standardisées, la contribution unique de *zecart* a augmenté de 7,5 % pour atteindre ,094 et celle de *zsurest* de 9,7 % pour atteindre ,102. Finalement, la contribution unique de l'indicateur *trans* voit sa contribution unique diminuer, passant de ,167 à ,055, une baisse de 11,2 %.

Pour l'ensemble des indicateurs de la composite, le passage de l'échelle brute à l'échelle standardisée se traduit par une augmentation de la somme des contributions uniques ($\sum sr^2$). De 20,5 % qu'elle était avec les scores bruts, elle passe à 27 % avec les scores standardisés. Théoriquement, plus la somme des contributions uniques augmente, plus la cote globale sera affectée par un changement de pondération. Une somme des contributions uniques élevée indique que les indicateurs sont faiblement corrélés. Dans une telle situation, la pondération nominale utilisée a un effet sur une plus grande partie de la variance expliquée par un indicateur de la composite. C'est ainsi que la standardisation diminue, toute chose étant égale par ailleurs, la robustesse de la composite.

Pour mesurer l'effet de la standardisation sur les scores composites, les cotes globales obtenues par l'agrégation des résultats bruts ont été comparées à celles obtenues grâce aux indicateurs standardisés. La standardisation procure un gain de précision moyen de 28,3 rangs (ET = 29,85) à la cote globale. L'utilisation d'une échelle de mesure commune apporte un gain de précision non négligeable à l'évaluation de la performance des écoles. Le rho de Spearman indique une forte corrélation entre le classement obtenu à l'aide des données brutes et celui obtenu à l'aide des données standardisées. La cote globale obtenue à l'aide des scores bruts (*cg1*) et celle obtenue à partir des scores standardisés (*zzcg1*) a un coefficient de corrélation de ,939 ($p < .01$).

En somme, la standardisation des indicateurs de la cote globale atténue la différence entre la pondération nominale et la pondération effective des indicateurs de la cote globale du *Bulletin*, diminue la robustesse de composite et améliore le précision des résultats.

5.3 Effets de la normalisation sur la composite.

Les différentes études recensées traitant de l'effet de la standardisation des données sur l'agrégation et la pondération des données formant une composite

(Stevens et Aleamoni, 1986; Terwilliger et Anderson, 1969; Gardner et Erdle, 1984) ont toutes démontré l'importance de l'utilisation d'une échelle de mesure commune. Cependant, aucune d'entre elles ne s'est penchée sur l'effet de l'échelle normale sur la composite.

Au chapitre II, nous avons montré que l'échelle normale est la seule échelle qui possède à la fois une unité de mesure fixe et une valeur zéro qui représente une réelle absence de la caractéristique mesurée (Angoff, 1971). Pourtant, ces deux caractéristiques de l'échelle normale sont des éléments essentiels à l'analyse mathématique ou statistique de données numériques (Michell, 1986). La présente section a pour but de mesurer l'amélioration qui résulte de l'utilisation de l'échelle normale au lieu de l'échelle standardisée. Une fois les effets associés à la standardisation des données montrés, l'effet de la normalisation des données a été mesuré. Comme mentionné au chapitre précédent, cette analyse comporte deux volets. Le premier concerne la nécessité de recourir à la normalisation alors que le second analyse les effets empiriques de cette transformation.

Tableau 5.4

Résultats des tests de normalité

	Asymétrie	Erreur standard d'asymétrie	Résultats du test de normalité	Aplatissement	Erreur standard d'aplatissement	Résultats du test de normalité
zresexam	,236	,125	1,89*	,837	,249	3,36
zechec	,797	,125	6,38	1,322	,249	5,31
ztrans	-,399	,125	-3,19*	,624	,249	2,51*
zsurest	3,172	,125	29,70	15,045	,249	60,42
zecart	,672	,125	5,38	,911	,249	3,66

* Résultats inférieurs au seuil de 3,29

Pour que la normalisation soit considérée, la courbe d'origine doit être quasi normale de sorte que sa déviation puisse être attribuable à une erreur de mesure et non à un phénomène réel (Cronbach,1990 et Angoff, 1971). Dans ce cas, nous pouvons poser l'hypothèse que l'habileté mesurée est distribuée de façon normale

dans la population à l'étude. Le tableau 5.4 présente les résultats des tests de normalité effectués pour déterminer quels indicateurs peuvent être normalisés. Les résultats sont à l'effet que seuls les indicateurs *zresexam* et *ztrans* peuvent être transformés. Les autres indicateurs présentent des distributions dont la forme s'éloigne de la normale et ne passent pas le test de normalité proposé par Tabachnick et Fidell (2001). Pour trois des cinq indicateurs (*zechec*, *zsurest* et *zecart*), il est difficile d'affirmer que l'habileté mesurée est distribuée de façon normale dans la population à l'étude et que les déviations de la distribution par rapport à la courbe normale peuvent être le résultat d'erreurs de mesure.

Tableau 5.5

Effet de la normalisation sur la pondération effective (r^2) et la contribution unique (sr^2) des indicateurs de la cote globale

	Standardisée			Normalisée et Standardisée	
	r^2	sr^2		r^2	sr^2
zresexam	,721	,009	*nresexam*	,722	,010
zechec	,743	,010	*zechec*	,740	,011
ztrans	,582	,055	*ntrans*	,576	,060
zsurest	,034	,094	*zsurest*	,034	,096
zecart	,181	,102	*zecart*	,181	,101
somme		,270			,278

Une fois les indicateurs *resexam* et *trans* normalisés, les variables *nresexam* et *ntrans* ont été utilisées pour former une nouvelle cote globale *zncg1*. Le tableau 5.5 montre l'effet de cette transformation sur la pondération effective (r^2) et la contribution unique (sr^2) des indicateurs de la cote globale. La normalisation des échelles de mesure des indicateurs *resexam* et *trans* n'a pas entraîné de changements importants pour les indicateurs de la cote globale.

Dans le tableau 5.5, la pondération effective (r^2) des indicateurs *zsurest* (,034) et *zecart* (,181) n'est pas affectée par la normalisation des indicateurs *ztrans* et *zresexam*. La pondération effective de l'indicateur *zechec* a diminué, passant de ,743 à ,740. Finalement, les pondérations effectives des deux indicateurs de performance transformés, soit les indicateurs *nresexam* et *ntrans*, ont pris des directions opposées. La pondération effective de *nresexam* a légèrement augmenté, passant de ,721 à ,722 et celle de *ntrans* a diminué, passant de ,582 à ,576.

Tableau 5.6

Matrice de corrélation des cinq indicateurs de performance de la cote globale *zncg1*

	nresexam	*zechec*	*ntrans*	*zsurest*	*zecart*
nresexam	1,000				
zechec	-,944*	1,000			
ntrans	,678*	-,639*	1,000		
zsurest	,152*	-,072	,116	1,000	
zecart	-,142*	,135	-,122*	-,084	1,000

* Corrélation significative à .01

La faiblesse de l'effet de la normalisation des indicateurs *resexam* et *trans* sur la pondération effective (r^2) des indicateurs de performance s'explique par le maigre effet de la normalisation sur la corrélation entre les indicateurs de la composite, comme en font fois les résultats du tableau 5.6. En effet, les valeurs obtenues dans les tableaux 5.6 et 5.3 sont quasi-identiques. Comme nous l'avons vu avec l'équation (5) ci-dessus, trois facteurs influencent le poids effectif d'un indicateur en particulier: son poids nominal, son écart-type et sa corrélation avec les autres indicateurs. Dans le cas présent, les poids nominaux (.2) et les écarts-types (10) sont demeurés identiques et les coefficients de corrélation ont très peu changé. C'est pourquoi, l'effet de la normalisation sur la pondération effective des indicateurs formant la composite est faible.

Quant à la contribution unique (sr^2) des cinq variables de la cote globale, on constate que la contribution unique des indicateurs de performance *nresexam*, *zechec*, *ntrans* et *zsurest* augmente dans le tableau 5.5 suite à la normalisation des variables *zresexam* et *ztrans*. Le coefficient sr^2 de *nresexam* et *zechec* augmente de ,001, celui de *ntrans* de ,005 et celui de *zsurest* de ,002. On attribue cette hausse à l'augmentation de l'intercorrélation moyenne de ces variables avec les autres variables observées dans le tableau 5.6, suite à la normalisation des variables *resexam* et *trans*. Finalement, la contribution unique de la variable *zecart* voit sa contribution unique diminuer de ,001 suite à la normalisation.

Le chapitre III montre que plus les variables sont corrélées, plus grande sera leur contribution partagée. À l'opposé, moins les variables sont corrélées, plus leur contribution unique sera grande. Le tableau 5.6 montre qu'en dépit de la normalisation des indicateurs *trans* et *resexam*, deux des cinq indicateurs de performance de la cote globale sont toujours fortement corrélés (*nresexam* et *zechec*), et que l'indicateur *ntrans* est moyennement corrélé avec ceux-ci. Finalement, les indicateurs *zecart* et *zsurest* sont faiblement corrélés entre eux et avec les trois autres indicateurs de la composite. Pour l'ensemble des variables indépendantes, le passage de l'échelle standardisée à l'échelle normalisée fait augmenter la somme des contributions uniques. Dans le tableau 5.5, elle est 27,8 % avec les scores normalisés alors qu'elle était de 27 % avec les scores standardisés. La pondération nominale utilisée a un effet sur une plus grande partie de la variance de variables indépendantes, faisant en sorte que la normalisation diminue la robustesse de la composite.

La différence de cote globale et de rang des écoles évaluées est quasi imperceptible. La différence moyenne au niveau de la cote globale est de ,59 (ET = .82). Quant aux effets de la normalisation sur les rangs octroyés aux écoles, ils ne sont pas beaucoup plus importants avec un changement moyen de 4,8 rang (ET = 5.99). La corrélation de Spearman entre les rangs attribués en utilisant les données

standardisées et ceux attribués à l'aide des données standardisées et normalisées est de .979 ($p < .01$).

En somme, la normalisation des indicateurs *zresexam* et *ztrans* modifie de façon marginale la pondération effective et la contribution unique des indicateurs de la cote globale du *Bulletin*. Elle diminue la robustesse de la composite et procure un gain de précision relativement faible. Vu la faiblesse de l'amélioration que procure la normalisation des indicateurs de la cote globale et les postulats que son utilisation nécessite, nous croyons qu'il est plus prudent d'adopter l'échelle standardisée comme échelle de mesure commune.

5.4 Effet d'un changement de pondération sur la composite.

Lorsqu'on mesure l'effet d'un changement de pondération sur la composite, on doit tenter de comprendre quels sont les effets d'un changement de pondération nominale sur la robustesse de la composite et ensuite déterminer l'effet d'un changement de pondération nominale sur la cote globale et les rangs attribués aux écoles.

La pondération utilisée par le *Bulletin* augmente la robustesse des résultats composites comme l'indique la somme des sr^2 du tableau 5.7 qui passe de 27 % à 10 %. En faisant passer le poids de la variable *zresexam* de 20 % à 40 %, celui des variables *zsurest* et *zecart* de 20 % à 10 % et en gardant la même pondération pour les variables *zechec* et *ztrans*, on donne plus d'importance aux variables fortement corrélées et moins à celles qui sont faiblement corrélées.

Le poids effectif (r^2) de *zresexam* augmente de ,721 à ,918, tout comme sa contribution unique (sr^2) qui elle, passe de ,009 à ,025. Quant aux variables *zsurest* et *zecart*, dont le poids nominal est diminué, leurs coefficients r^2 et sr^2 diminuent considérablement puisqu'ils sont faiblement corrélés avec les autres variables indépendantes. Les poids effectifs de *zsurest* et *zecart* diminuent respectivement de

,034 à ,000 et de ,181 à ,071. Leur contribution unique passe elle de ,094 à ,016 et de ,102 à ,017 respectivement.

Finalement, les deux variables dont le poids reste identique, *zechec* et *ztrans*, voient aussi leurs coefficients r^2 et sr^2 être altérés par le changement de pondération nominale effectué. Le poids effectif de *zechec* augmente de ,743 à ,891 et sa contribution unique diminue de ,010 à ,007. Le poids effectif de *ztrans* augmente de ,582 à ,645 et sa contribution unique diminue de ,055 à ,036. Ces variations sont attribuables uniquement au fait qu'elles présentent une forte corrélation avec la variable *zresexam*, qui est maintenant pondérée plus fortement. C'est pourquoi on note une augmentation du r^2 et une diminution du sr^2.

Tableau 5.7

Effet de la pondération nominale sur la pondération effective (r^2) et la contribution unique (sr^2) des indicateurs de la cote globale

	Pondération 1		Pondération 2	
	r^2	sr^2	r^2	sr^2
zresexam	,721	,009	,918	,025
zechec	,743	,010	,891	,007
ztrans	,582	,055	,645	,036
zsurest	,034	,094	,000	,016
zecart	,181	,102	,071	,017
somme		,270		,100

Les travaux de Terwilliger et Anderson (1969), Fralick et Raju (1982) et Aamodt et Kimbrough (1985) montrent qu'une composite ayant un petit nombre d'indicateurs (si K se situe entre 3 et 10 par exemple) et dont l'intercorrélation moyenne est faible (.5 ou moins) est généralement sensible à l'ensemble de pondérations choisies. La pondération de notre composite devrait avoir un effet

important sur les résultats obtenus puisque 5 indicateurs sont combinés et que la corrélation moyenne se situe aux alentours de .5.

Le changement moyen de la cote globale provoqué par l'adoption d'une pondération non équivalente est de 2,1 points (ET = 1,95). Le changement de pondération provoque un changement moyen de 23,4 rangs (ET = 25,87). Par contre, le coefficient de corrélation de Spearman de ,950 ($p < ,01$) indique que les classements faits à partir de chacune de ces pondérations sont fortement corrélés.

Alors qu'il était possible d'affirmer que la différence de rang provoquée par l'ajout d'indicateurs de performance et par la standardisation avait pour effet d'améliorer la justesse de l'évaluation de la performance faite à partir de la composite, on ne peut affirmer que la différence que provoque un changement de pondération améliore ou non cette mesure. On ne peut que noter une certaine fragilité de la composite à tout changement de pondération significatif. On sait également que la pondération adoptée par le *Bulletin* rend les résultats plus robustes face à des changements de scores des variables utilisées.

5.5 Conclusion.

Les résultats des diverses analyses effectuées montrent que la mesure de la performance des écoles québécoises devrait : 1) comprendre plus d'un indicateur, 2) utiliser une échelle de mesure standardisée et 3), pondérer les indicateurs utilisés par la composite de manière a donner plus d'importance aux indicateurs fortement corrélés.

Premièrement, en ce qui concerne le nombre d'indicateurs utilisés pour la composite, les analyses de la première section de ce chapitre ont démontré que le fait d'ajouter des indicateurs de performance pertinents améliore la justesse de l'évaluation. Il est important de noter que pour pouvoir conclure à l'amélioration de la mesure, les indicateurs ajoutés à la composite doivent être justifiés théoriquement.

Cette justification nous permet de croire que leur intégration à la composite contribue à l'amélioration de la mesure de la performance des écoles.

Deuxièmement, en ce qui concerne l'utilisation d'une échelle de mesure standardisée, la deuxième section du présent chapitre a montré que la standardisation est nécessaire à la neutralisation des instruments de mesure utilisés par la composite. Cependant, il ne faut pas croire qu'une échelle standardisée assure à l'évaluateur une pondération effective égale à la pondération nominale qu'il attribue à chacun des indicateurs. La corrélation des variables indépendantes a un effet important sur le calcul de la cote globale. La corrélation entre indicateurs est à l'origine des différences du poids effectif des variables prédictives une fois la standardisation des données complétée. La standardisation, en dépit du fait qu'elle diminue la robustesse de la composite, s'avère nécessaire puisque l'on se doit d'utiliser une échelle d'intervalle. De plus, l'amélioration de la précision qu'elle apporte est importante et justifie, à elle seule, son adoption.

Quant à l'utilisation de l'échelle normalisée, une première analyse a montré que seulement deux variables peuvent être normalisées. À la lumière de ces résultats, nous croyons que le nombre d'indicateurs pouvant être normalisés n'est pas suffisamment important pour nous permettre de poser l'hypothèse de la normalité de la distribution des performances dans la population à l'étude. Aussi, la contribution marginale apportée à l'évaluation de la performance des écoles par la normalisation de ces deux indicateurs est mince. Ainsi, nous croyons qu'il est préférable et plus prudent d'utiliser l'échelle standardisée plutôt que l'échelle normalisée lors de la mesure de la performance des écoles.

Finalement, en ce qui concerne l'effet d'un changement de pondération nominale, la composite utilisée par le *Bulletin* est sensible aux changements de pondérations. Par contre, la pondération adoptée lui assure une plus grande robustesse en donnant aux indicateurs qui sont fortement corrélés des pondérations plus élevées

que les indicateurs faiblement corrélés. Ce choix réduit la somme des contributions uniques des variables indépendantes et donc, augmente la robustesse de la composite.

BIBLIOGRAPHIE

AAMODT, Michael G. et Wilson W. KIMBROUGH (1985). « Comparaison of four methods for weighting multiple predictors », *Educational and psychological measurement*, vol. 45, p. 477-482.

ANGOFF, William H. (1971). « Scales, norms, and equivalent scores », dans Robert L. Thorndike (dir.), *Educational Measurement*, 2e éd., Washington, D.C., American Council on Education, p.508-600.

BRADLEY, J. V. (1984). « The complexity of nonrobustness effects » *Bulletin of the psychonomic society*, vol. 22, n° 3, p. 250-253.

BRADLEY, J. V. (1982). « The insidious L-shaped distribution », *Bulletin of the psychonomic society,* vol. 20, n° 2, p. 85-88.

BROGDEN, H. E. et E. K. TAYLOR (1950). « The dollar criterion : applying the cost accounting concept to criterion construction », *Personnel psychology*, vol. 3, p. 133-167.

CATTELL, R. B. (1957). *Personality and motivation : structure and measurement.* New York, Harcourt, Brace and World.

COUNCIL OF CHIEF STATE SCHOOL OFFICERS (1999). *Annual survey: state student assessment programs: a summary report fall 1999.* Washington, DC, Council Of Chief State School Officers.

COWLEY, Peter et Richard MARCEAU (2000). *Bulletin de écoles secondaires du Québec*, éd. 2000, Vancouvert, Institut Fraser, « Études sur les politiques éducationnelles ».

CREAGER, J. A. et L. D. VALENTINE (1962). « Regression analysis of linear composite variance » *Psychometrika*, vol. 27, p. 31-38.

CRONBACH, Lee J. (1990). *Essentials of psychological testing*, 5e éd., New York, Harper and Row.

CURETON, E. E. (1951). « Approximate linear restraints and best predictor weights » *Psychometrika*, vol. 11, p. 12-15.

DARLINGTON, Richard B. (1968) « Multiple regression in psychological research and practice », *Psychological Bulletin*, vol. 69, p. 161-182.

DAWES, R. M. et B. CORRIGAN (1974). « Linear models in decision making », *Psychological Bulletin*, vol. 81, p. 95-106.

DUNNETTE, M. D. (1963). « A modified model for test validation and selection research », *Journal of applied psychology*, vol. 47, p. 317-323.

DUNNETTE, M. D. et A. C. HOGGATT (1957). « Deriving a composite score from several measures of the same attribute » *Educational and psychological measurement*, vol. 17, p. 423-434.

Fitz-GIBBON et KOCHAN (2000). Dans C. TEDDLIE et D. REYNOLDS (Eds.). *The international handbook of school effectiveness research.* New York, Falmer Press.

FRALICK, Rodney D. et Nambury S. RAJU (1982). « A comparaison of five methods for combining multiple criteria into a single composite », *Educational and psychological measurement*, vol. 42, p. 823-827.

GARDNER, R. C. et S. ERDLE (1984). « Aggregating scores : to standardize or not to standardize? », *Educational and psychological measurement*, vol. 44, p. 813-821.

GHISELLI, E. E. (1956). « Dimensional problems of criteria. », *Journal of applied psychology*, vol. 40, p. 1-4.

GOLDSTEIN, H. (1995). *Multilevel models in educational and social research: a revised edition.* London, Edward Arnold.

GRAY, J., D. JESSON, H. GOLDSTEIN, K. HECKER, J. RASBASH (1995). « A multi-level analysis of school improvement : changes in school's performance over time ». *School effectiveness and school improvement*, vol. 6, n° 2, p. 97-114.

GUBA, Egon G. et Yvonna S. LINCOLN (1989). *Fourth generation evaluation*, Sage publications.

GUILFORD, Joy Paul et Benjamin FRUCHTER (1978). *Fundamental statistics in psychology and education*, 6e éd., McGraw-Hill Inc., « McGraw-Hill series in psychology ».

GULLIKSEN, Harold (1962). *Theory of mental test*, 4e éd., New York, John Willey & Sons.

HORST, Paul (1936). « Obtaining a composite measure from a number of different measures of the same attribute » *Psychometrika*, vol. 1, p. 53-60.

HOUSE, Ernest R. et Kenneth R. HOWE (1999). *Values in evaluation and social research*, Sage Publications.

LINCOLN, Yvonna. S. et Egon G. GUBA (2000). « Paradigmatic controversies, contradictions, and emerging confluences » dans N. K. Denzin et Y. S. Lincoln (dirs.), *Handbook of Qualitative Research*, 2e éd., Thousand Oaks, Sage Publications, p. 163-188.

MANDEVILLE, G. K. et L. W. ANDERSON (1987). « The stability of school effectiveness indices across grade levels and subject areas » *Journal of educational measurement*, vol. 24, n° 3, p.203-216.

MARCEAU, Richard (2000). « Le palmarès maudit? », *La Presse*, 11 novembre.

MEEHL, Paul E. (1966). *Clinical versus statistical prediction : a theoretical analysis and a review of the evidence*, 6e éd., Minneapolis, University of Minneapolis Press.

MEHRENS, William A. (1990). « Combining evaluation data from multiple sources », dans Jason Millman et Linda Darling-Hammond (éd.), *The new handbook of teacher evaluation : assessing elementary and secondary school teachers*, Sage Publications, p.322-334.

MESSICK, S. (1994). « The interplay of evidence and consequences in the validation of performance assessments ». *Educational Researcher*, vol. 23, n° 2, p. 13-23.

MICHELL, Joel (1986). « Measurement sclaes and statistics : a clash of paradigms » *Psychological bulletin*, vo. 100, n° 3, p. 398-407.

MONROE, W. S. et D. B. STUIT (1935). « Correlation analysis as a means of studying contributions of causes » *The journal of experimental education*, vol. 3, p. 155-165.

MORTIMORE, P., P. SAMMONS, L. STOLL, D. LEWIS et R. ECOB (1988). *School matters : the junior years*, Wells, Open books.

NISBETT, R. E. et L. ROSS (1980). *Human inference : strategies and short comings of human judgement.* Prentice-Hall.

OCDE. CENTRE POUR LA RECHERCHE ET L'INNOVATION DANS L'ENSEIGNEMENT (1995). *Gros plan sur les écoles*, Paris, OCDE.

PETERSEN, Nancy S., Michal J. KOLEN et H. D. HOOVER (1989). « Scaling, Norming and Equating » dans Robert L. Linn (éd.), *Educational measurement*, National Council on Education and Macmillan Publishing Company, « Series on higher education », p. 221-262.

QUÉBEC. CONSEIL SUPÉRIEUR DE L'ÉDUCATION (1999). *L'évaluation institutionnelle en éducation : une dynamique propice au développement*, Québec, le Conseil.

QUÉBEC. GOUVERNEMENT DU QUÉBEC (1999). *Pour de meilleurs services aux citoyens –Un nouveau cadre de gestion pour la fonction publique*. Énoncé de politique sur la gestion gouvernementale, Québec.

QUÉBEC. MINISTÈRE DE L'ÉDUCATION (2001). *Résultats aux épreuves uniques de juin 2000 par commission scolaire et par école pour les secteurs public et privé et diplomation par commission scolaire*, Québec, Direction de la sanction des études.

ROSSI, Peter H. (1982). « Standards for evaluation practice » *New directions for program evaluation*, Jossey-Bass Publishers, n° 15.

ROZEBOOM, W. W. (1965). « Linear correlations between sets of variables » *Psychometrika*, vol. 30, p. 57-71.

SCHMIDT, Frank L. et Leon B. KAPLAN (1971). « Composite vs. multiple criteria : a review of the controversy », *Personnel psychology*, vol. 24, p. 419-434.

SCRIVEN, Michael (1967). « The methodology of evaluation » dans *AERA Monograph Series in Curriculum Evaluation*, n° 1, p. 39-83.

SCRIVEN, Michael (1983). « Evaluation ideologies » dans G. F. Madaus, M. S. Scriven et D. L. Stufflebeam (éds.), *Evaluation models*, Boston, Kluwer-Nijhoff, p. 229-260.

SCRIVEN, Michael (1991). *The thesaurus*, 4e éd., Sage Publications.

SCRIVEN, Michael (1993). « Hard-won lessons in program evaluation », *New directions for program evaluation*, Jossey-Bass Publishers, n° 58, p. 1-37.

STEVENS, Joseph J., Susan ESTRADA et Jay PARKES (2000). *Measurement issues in the design of state accountability systems*, Rapport présenté à la conférence annuelle de l'American Educational Research Association.

STEVENS, Joseph J. et Lawrence M. ALEAMONI (1986). « The role of weighting in the use of aggregate scores », *Educational and psychological measurement*, vol. 46, p. 523-531.

STEVENS, S. S. (1946). « On the theory of scales of measurement », *Science*, vol. 103, p. 667-680.

STEVENS, S. S. (1951). « Mathematics, measurements, and psychophysics », dans S. S. Stevens (éd.), *Handbook of experimental psychology*, New York, Wiley, p. 1-49.

STUIT, D. B. (1947). *Personnel research and test development in the bureau of naval personnel*. Princeton, N. J., Princeton University Press.

SUPPES, P. et J. L. ZINNES (1963). « Basic measurement theory » dans R. D. Luce, R. R. Bush et E. Galanter (éds.), *Handbook of mathematical psychologie*. New York, Wiley, p. 3-76.

TABACHNICK, Barbara G. et Linda S. FIDELL (2001). *Using multivariate statistics*, 4e éd., Needham Heights, MA, Allyn & Bacon.

TERWILLIGER, James S. et Douglas H. ANDERSON (1969). « An empirical study of the effects of standardizing scores in the formation of linear composites », *Journal of educational measurement*, vol. 6, n° 3, p. 145-154.

WILKS, S. S. (1938). « Weighting systems for linear functions of correlated variables when there is no dependent variables », *Psychometrika*, vol. 3, n° 1, p. 23-40.

YEN, Wendy M. (1986). « The choice of scale for educational measurement : an IRT perspective », *Journal of educational measurement*, vol. 23, n° 4, p. 299-325.