

Université de Montréal

Application de la théorie des ensembles flous à  
l'élaboration d'un modèle pour prédire la réussite dans  
une école de génie à partir du score à une épreuve  
diagnostique

par

Paul Martin

Département de l'administration et des fondements de  
l'éducation

Faculté des sciences de l'éducation

Thèse présentée  
en Sciences de l'éducation  
option Mesure et évaluation (Ph. D.)

Avril, 2007

© Paul Martin, 2007





## AVIS

L'auteur a autorisé l'Université de Montréal à reproduire et diffuser, en totalité ou en partie, par quelque moyen que ce soit et sur quelque support que ce soit, et exclusivement à des fins non lucratives d'enseignement et de recherche, des copies de ce mémoire ou de cette thèse.

L'auteur et les coauteurs le cas échéant conservent la propriété du droit d'auteur et des droits moraux qui protègent ce document. Ni la thèse ou le mémoire, ni des extraits substantiels de ce document, ne doivent être imprimés ou autrement reproduits sans l'autorisation de l'auteur.

Afin de se conformer à la Loi canadienne sur la protection des renseignements personnels, quelques formulaires secondaires, coordonnées ou signatures intégrées au texte ont pu être enlevés de ce document. Bien que cela ait pu affecter la pagination, il n'y a aucun contenu manquant.

## NOTICE

The author of this thesis or dissertation has granted a nonexclusive license allowing Université de Montréal to reproduce and publish the document, in part or in whole, and in any format, solely for noncommercial educational and research purposes.

The author and co-authors if applicable retain copyright ownership and moral rights in this document. Neither the whole thesis or dissertation, nor substantial extracts from it, may be printed or otherwise reproduced without the author's permission.

In compliance with the Canadian Privacy Act some supporting forms, contact information or signatures may have been removed from the document. While this may affect the document page count, it does not represent any loss of content from the document.

Université de Montréal  
Faculté des études supérieures

Cette thèse intitulée :

*Application de la théorie des ensembles flous à l'élaboration d'un modèle pour  
prédire la réussite dans une école de génie à partir du score à une épreuve  
diagnostique*

présentée par :

*Paul Martin*

a été évaluée par un jury composé des personnes suivantes :

*Michel Laurier*

président-rapporteur

*Jean-Guy Blais*

directeur de recherche

*Michel Desmarais*

membre du jury

*Pierre Valois*

examineur externe

.....

représentant du doyen de la FES

## Résumé

À l'aube de ce XXI<sup>e</sup> siècle, la réussite scolaire constitue un enjeu socio-économique important et le diagnostic des élèves en difficulté d'apprentissage est devenu une tâche prioritaire pour l'éducateur. Un diagnostic repose sur des informations qui prennent parfois la forme de scores à une épreuve. Alors, il est préférable d'interpréter ces scores en référence au critère de la réussite future. Par conséquent, la tâche de l'éducateur serait grandement facilitée s'il pouvait compter sur un modèle qui illustre la relation entre le diagnostic posé et le pronostic de la réussite.

Les théories des tests existantes restent limitées quant à l'interprétation des scores en référence à un critère. En effet, le diagnostic traditionnel de réussite/échec obtenu en comparant le score brut à un score de césure représente un modèle incomplet. La théorie des ensembles flous vient nuancer ce modèle en y ajoutant un second paramètre afin d'illustrer le lien entre le score brut et le degré de certitude pour un diagnostic de réussite. Ainsi, la donnée obtenue est classée dans un ensemble flou sans qu'on ait besoin d'assumer que le trait soit mesuré sur une échelle d'intervalle. De plus, le traitement des données s'effectue par des opérations sur les ensembles flous et la valeur vraie du degré d'inclusion d'un sous-ensemble dans un ensemble vient inférer une règle d'implication logique du type : «Si l'élève a été diagnostiqué positivement alors il réussira.»

L'expérimentation du modèle pour cette recherche a été faite à partir de données qui nous proviennent de l'École Polytechnique de Montréal. En effet, depuis plusieurs années les nouveaux étudiants passent une épreuve de mathématique lors de l'admission et nous présumons qu'il existe un lien entre la réussite des études de génie et la capacité en mathématique. En ce sens, le modèle a été élaboré à partir des scores pour la cohorte de 1997 à cette épreuve de mathématique et de leurs cotes au cours «Calcul I». Afin de valider ce modèle, trois hypothèses ont été formulées et chacune prend la forme de deux règles

d'implication logique: l'une pour prédire la réussite et l'autre pour prédire l'échec. Enfin, d'autres données pour les cohortes de 1997 et de 1999 ont permis de vérifier ces hypothèses.

Les résultats nous montrent que le modèle est valide pour prédire la réussite future d'un étudiant mais non l'échec. Conséquemment, il ne peut être utilisé directement pour diagnostiquer les étudiants en difficulté. Même si les résultats obtenus semblent mitigés au premier abord, il reste que le modèle élaboré à partir de la théorie des ensembles flous constitue un outil important pour favoriser la réussite scolaire.

### **Mots clés**

Théorie des ensembles flous – modèle – interprétation des scores en référence à un critère – réussite scolaire – diagnostic – pronostic – implication logique

### **Abstract**

*At the dawn of the twenty-first century, academic achievement has become a growing socio-economic concern. Being able to properly identify and assess students' learning problems now constitutes a task of prime importance for the instructor. Since such diagnosis often stems from information pertaining to a given score on a test or other evaluation, it becomes essential for those scores to be analysed and interpreted based on criteria which underlie future achievement. Consequently, the educator's task would be greatly facilitated if he could benefit from a model that can highlight the cause-and-effect relationship linking both initial diagnosis and achievement forecast.*

*Current theories on measure and evaluation have remained hitherto relatively limited in their ability to deal with the interpretation of scores in comparison with a criterion-based reference. Indeed, the traditional success/failure diagnosis outcome, which consists of comparing a gross score with a threshold*

score, is an incomplete model. However, the fuzzy sets theory adds further nuance to this model by adding a second parameter which shows the link between the gross score and the degree of certainty behind a success diagnosis. Thus, the obtained data is classified in a fuzzy set —no need to resort to an interval scale for that matter. Furthermore, the data processing is performed by operations onto fuzzy sets and the true value of inclusion of a subset within a set so as to yield inferences of logical implication rules the like of, « If the student was diagnosed positively, then he will succeed. »

The model experimentation for this research was conducted with data from Ecole Polytechnique de Montréal because for the past many years, new students to that institution have had to write a mathematics test for it is assumed that there is a positive correlation between maths proficiency and achievement in engineering studies. This model experimentation was elaborated from the scores obtained on the aforementioned maths test by the 1997 group and also from their grade in the « Calculus I » course. As a means to validate this model, three hypotheses were formulated, each of which taking the form of two logical implication rules: one to forecast success and the other to forecast failure. At last, additional data from the 1997 and 1999 groups allowed to put these hypotheses to the test.

Results have revealed that the model is valid to forecast future students' success but not their failure. Consequently, we cannot use it directly to diagnose students in difficulty. Nonetheless, even if outcomes seem mitigated at first glance, the model elaborated with fuzzy sets theory constitutes an important step or tool to foster academic achievement.

### Keys words

*Fuzzy sets theory – model – criterion-based referenced scores – academic achievement – diagnosis – forecast – logical implication*

## Table des matières

	Page
Liste des tableaux .....	vii
Liste des figures.....	viii
Liste des sigles.....	ix
Liste des symboles mathématiques.....	ix
Remerciements .....	x
Introduction .....	1
Chapitre premier: La problématique.....	5
1. La réussite.....	5
2. Le plan de réussite.....	8
3. Le problème de validité de la mesure critériée en éducation.....	15
4. Les visées de la recherche.....	22
Chapitre second: Le cadre conceptuel .....	27
1. La mesure en éducation .....	27
2. La théorie classique des tests (TCT) .....	36
3. La théorie de réponse à l'item (TRI).....	41
4. La théorie de la généralisabilité .....	50
5. La mesure critériée et la théorie des ensembles.....	53
6. La théorie des ensembles flous, un changement de paradigme .....	69
7. La théorie des ensembles flous : une théorie de la mesure critériée .....	80
8. Les concepts de la TEF.....	92
9. Application de la théorie des ensembles flous à l'élaboration d'un modèle prédictif de la réussite .....	104
Chapitre troisième: La méthodologie .....	112
1. Le choix méthodologique .....	112
2. La provenance des données .....	115
3. Les hypothèses de recherche.....	120
4. La méthode d'élaboration du modèle pronostique.....	122
5. La méthode de validation du modèle pronostique .....	126
6. Les limites à la généralisation des résultats.....	132

Chapitre quatrième: Analyse des données .....	137
1. La description des données .....	137
2. Le classement des données .....	143
3. L'inférence des relations de causalité .....	147
4. La vérification des limites pour la généralisation .....	156
Chapitre cinquième: Discussion des résultats .....	159
1. La fidélité dans l'interprétation des données.....	159
2. La validité du modèle pronostique .....	165
Conclusion.....	170
Bibliographie .....	173
Annexe I.....	182
L'entropie d'une distribution uniforme.....	182
1. La distribution uniforme de l'appartenance à un ensemble flou .....	182
2. L'intégrale du numérateur .....	182
3. L'intégrale du dénominateur .....	183
4. L'entropie de la distribution uniforme.....	183
Annexe II.....	184
Devis pour les données de la recherche à l'École Polytechnique .....	184
1. Le but.....	184
2. Méthode .....	184
3. Requête de données.....	184
Annexe III .....	186
Apparition des nouveaux sigles (cours) et correspondance avec les anciens sigles .....	186

## Liste des tableaux

	Page
Tableau I: Propriétés des nombres.....	29
Tableau II: Scalogramme d'un test à quatre items .....	42
Tableau III: Lois de l'algèbre de Boole.....	60
Tableau IV: L'appartenance à un ensemble flou .....	84
Tableau V : Les échelles pour une situation de classement (Ragin, 2000).....	85
Tableau VI: Un exemple de calcul pour la cardinalité scalaire.....	97
Tableau VII: Un exemple de calcul de l'entropie .....	99
Tableau VIII: Exemple de la valeur vraie d'une implication logique.....	102
Tableau IX: Le bilan des appartenances .....	105
Tableau X: Les degrés de réussite au cours «Calcul I» .....	127
Tableau XI : Le degré de réussite dans un programme pour la cohorte 1997..	129
Tableau XII: Comparaison des distributions précédentes.....	139
Tableau XIII : La diplomation dans les programmes de génie .....	142
Tableau XIV: Groupes contraires pour déterminer la césure.....	144
Tableau XV: Le calcul de l'indice de Livingston (N=426) .....	145
Tableau XVI: La description des ensembles flous .....	146
Tableau XVII : Bilan pour la réussite au cours 105 (N=218) .....	148
Tableau XVIII : Bilan pour l'échec au cours 105 (N=218) .....	149
Tableau XIX : Bilan pour la réussite dans un programme (N=340) .....	151
Tableau XX : Bilan pour l'échec dans un programme (N=340) .....	152
Tableau XXI : Bilan pour la réussite dans le cours ING1001 (N=389) .....	154
Tableau XXII : Bilan pour l'échec dans le cours ING1001 (N=389).....	155
Tableau XXIII : L'exactitude du pronostic de réussite .....	157
Tableau XXIV : La généralisation transversale .....	158

## Liste des figures

	Page
Figure 1: Le modèle d'un processus d'enseignement différencié .....	14
Figure 2: Le schéma de la communication .....	32
Figure 3: Ordonnancement des items .....	43
Figure 4: Le modèle de Rasch .....	45
Figure 5: Mise en ordre des items selon le modèle de Rasch .....	46
Figure 6 : Le plan de mesure $a \times i$ .....	51
Figure 7: Le diagramme de Venn pour une mesure critériée .....	56
Figure 8: Les diagnostics à des tests parallèles $X_1$ et $X_2$ .....	62
Figure 9: Les erreurs avec une mesure critériée .....	63
Figure 10 : Le modèle d'une mesure critériée floue .....	89
Figure 11: La fonction caractéristique d'une mesure critériée .....	90
Figure 12: Le diagramme de Venn pour la validité pronostique .....	93
Figure 13 : Les zones d'incertitude pour la cause X et l'effet Y .....	107
Figure 14: Le plan d'analyse .....	119
Figure 15: Distributions des scores pour des groupes contraires .....	123
Figure 16 : Plan pour la généralisation du modèle pronostique .....	134
Figure 17: Distribution des scores à l'épreuve de juin 1997 (N=426) .....	138
Figure 18 : Distribution de juin 1997 pour les sujets du cours 105 (N=344) .....	139
Figure 19 : Distribution des cotes au cours 105 (N=344) .....	140
Figure 20 : Résultats obtenus à l'épreuve de juin 1999 (N=478) .....	143
Figure 21 : Modèle pronostique .....	161

### Liste des sigles

- ANOVA : Analysis of variance
- B. ing. : Baccalauréat en ingénierie
- IEEE : Institute of Electrical and Electronical Engineers
- L.R.Q. : Lois et règlements du Québec
- MEQ : Ministère de l'éducation, des loisirs et des sports du Québec (le nouveau sigle officiel est MELS mais nous avons préféré conserver l'ancien sigle MEQ)
- RHDC : Ressources humaines et développement des compétences du  
Gouvernement du Canada
- SAT : Scholastic Assessment Test
- TCT : Théorie classique des tests
- TEF : Théorie des ensembles flous
- TG : Théorie de la généralisabilité
- TRI : Théorie de réponse à l'item

### Liste des symboles mathématiques

- U : univers des sujets pour une cohorte donnée
- X : ensemble flou des sujets ayant été diagnostiqués positifs
- X' : ensemble flou des sujets ayant été diagnostiqués négatifs
- Y : ensemble flou des sujets ayant réussi
- Y' : ensemble flou des sujets ayant échoué
- $\mu_X(a)$  :  $U \rightarrow [0, 1]$  degré d'appartenance du sujet «a» à l'ensemble flou X
- $|X|$  : cardinal scalaire de l'ensemble flou X qui est égal à la somme des degrés d'appartenance,  $\mu_X(a)$ , pour tous les sujets de l'univers U
- $\|X \subseteq Y\|$  : valeur vraie, située entre 0 et 1, que l'ensemble X soit inclus dans l'ensemble Y

## Remerciements

Même si cette thèse semble à première vue le fruit d'une seule personne, il reste que je n'aurais pu arriver à bon port sans le soutien de plusieurs personnes. Je remercie Carole Burny-Vincent, Yvon Bouchard et Jean-Guy Deschesnes de l'École Polytechnique de Montréal pour m'avoir procuré les données utilisées dans cette recherche. Un merci à Nathalie Loye pour son optimisme lorsqu'on travaillait sur les données. Il n'est pas évident d'entreprendre une recherche sur une théorie aussi exotique que la théorie des ensembles flous ; un grand merci à Jean-Guy Blais qui a su me diriger avec une très grande rigueur et qui a toujours été très disponible.

Enfin, un grand merci à ma famille: à Julie pour m'avoir donné le goût d'entreprendre des études de doctorat ; à Annie pour m'avoir dit les bons mots à un moment de découragement ; et particulièrement un très grand merci à Gisèle, mon épouse, qui a révisé le français de la thèse et qui m'a toujours supporté.

## Introduction

Avant le XIX<sup>e</sup> siècle, le système d'éducation de la plupart des pays visait essentiellement à instruire une élite. Les individus privilégiés accédant à une éducation supérieure s'instruisaient pour exercer une profession libérale telle : médecin, avocat, notaire, prêtre, etc. Il existait en parallèle des artisans qui formaient eux-mêmes des apprentis pour assurer la pérennité de leur métier à travers les générations.

Avec l'avènement de l'ère industrielle, les artisans à eux seuls ne pouvaient pas combler tous les besoins en main d'œuvre spécialisée ; désormais, il faudrait aussi des ingénieurs, des techniciens, des enseignants et des scientifiques de tout acabit. Afin de combler cette lacune, quelques pays, dont la France, rendirent obligatoire l'instruction dans la seconde moitié du XIX<sup>e</sup> siècle. En d'autres mots, on a implanté un système d'éducation de masse par opposition au système élitiste qui existait auparavant. Ainsi, l'éducation devenait accessible à tous les individus et les meilleurs d'entre eux pouvaient accéder à une éducation supérieure. Des écoles techniques prirent donc le relais des artisans pour l'instruction des métiers. Au Québec, on ne mit en place un système d'éducation de masse qu'avec la révolution tranquille des années 60.

Maintenant, au début du XXI<sup>e</sup> siècle, le système d'éducation de masse ne suffit plus. En effet, les besoins pour une main d'œuvre de plus en plus spécialisée agissent comme une contrainte sur le système d'éducation pour faire réussir le plus grand nombre possible d'élèves et d'étudiants. Il ne faut plus seulement l'accessibilité pour tous, il faut aussi la réussite pour tous.

Or, la responsabilité de faire réussir le plus grand nombre d'individus possible n'incombe plus seulement à l'enseignant. En effet, il devient également nécessaire de dispenser un enseignement différencié, c'est-à-dire offrir un enseignement qui prend en considération les rythmes d'apprentissage des

individus. Pour ce faire, l'éducateur responsable du curriculum se doit de recueillir des informations valides afin de poser un diagnostic envers chaque élève ou étudiant et l'orienter dans un cheminement scolaire adéquat. Toutefois, ce diagnostic doit s'appuyer sur une prédiction de la réussite ; or, la mesure de la capacité nécessaire à un individu pour assurer sa réussite constitue un problème en éducation.

Il faut ajouter que la plupart des recherches pour solutionner ce problème ont toujours été initiées à partir de fondements diamétralement opposés, soient les deux courants épistémologiques contraires que sont : le positivisme proposé par Auguste Comte et le dualisme épistémologique amené par Kant. D'une part, avec le positivisme, on cherche à unifier la connaissance et à reproduire tous les phénomènes à l'image des phénomènes physiques ; cela conduit indéniablement à adopter une méthodologie quantitative qui ne peut prendre en considération toute la complexité de l'être humain. D'autre part, avec le dualisme épistémologique, on considère différemment les sciences de la nature et les sciences humaines ; dans le cas des sciences humaines, on privilégiera une méthodologie qualitative mais encore là, on se retrouve avec le problème de ne pas pouvoir généraliser les résultats.

Au XXe siècle, le philosophe Karl Popper se distingua des deux courants précédents par un discours sur la relativité épistémologique dont Debays (1998, p.8) résume la pensée :

«Les théories scientifiques ne doivent pas être considérées comme des énoncés de vérité immuables, mais bien comme des conjectures provisoires qui peuvent à tout moment être falsifiées et remplacées par des théories plus performantes.»

Il est possiblement vrai que le monde réel soit une gigantesque mécanique comme le prétend le courant du positivisme mais son comportement ne peut être prévu avec précision car l'image que nous en percevons est une image floue; nous pouvons distinguer les formes sans pouvoir nécessairement identifier tous les détails. Pour le courant de la relativité épistémologique, les théories agissent comme un télescope pour venir affiner l'image de l'univers. Cependant, le

discours scientifique ne peut jamais être pure vérité, il doit plutôt montrer la conscience de l'erreur en intégrant l'incertitude comme c'est le cas pour la mécanique quantique ; voilà la position épistémologique de Popper qui sera nôtre.

Une autre théorie issue de ce courant du relativisme épistémologique a émergé dans les années 1960, il s'agit de la théorie des ensembles flous. Cette théorie a depuis trouvé de nombreuses applications dans d'autres domaines y compris celui de l'éducation et elle a l'avantage de tenir compte de l'incertitude dans les données. Aussi, notre recherche va recourir à cette théorie des ensembles flous pour élaborer un modèle qui vient appuyer l'éducateur responsable d'un curriculum lorsque celui-ci diagnostiquera un étudiant ou un élève.

Le premier chapitre portera sur la problématique qui nous conduira à définir l'objectif et établir la question à laquelle la recherche vise à répondre.

Dans le second chapitre portant sur le cadre conceptuel, une recension des écrits viendra appuyer la remise en question des approches standard de la modélisation des scores aux tests et la pertinence de recourir à cette théorie des ensembles flous.

Le troisième chapitre abordera la méthodologie qui explicitera la démarche de la recherche et des données, acquises depuis plusieurs années à l'École Polytechnique de Montréal avec une épreuve de mathématique administrée lors de l'admission, serviront à élaborer le modèle. Plusieurs hypothèses seront énoncées afin de vérifier l'efficacité de la prédiction de la réussite. De plus, la confidentialité des sujets sera assurée par le fait qu'ils ne seront en aucun temps identifiés dans cette recherche.

Le quatrième chapitre touchera l'analyse des données et la discussion des résultats pour deux cohortes d'étudiants se fera au cinquième chapitre, ceci afin d'élaborer et de valider le modèle. Pour cela, chacune des variables sera

associée à un ensemble et, une donnée relative à un sujet sera traduite en un degré d'appartenance à un ensemble flou. Bref, c'est à partir de ce modèle qu'un diagnostic pourra être posé sur chaque sujet.

Finalement, les conclusions de cette recherche pourraient être généralisées à d'autres situations éducatives que l'on retrouve autant au secondaire qu'au post-secondaire.

## Chapitre premier

### La problématique

Dans ce chapitre, nous illustrerons comment la problématique de la réussite pour tous est liée au problème de la prédiction de la réussite pour chacun.

#### *1. La réussite*

Depuis quelques années au Québec, le discours sur l'éducation laisse beaucoup de place au concept de réussite. En effet, le discours sur l'accessibilité pour tous qui avait émergé de la révolution tranquille des années 60 s'est transformé au fil des ans en un discours sur la réussite. Alors que le premier faisait référence à la mise en place de structures, le second nous renvoie plutôt à la performance du système d'éducation quant à la réussite du plus grand nombre d'élèves possible.

- *Le concept de réussite*

Qu'entendons-nous par le concept de réussite ? Le dictionnaire Le Petit Larousse (1991) définit le mot réussite comme étant : «un succès, un résultat favorable». Le dictionnaire de l'évaluation et de la recherche en éducation (De Landsheere, 1979; p.2) associe plutôt la réussite à l'accomplissement et plus spécifiquement à : «la poursuite d'un but couronnée de succès, à une chose accomplie, au résultat d'une performance». Legendre (1988, p.492) parle du rendement scolaire qu'il associe au : «degré de réussite d'un sujet ou d'un ensemble de sujets en regard des objectifs spécifiques des divers programmes d'études».

On peut aussi qualifier la réussite en ajoutant l'adjectif scolaire ou professionnel. Dans le cas de la réussite scolaire, il y a au moins deux acceptions possibles : réussir en obtenant un diplôme ou réussir des exercices, des épreuves. Nous parlons aussi de la réussite professionnelle d'une personne lorsqu'elle gradue les échelons d'une organisation ou encore lorsqu'elle se montre compétente dans son travail. Dans la présente recherche, le concept de

réussite sera plutôt attaché à la réussite scolaire observée à la fin d'un processus d'enseignement, c'est-à-dire réussir un cours ou obtenir un diplôme à la fin d'un programme d'études.

- Quelques données sur la réussite

Le discours politique et médiatique pourrait nous laisser croire que la situation en ce qui concerne la réussite scolaire au Québec va de mal en pis. Pourtant, certains indicateurs de l'éducation (MÉQ, 2003) laissent entendre le contraire.

Depuis quelques années, les médias nous font voir le décrochage scolaire comme étant un problème qui est nouveau; est-ce la réalité ou s'agit-il d'un mythe ? La proportion des élèves qui redoublent une classe au primaire est passée de 4,7% en 1983-1984 à 3,1% en 2001-2002 (MÉQ, 2003 ; p.109). En 1982-1983, 64,1% des jeunes étaient admis en 4<sup>e</sup> secondaire ; en 2001-2002, cette proportion est passée à 84,4% (MÉQ, 2003 ; p.95). À l'université, la proportion des sortants avec diplôme est passée de 55,9% en 1987-1988 à 67,0% en 2000-2001 (MÉQ, 2003 ; p.148). Or, seul l'ordre d'enseignement collégial stagne dans sa performance depuis 1980-1981 avec des taux de diplomation qui tournent autour de 67% pour le secteur pré-universitaire et de 58% pour le secteur technique alors qu'il y a eu amélioration au primaire, secondaire et universitaire.

Les filles réussissent mieux que les garçons, cela est vrai mais est-ce un phénomène récent ou cela a-t-il toujours été ? L'écart entre les sexes pour le taux de redoublement au primaire est passé de 2,4 points en 1983-1984 à 1,5 points en 2001-2002 (MÉQ, 2003 ; p.109). Pour les taux d'accessibilité en 4<sup>e</sup> secondaire, l'écart était de 8,7 points en 1982-1983 et de 8,1 points en 2001-2002 (MÉQ, 2003 ; p.95). À l'université, les écarts se sont accentués un peu mais cela reste négligeable puisqu'en 2000-2001, il y avait 64,4% des sortants et 68,9% des sortantes qui décrochaient un diplôme (MÉQ, 2003 ; p.148).

Est-ce que le fait de rendre la scolarité obligatoire et d'admettre un plus grand nombre d'individus aux études post-secondaires a eu un effet sur la qualité de

l'éducation? Les élèves québécois de 13 ans ont obtenu des résultats légèrement supérieurs aux élèves canadiens du même âge pour deux épreuves de mathématique PIRS tenues au printemps 2001 (MÉQ, 2003 ; p.168). Et le taux de chômage pour les finissants en formation technique au collégial n'était que de 5,6% en 2003 (MEQ, 2004-1 ; p.9) comparativement à 10% pour la population active en janvier 2004 (Statistique Canada, 2005) ; de plus, 83,5% des emplois obtenus par les diplômés étaient en rapport avec leur formation.

Voilà des données qui viennent baliser la réalité sur la réussite. Mais alors, la réussite scolaire au Québec est-elle véritablement un problème?

- *La réussite scolaire : un enjeu économique*

La réussite scolaire devient un problème sérieux si on l'aborde sous l'angle de l'enjeu économique. Le développement du secteur des services au détriment du secteur de la transformation a fait en sorte que les emplois vont requérir de plus en plus une formation de niveau post-secondaire.

En effet, selon le ministère Ressources humaines et Développement des compétences du gouvernement du Canada (RHDC, 2004), les perspectives d'emploi bonnes et satisfaisantes pour l'année 2007 vont demander pour 57% des catégories une formation de niveau universitaire. Par contre, les perspectives limitées vont se retrouver pour 75% des catégories où il existe des emplois ne demandant qu'une formation professionnelle de niveau secondaire.

Le recensement de Statistique Canada pour l'année 1996 (Statistique Canada, 2001) nous informe que 12,2% de la population active du Québec détenait un grade universitaire comparativement à 13,3% pour le pays. Avec le recensement de 2001, ces proportions ont passé respectivement à 14,0% et 15,4%. Malgré le fait qu'il y ait eu amélioration, le Québec se situe toujours à 1,4 point sous la moyenne nationale; cela peut s'expliquer par la particularité du système collégial qui fait en sorte que les québécois prennent une année de plus avant d'accéder à des études universitaires. De plus, la proportion de la population ayant fait des

études universitaires en 2001 est de 22,6% pour le Québec et de 25,8% pour le Canada.

Étant donné ce qui vient d'être dit ci-dessus, la réussite scolaire au Québec peut être vue comme un problème. Il reste cependant la question du: «Comment s'y prendre pour améliorer la performance du système d'éducation?»

## *2. Le plan de réussite*

Afin de mettre en pratique son discours, le gouvernement du Québec a introduit en 2002 des dispositions légales pour établir des plans de réussite ou des contrats de performance dans tous les établissements scolaires et ce, pour chacun des ordres d'enseignement. Cette façon de faire postule qu'une décentralisation de la responsabilité à l'établissement accompagné d'une obligation de reddition de comptes permettra d'augmenter l'efficacité du système d'éducation.

- *La législation*

Ainsi, pour les ordres d'enseignement primaire et secondaire, la Loi sur l'instruction publique (L.R.Q., c. I-13.1; a.96.13) stipule que:

*«Le directeur d'école assiste le conseil d'établissement dans l'exercice de ses fonctions et pouvoirs et à cette fin:*

*1.1° il coordonne l'élaboration, la révision et, le cas échéant, l'actualisation du plan de réussite de l'école.»*

Ce plan de réussite doit s'inscrire dans un projet éducatif qui tient compte du plan stratégique de la commission. En effet, la Loi sur l'instruction publique (L.R.Q., c. I-13.1; a.209.1) stipule aussi:

*«Pour l'exercice de ses fonctions et pouvoirs, chaque commission scolaire établit un plan stratégique couvrant une période de plusieurs années qui comporte (.):*

*2° les principaux enjeux auxquels elle fait face, entre autres en matière de réussite, qui tiennent compte des indicateurs nationaux établis par le ministre (.).»*

Cet organisme public a aussi l'obligation de rendre des comptes sur la réussite scolaire au ministre et à la population qu'elle dessert (L.R.Q., c. I-13.1; a.220):

*«La commission scolaire informe la population de son territoire des services éducatifs et culturels qu'elle offre et lui rend compte de leur qualité.*

*La commission scolaire prépare un rapport annuel qui rend compte à la population de son territoire de la réalisation de son plan stratégique.*

*(.) Une copie de ce rapport est transmise au ministre.»*

Aux Etats-Unis, une telle législation existe depuis 1965. En effet, la section 124 du «Elementary and Secondary Education Act» prévoyait qu'une école pouvait recevoir des fonds seulement si des procédures étaient adoptées pour évaluer l'efficacité de ses programmes d'enseignement (Echternacht, 1980; p.1).

Au niveau collégial, la Loi sur les collèges d'enseignement général et professionnel ( L.R.Q., c. 29; a.16.1 ) stipule:

*«Le conseil de chaque collège établit, en tenant compte de la situation du collège et des orientations du plan stratégique établi par le ministère de l'Éducation, un plan stratégique couvrant une période de plusieurs années. Ce plan comporte l'ensemble des objectifs et des moyens qu'il entend mettre en œuvre pour réaliser la mission du collège. Il intègre un plan de réussite, lequel constitue une planification particulière en vue de l'amélioration de la réussite des étudiants.»*

Enfin, le plan de réussite prend la forme d'un contrat de performance pour les universités. En effet, la Loi sur les établissements d'enseignement de niveau universitaire (L.R.Q., c. E-14.1; a.4.1) mentionne que:

*«Tout établissement (.) doit joindre aux états financiers qu'il transmet annuellement au ministre (.) un rapport sur sa performance établi conformément aux dispositions de l'article 4.6.»*

Et les dispositions de l'article 4.6 sont:

*«Le rapport sur la performance indique notamment:*

- 1° le taux de réussite, par secteur disciplinaire, pour chaque grade universitaire;*
- 2° la durée moyenne des études, par secteur disciplinaire, pour l'obtention d'un grade universitaire;*
- 3° les mesures prises pour l'encadrement des étudiants;*
- 4° les programmes d'activités de recherche.»*

D'une part, la législation et la reddition de comptes qui y est afférente permettent de réguler le système d'éducation au niveau global vers la finalité de la réussite du plus grand nombre d'individus possible. D'autre part, l'établissement a l'obligation, au niveau intermédiaire, de piloter son processus d'enseignement dans le but de le rendre le plus efficace possible.

- *Le processus d'enseignement*

L'efficacité du processus d'enseignement est implicite à l'élaboration d'un plan de réussite. Un processus est selon De Landsheere (1979, p.215) un: «ensemble de fonctions plus ou moins coordonnées et régulières et aboutissant à un phénomène qui est le produit.» Alors un processus d'enseignement devrait aboutir à la réussite du plus grand nombre de sujets<sup>1</sup>.

---

<sup>1</sup> Le mot sujet sera utilisé tout au long de cette recherche pour désigner à la fois l'élève du niveau primaire ou secondaire et l'étudiant de niveau collégial ou universitaire.

Dans un système d'éducation de masse<sup>2</sup>, le processus d'enseignement vise à ce que tous les sujets admis réussissent dans les délais prescrits. Or, en pratique ces derniers n'ont pas tous la même capacité pour mener à terme leurs études.

De Landsheere (1979, p.31) définit la capacité comme étant le :

*«pouvoir d'accomplir un acte, de produire un comportement ou un ensemble de comportements adéquats dans une situation donnée.»*

Quant à B.S. Bloom cité dans Legendre (1988, p.75), la capacité est plutôt la:

*«compétence de l'individu, en face de nouveaux problèmes et de nouvelles situations, à faire appel à des renseignements et des techniques déjà utilisés lors d'expériences antérieures.»*

Dans le cadre de cette recherche, le concept de capacité va être plutôt pris dans le sens du potentiel que possède un individu au début d'un processus d'enseignement. En effet, le plan de réussite d'un établissement doit proposer des modifications au processus d'enseignement pour l'adapter à la capacité des sujets et ce, afin d'augmenter leur chance de réussite. Nous dirons alors que le processus d'enseignement est différencié et non plus uniforme.

- *Le modèle d'un processus d'enseignement différencié*

Pour mener à bien sa tâche d'organiser l'enseignement différencié, l'éducateur responsable d'un processus d'enseignement doit pouvoir compter sur un modèle. Plusieurs définitions d'un modèle existent. Pour Lord et Novick (1968, p.16-17), un modèle est tout simplement un système logique abstrait alors que pour Legendre (1988, p.379), il est une:

*«représentation fonctionnelle et simplifiée d'une classe d'objets ou de phénomènes à l'aide de symboles, organisés en une forme plus ou moins structurée, dont l'exploration et la manipulation, effectuées concrètement ou par la pensée, entraînent une compréhension accrue et permettent l'énoncé d'hypothèses de recherche.»*

---

<sup>2</sup> Système qui rend l'éducation secondaire et supérieure accessible à toute la population.

Selon cette définition, le modèle d'un processus d'enseignement différencié devrait inclure comme objets : la mesure de la capacité des sujets et les décisions prises à partir de cette mesure. Mais avant d'élaborer un tel modèle, il serait bon de regarder un exemple dans un autre domaine.

Plaisance et Vergnaud (1993) proposent de comparer une situation en éducation avec une situation en médecine. Tout comme pour le médecin qui diagnostique une maladie et offre un traitement au patient, l'éducateur doit poser un diagnostic et offrir une aide si nécessaire. Dassa et Laurier (2003, p.119) viennent confirmer ce rôle du diagnostic en éducation :

*«..en tant que mécanisme de régulation, l'évaluation formative ne vise pas à porter un jugement sur le degré de développement des compétences ; elle doit plutôt, par le diagnostic, suggérer des avenues pour améliorer la qualité de l'enseignement/apprentissage.»*

Prenons l'exemple du test sanguin. À partir des résultats pour un patient, le médecin établit un diagnostic sur son taux de cholestérol; ce diagnostic se base sur le pronostic que le taux de cholestérol d'un patient permet de prédire une maladie cardio-vasculaire. Un diagnostic de cholestérol élevé devrait entraîner la décision de prescrire un traitement ou un médicament.

À l'instar du médecin qui fonde son diagnostic sur les résultats à un test, l'éducateur pourrait lui aussi utiliser les informations provenant d'un test de capacité administré au début du processus d'enseignement afin d'établir un diagnostic envers chaque sujet admis. Évidemment, il y a des différences entre les deux situations et plus particulièrement en ce qui concerne l'instrument de mesure. En effet, l'instrument utilisé en éducation mesure le comportement d'un être humain alors que le test sanguin mesure la composition chimique du sang.

- *Le graphe du processus d'enseignement différencié*

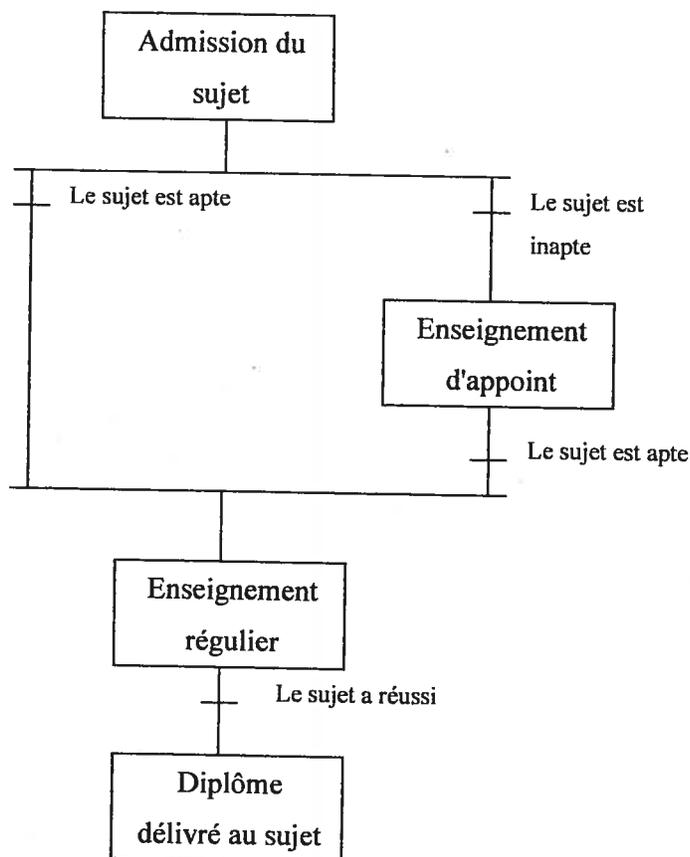
On peut s'inspirer de l'analogie avec une situation en médecine pour établir un modèle d'enseignement différencié dans le cadre d'un plan de réussite. Ainsi, en

éducation un diagnostic positif serait du type *le sujet est apte* et un diagnostic négatif du type *le sujet est inapte*. Le modèle d'enseignement différencié qui doit mettre en œuvre un plan de réussite devrait prendre la forme discursive suivante: «Si le sujet était apte ou s'il était inapte mais qu'il avait suivi un enseignement d'appoint alors il devrait réussir.»

Il devrait exister une relation directe de cause-à-effet entre la capacité et la réussite du sujet dans le sens qu'il devrait réussir s'il a la capacité requise, capacité qui est déterminée suite à un diagnostic positif. À l'inverse, un diagnostic négatif devrait diriger le sujet vers un enseignement d'appoint si on veut diminuer la possibilité d'échec.

Le modèle peut aussi être représenté sous la forme d'un graphe (voir figure 1). Martin (1986) proposa d'utiliser un graphe avec des symboles standards pour aborder une situation en éducation selon une approche systémique. Ce mode de représentation permet de voir les interactions entre les acteurs du système et les sujets. Le rectangle représente une étape où il y a une intervention des acteurs. Le système évolue dans le temps du haut vers le bas du graphe. Ainsi, l'admission est le premier acte sanctionné alors que la délivrance du diplôme en constitue le dernier. Les conditions d'évolution, aussi appelées transitions, sont reliées à l'état d'un sujet et sont indiquées par des arcs qui relient les interventions entre elles.

Ce graphe montre une divergence et une convergence sélectives à deux branches, la sélection étant déterminée par le diagnostic sur la capacité du sujet lors de l'admission. Un diagnostic positif fera cheminer le sujet dans la voie régulière de l'enseignement de masse alors qu'un diagnostic négatif l'orientera vers un enseignement d'appoint.



**Figure 1: Le modèle d'un processus d'enseignement différencié**

Au Québec, comme on le verra ci-dessous, l'expérimentation de ce modèle se fait déjà mais d'une manière implicite.

- Les expériences québécoises

Au niveau primaire, on diagnostique les élèves en difficulté d'apprentissage au début d'un cycle d'apprentissage de deux ans et l'enseignement d'appoint est dispensé par un orthopédagogue. Au secondaire, on dirigera les élèves en difficulté vers un cheminement particulier. Dans certains collèges, on fait même passer une dictée aux nouveaux étudiants admis et on dirige ceux qui ont des faiblesses en français vers un enseignement d'appoint. Dans quelques facultés universitaires non contingentées, on offre des cours de mise à niveau après avoir fait passer un test de capacité aux étudiants admis ; c'est le cas de l'École Polytechnique de Montréal et des facultés des sciences de l'éducation des universités francophones du Québec. Plus particulièrement dans ces dernières,

on fait passer des tests de français pour les futurs enseignants ; mais récemment la validité de ces instruments a été mise en doute (Voisard, 2004; p.A4):

*«C'est un vrai fouillis. Les tests qui visent à savoir si les aspirants au métier d'enseignant connaissent ou non leur français ont tout pour embrouiller. Que ce soit à l'université ou à l'embauche.*

*Ce n'est donc pas sans raison que le ministère de l'Éducation (MEQ) (.) et les doyens des départements ou facultés des sciences de l'éducation de toutes les universités se sont assis à la même table. Le but: évaluer les différents tests.»*

Le cas des tests de français représenterait-il la face visible du problème plus vaste de la validité de la mesure critériée en éducation ?

### ***3. Le problème de validité de la mesure critériée en éducation***

Le médecin pose un diagnostic de taux de cholestérol élevé chez un patient parce que cette variable dépasse un certain seuil, il fait alors une interprétation critériée de la mesure. De même, l'éducateur doit auparavant faire une interprétation critériée de la mesure de la capacité d'un sujet avant de poser un diagnostic. Toutefois, il faut que cette mesure de capacité ait une validité pronostique et il est plus difficile de prouver la validité d'un instrument de mesure en éducation que la validité d'un test sanguin parce que les conditions d'administration sont moins stables.

- ***L'instrument de mesure***

En éducation, une technologie particulière est utilisée pour recueillir des données sur les individus. Plus précisément, les données sont recueillies en administrant un questionnaire à un groupe de sujets dans des conditions déterminées. Dans le cas particulier d'un processus d'enseignement, l'interprétation des données obtenues devrait résulter en l'évaluation de l'apprentissage des sujets.

L'évaluation, selon Cardinet (1986, p.74), a une fonction qui dépend du moment où la mesure a été effectuée ; ainsi nous aurons une fonction d'orientation en début de processus d'enseignement, une fonction de régulation pendant et une fonction de certification à la fin. Dans le cas du processus d'enseignement différencié, nous sommes intéressés à poser un diagnostic au début de ce processus pour pouvoir orienter par la suite le sujet vers un cheminement adapté. Cependant, le type de validité exigée pour un instrument de mesure va différer d'une fonction d'évaluation à l'autre.

Dans le milieu de l'éducation, les instruments de mesure prennent différentes appellations. Les trois principales en sont : l'examen, l'épreuve et le test.

L'examen est un instrument de mesure à usage restreint qui est normalement conçu par l'enseignant et qui est administré à un ou plusieurs groupes-classes. Lorsque ces instruments sont utilisés pendant le processus d'enseignement, l'interprétation des résultats a une fonction de régulation, alors leur fidélité et validité n'ont pas à être démontrées. De Landsheere (1979, p.281) affirme qu'en anglais, le mot test signifie un examen alors qu'en français, une épreuve ou un examen ne peuvent être appelés test que s'ils sont standardisés.

L'épreuve, quant à elle, est un instrument qui exige un minimum de validité. Au Québec, le terme «épreuve» est utilisé par la Ministère de l'éducation (MEQ) et les commissions scolaires pour désigner un instrument de mesure administré à un grand groupe à la fin d'un processus d'enseignement. L'interprétation des résultats a ici une fonction de certification et on exige une validité de contenu pour démontrer la congruence entre les objectifs d'apprentissage et les items. Bref, l'épreuve est un instrument pour cueillir des données sur un grand groupe de sujets sans nécessairement que ce soit une population. Lorsque cet instrument est administré dans un contexte d'orientation, on exigera qu'il ait une validité pronostique.

Alors que le test est selon Legendre (1988, p.604) : «une épreuve ou série d'épreuves psychotechniques normalisées impliquant une tâche à remplir, identique pour tous les sujets examinés.» Pour De Landsheere (1979, p.265), le test est :

*«une situation standardisée servant de stimulus à un comportement qui est évaluée par comparaison avec celui d'individus placés dans la même situation, afin de classer le sujet, soit quantitativement, soit typologiquement.»*

Donc, le test est un ensemble d'items agissant comme un stimulus qui nécessite une réponse de la part du sujet sous la forme d'une tâche à remplir. Selon De Landsheere (1979, p.265) : «Pour mériter le nom de test, un examen doit être standardisé, fidèle, valide et étalonné.» La standardisation d'un test est «la définition précise des modalités d'utilisation» pour De Landsheere (1979, p.248). On retiendra que le test a un caractère plus universel que l'épreuve.

Dans le cas d'un processus d'enseignement différencié particulier à un établissement et où l'on a besoin de mesurer la capacité des sujets pour poser un diagnostic, on utilisera le terme «épreuve». Néanmoins, il faudra démontrer sa validité pronostique.

- *Pronostic et diagnostic*

Un pronostic est la «prédiction d'un comportement, d'un résultat.» (De Landsheere, 1979 ; p.218). Le Petit Larousse (1991) définit le verbe «prédire» par:

*«Annoncer d'avance ce qui doit se produire soit par intuition ou divination, soit par des règles certaines, soit par conjecture ou raisonnement.»*

La prédiction par intuition ou divination n'a pas une valeur scientifique et la prédiction avec des règles certaines n'est pas réaliste. Par contre, une conjecture est une supposition fondée sur des probabilités et il devient possible d'établir un modèle à partir des données provenant d'un grand nombre de sujets. Alors, la

valeur passée d'une ou plusieurs variables indépendantes pour un sujet détermine la valeur future d'une variable dépendante pour ce même sujet avec une marge d'erreur. Avec le modèle de la figure 1, le résultat attendu est la réussite du sujet après le processus d'enseignement. Donc, en éducation la capacité du sujet pourrait prédire sa réussite future. Toutefois, Cardinet (1986, p.101) fait une mise en garde pour un pronostic à long terme en éducation :

*«Cependant, quelles que soient les précautions prises, la précision de mesures pronostiques restera faible. À l'expérience, on verra qu'il est difficile de prévoir le classement des élèves, même dans une classe identique, à plus de trois ans de distance, par exemple. Ceci résulte naturellement de la multitude d'effets imprévisibles et incontrôlables qui viennent perturber ou faciliter l'apprentissage de chaque élève.»*

La mesure pronostique permet d'attribuer une valeur à la variable indépendante principale, en l'occurrence la capacité du sujet, tandis que la variable dépendante se confond au critère à prédire, c'est-à-dire la réussite. De plus, De Landsheere (1988, p.24) associe la réussite à une compétence minimale que le sujet atteint ou non. Donc, pour le cas du modèle d'un processus d'enseignement différencié, la capacité a un caractère continu alors que la réussite est plutôt de nature dichotomique (réussite ou échec) lorsqu'elle est associée au fait d'obtenir un diplôme ou non.

Le pronostic de la réussite à partir de la mesure de la capacité permettra d'établir un score de césure pour émettre un diagnostic. Par exemple, en médecine un taux de cholestérol de 5,2 constitue un seuil pour émettre un diagnostic. Il pourrait en être de même en éducation où le score de césure permettrait d'émettre un diagnostic de *sujet apte* ou *sujet inapte*. Par contre, la validité du diagnostic reposerait sur la validité de la mesure pronostique.

- La validité d'un instrument de mesure

La définition classique de la validité d'un instrument de mesure est le fait que cet instrument mesure véritablement ce qu'il prétend mesurer. Une définition plus récente proposée par Messick (1993) énonce que la validité d'un test est un jugement qui démontre jusqu'à quel point l'évidence empirique et le construit théorique supportent les interprétations des scores.

Selon De Landsheere (1979, p.55), le construit est un concept hypothétique, une :

*«entité ou modèle imaginé pour expliquer certains phénomènes, pour les rattacher à un facteur causal inobservable.»*

Les réponses du sujet aux items du test constituent l'observation alors que le construit est ce que nous voulons mesurer chez ce même sujet. Dans la présente recherche, le construit sera rattaché à la capacité. En d'autres mots, poser la question de la validité d'un test, c'est donc se demander si ce test mesure bien la capacité et si les résultats obtenus peuvent être confirmés dans la réalité. Cette confirmation ne peut prendre effet que si on réfère le score à un critère qui peut être concomitant ou situé dans le futur.

En sciences de la nature par exemple, on étalonne très souvent un instrument de mesure en comparant sa lecture avec une grandeur étalon ; cela signifie établir la validité concomitante de l'appareil. Dans les sciences sociales et de l'éducation la situation est tout à fait différente car il n'existe pratiquement pas de grandeur étalon. Le repère n'est plus un objet physique mais plutôt un comportement observable après un certain délai. Ainsi, dans le modèle du processus d'enseignement différencié de la figure 1 la réussite du sujet est à la fin de ce même processus. Donc, le critère, que nous assimilions plus tôt à un objectif d'apprentissage, sera relié dans la réalité à la réussite et constituera le repère pour établir la validité pronostique de l'épreuve diagnostique administrée au début du processus.

Ainsi pour effectuer une étude de validité, l'éducateur, responsable d'un processus d'enseignement, devrait pouvoir compter sur un modèle élaboré à partir d'une théorie de la mesure pertinente.

- Une théorie de la mesure pertinente

Une théorie est (Legendre; 1988, p. 609):

*« un ensemble de concepts, de définitions et de propositions,(.) qui propose une vue systématique d'un phénomène,(.) dans le but d'expliquer et de prédire. »*

Le but d'une théorie de la mesure est d'expliquer les scores à partir du construit. Il existe actuellement trois théories de la mesure qui sont appliquées en éducation: la théorie classique des tests (TCT), la théorie de réponse à l'item (TRI) et la théorie de la généralisabilité (TG). Cependant, les auteurs Fredericksen *et al*, (1993, p.19) doutent de la pertinence de ces théories dans un cadre moderne de la mesure en éducation. En effet, avec l'émergence de la psychologie cognitive depuis les dernières décennies, la pertinence de ces théories a été questionnée.

Avec la TCT, le score est indépendant du schéma de réponses d'un sujet à un test alors qu'avec la TRI, le score devrait idéalement ne dépendre que du schéma de réponses ; dans tous les cas, une théorie de la mesure doit rendre l'interprétation du score explicite et non équivoque. Dans ce sens, Brown (1980, p.74) affirme qu'il existe globalement trois façons principales pour interpréter les scores à un test : la référence à une norme, la référence à un contenu et la référence à un critère.

Les scores référés à une norme sont les plus couramment employés et leur interprétation est faite en se référant à la performance des autres sujets d'un groupe. Cette interprétation est très utilisée en psychométrie pour mesurer un trait psychologique comme par exemple le quotient intellectuel. En effet, la performance du sujet peut être comparée à une population pour juger de sa «normalité» un peu comme on le fait pour les mesures physiques d'un enfant.

À l'inverse, les scores référés à un contenu sont interprétés en termes de niveau de performance du sujet sur un continuum de rendement mesuré par le test. Cette interprétation est utilisée autant en psychométrie qu'en éducatrice (Cardinet, 1986 ; p.61) pour mesurer la capacité en éducation. Ainsi, un sujet pourrait obtenir un score dans une discipline scolaire qui serait en termes de niveau d'habileté.

Enfin, les scores référés à un critère sont interprétés en termes de performance à un critère externe. Généralement, ils ne peuvent prendre que deux valeurs : maîtrise ou non-maîtrise. Cette interprétation est aussi utilisée en éducatrice mais cette fois-ci, pour poser un diagnostic à partir de la capacité mesurée.

La théorie classique des tests (TCT) a surtout été utilisée pour une interprétation normative des scores à un test. Elle permet par exemple de déterminer le quotient intellectuel d'un sujet à partir de son score à un test de Q.I.

Quant à la théorie de réponse à l'item (TRI), elle propose une interprétation des scores sur un continuum. Plus spécifiquement, cette théorie permet de mesurer l'habileté du sujet pour un domaine disciplinaire donné ; de ce fait, nous pouvons affirmer qu'il s'agit d'une interprétation des scores en référence à un contenu. Cette théorie est surtout très utilisée dans le cas de tests adaptatifs effectués par ordinateur car alors elle permet d'estimer avec une grande précision l'habileté d'un sujet. L'un des tout premiers modèles issus de la TRI a été le modèle de Rasch qui selon Andrich (1978, p.449), permet d'étalonner la mesure sur une échelle linéaire lorsqu'on compare les réponses d'un sujet à plusieurs items. Plus il y a d'items qui mesurent la performance d'un sujet, meilleure est la précision.

La théorie de la généralisabilité est une théorie de la mesure qui tient compte du contexte et de ce fait, tente d'expliquer les scores à partir de plusieurs facettes de la mesure. Elle permet d'interpréter les trois types de scores mentionnés y

compris les scores en référence à un critère mais dans ce dernier cas, la perspective est limitée.

Ces trois théories de la mesure ne sont que d'une utilité limitée dans le cas d'une interprétation critériée des scores pour diagnostiquer le sujet au début du processus d'enseignement. Toutefois, une théorie en émergence pourrait combler le vide actuel, il s'agit de la théorie des ensembles flous (TEF). Le pronostic ne se fonderait plus alors sur une conjecture mais plutôt sur un raisonnement.

#### *4. Les visées de la recherche*

La théorie des ensembles flous (TEF) a déjà été utilisée dans plusieurs domaines et même dans les sciences de l'éducation.

- *La théorie des ensembles flous, un nouveau paradigme*

Selon Smithson (1989), la théorie des ensembles flous (TEF) s'inscrit dans un nouveau paradigme et épouse le langage du sens commun, elle devrait permettre l'élaboration de modèles mieux adaptés aux situations complexes que l'on retrouve en éducation et en sciences sociales. Pour Legendre (1988, p. 443):

*«Un paradigme est un super modèle, une conception globale et cohérente d'un domaine de savoir et d'activité, laquelle vision de l'esprit est partagée par les chercheurs et les praticiens qui oeuvrent dans ce domaine».*

Ainsi, la théorie des ensembles flous (TEF) se distingue des autres théories de la mesure par le fait qu'elle ne s'inscrit pas dans le paradigme des probabilités. Zadeh (1965) fut le premier à établir les fondements de la théorie des ensembles flous (TEF) qui permettent d'interpréter des données en raisonnant avec une logique floue. Le dictionnaire Le Petit Larousse (1991) définit ainsi la logique:

*«La logique constitue une langue, c'est-à-dire un système de signes avec les règles de leur emploi. Cette langue est constituée d'un système de symboles et de variables liés par des opérateurs qui déterminent la structure interne des propositions et les relations entre les propositions.»*

Avec la logique classique, chaque proposition d'un discours peut être évaluée comme étant vraie ou fausse. Pour Kosko (1993, p.8) : «Le monde est gris mais la science est en noir et blanc.»<sup>3</sup> Les modèles traditionnels ne laissent que peu de place pour l'incertitude dans l'interprétation des données alors que la logique floue ajoute des adverbes pour qualifier une proposition d'un discours comme par exemple: certainement fausse, possiblement fausse, incertain, possiblement vraie, certainement vraie. Autrement dit, la logique floue teintera en gris la science pour qu'elle ressemble davantage à la réalité.

La théorie des ensembles flous (TEF) considère un ensemble «ordinaire» comme un cas particulier ; le corollaire étant que la logique classique est un cas particulier de la logique floue. Avec la logique classique le diagnostic est une proposition que l'on peut évaluer par vraie ou fausse alors qu'avec la logique floue cette proposition pourra être qualifiée en tenant compte des réponses du sujet au test.

Logiquement, nous affirmerions qu'un sujet ayant obtenu un score élevé à l'épreuve de capacité serait certainement diagnostiqué apte à entreprendre le processus d'enseignement régulier et à l'inverse, un sujet ayant obtenu un faible score serait certainement diagnostiqué inapte. La logique floue va tenir compte de l'incertitude du diagnostic pour un sujet dont le score brut se situe près du score de césure. Ainsi, nous pourrions avoir cinq messages distincts pour le diagnostic: certainement apte, possiblement apte, ni apte ni inapte, possiblement inapte, certainement inapte. Dans les faits, nous aurons un degré d'appartenance

---

<sup>3</sup> Traduction libre de : *The world is gray but science is black and white.*

à un ensemble flou qui graduera de 0 à 1; les messages «certainement inapte» et «certainement apte» donnant respectivement des valeurs de 0 et 1. Le diagnostic serait alors plus nuancé et serait fonction du score au test.

- Objectif de la recherche : élaborer un modèle pronostique pour une mesure référée au critère de réussite en recourant à la théorie des ensembles flous

La théorie des ensembles flous (TEF) propose une méthode inductive qui devrait permettre d'établir un modèle. Une méthode inductive est un (Legendre, 1988; p.372):

*«mode de raisonnement logique qui consiste à généraliser à un ensemble les données particulières obtenues à partir d'un nombre limité d'éléments.»*

L'originalité de cette théorie est qu'on peut produire directement un discours à partir de données empiriques et qu'il est possible d'y associer un modèle avec des opérateurs logiques pour prédire le critère.

Les théories usuelles des tests (TCT et TRI) utilisent la prédiction par conjecture pour valider un instrument de mesure. Avec la théorie des ensembles flous (TEF), la prédiction se fera par un raisonnement en logique floue. Ce raisonnement en logique floue devra quand même refléter la réalité pour la très grande majorité des cas possibles.

- L'expérimentation du modèle pronostique

Depuis plusieurs années, les autorités de l'École Polytechnique de Montréal font passer une épreuve de capacité en mathématiques pour tous les nouveaux étudiants admis. Des données ont été amassées sur les réponses des sujets à chaque item de cette épreuve et de ce fait, elles seront utilisées pour expérimenter le modèle. La capacité et la réussite deviendront des variables et plus particulièrement, elles seront considérées comme des attributs pour des ensembles flous.

L'École Polytechnique de Montréal est un établissement universitaire reconnu au sens de la Loi sur les établissements d'enseignement de niveau universitaire (L.R.Q., c. E-14.1; a.1). Il offre plusieurs programmes d'études en génie qui conduisent à la profession d'ingénieur. Sur le site internet de cette institution, on peut y lire que les programmes de formation comprennent un tronc commun d'une durée normale d'un an, un tronc de spécialité d'une durée de deux ans et demie, et un dernier trimestre de projet et de cours d'orientation. Cela mène normalement à un diplôme de baccalauréat en ingénierie (B. ing.) après quatre ans.

Selon les autorités de cette institution, l'étudiant qui veut entreprendre des études d'ingénieur doit démontrer une compétence pour les mathématiques et les sciences car ces disciplines constituent le fondement des programmes de génie. D'ailleurs, les critères d'admission exigent la réussite de cours de niveau collégial en mathématiques, physique et chimie.

Présentement, aucune décision n'est prise à partir des résultats obtenus à l'épreuve. En effet, cet instrument n'a pas une fonction de sélection, ni de classement; il vise plutôt à informer le sujet sur ses lacunes. Le modèle proposé devrait pouvoir apporter une réponse à la question de la recherche.

- *La question de la recherche : «Est-ce que le modèle, issu de la théorie des ensembles flous et appliqué aux scores obtenus par une cohorte de l'École Polytechnique de Montréal à une épreuve diagnostique de mathématique, permet de prédire adéquatement la réussite des sujets dans leur programme?»*

Dans le cas où le modèle ne pourrait prédire la réussite dans le programme, il faudrait chercher s'il peut prédire la réussite d'un cours de mathématiques dispensé en première année. Ainsi, on aurait une idée si le construit «capacité en mathématiques» était trop spécifique. On appliquera aussi le modèle à une autre cohorte pour vérifier s'il peut prédire aussi bien la réussite. Cela nous permettra de discuter de sa généralisation.

Assez souvent, les hypothèses de recherche sont amenées à la fin de la problématique mais ce ne sera pas le cas pour la présente recherche. En effet, la traduction de la question en hypothèses de recherche ne peut se faire à ce moment-ci puisque les concepts de la théorie des ensembles flous qui sont nécessaires à leur élaboration ne seront abordés que dans le chapitre second sur le cadre conceptuel. Aussi, les hypothèses de recherche seront énoncées dans le chapitre trois portant sur la méthodologie.

En résumé, le modèle pronostique devrait devenir un outil pour l'éducateur lorsqu'il oriente l'élève ou l'étudiant dans un processus d'enseignement différencié.

## Chapitre second

### Le cadre conceptuel

Quels concepts sont en jeu lorsqu'on désire prédire la réussite? En quoi la théorie des ensembles flous peut faire mieux que les théories de la mesure traditionnelles pour prédire la réussite? Voilà des questions que nous voulons explorer dans le cadre de ce chapitre.

#### *1. La mesure en éducation*

Selon le chapitre précédent, le problème d'établir un pronostic sur la réussite est directement relié au besoin d'avoir une mesure valide de la capacité. En effet, la capacité constituera la variable pronostique principale alors que la réussite sera plutôt la variable prédite. Nous utilisons l'expression «variable pronostique» plutôt que «variable prédictive» parce qu'un pronostic est une prédiction qui se fonde principalement sur un diagnostic. Ainsi, une erreur de mesure pour la variable pronostique entraînerait sans doute une erreur de prédiction.

- *Le concept de mesure*

En éducation, le concept de mesure se distingue de celui véhiculé généralement dans les sciences de la nature. Blais (2003, p.13) suggère deux définitions de la mesure:

*«La première définition, celle qui met de l'avant les concepts d'unité de mesure et de rapports entre des grandeurs, est considérée comme la définition classique (.). La deuxième définition, celle que l'on attribue à Stevens (1951), est en fait une version operationaliste d'une théorie de la mesure où les nombres représentent des propriétés des attributs visés.»*

Pour nous, le concept de mesure ne peut pas se restreindre à la définition classique, il faut plutôt l'étendre à la lumière de la définition de Stevens. Les

auteurs Flood et Carson (1990) adoptent aussi cette position lorsqu'ils prétendent généraliser le concept de mesure à une science des systèmes s'inscrivant dans une perspective multidisciplinaire. Blais (2003, p.16) vient préciser la pensée de Stevens:

*«En effet, pour Stevens l'élément central de la mesure consiste à établir une règle bien définie pour attribuer les nombres, le problème central de la mesure devenant le développement de ces règles opérationnelles.»*

Ainsi, avec la proposition de Stevens pour produire une mesure, il devra y avoir une règle pour assigner des nombres à des attributs qui ont la totalité ou une partie seulement de leurs propriétés ; de plus, une règle d'assignation des nombres correspondra à une échelle de mesure (Flood et Carson, 1990 ; p.65). Plus particulièrement, les échelles de mesure viendront indiquer quelles propriétés des nombres seront assignées aux attributs du monde réel.

- Les échelles de mesure

Les nombres sont des éléments qui ont des propriétés nettement définies et de ce fait, ils constituent le fondement de la plupart des modèles de mesure. Plus particulièrement, on retrouve au tableau I les principales propriétés de ces nombres ; les variables a, b, c, d pouvant être des nombres quelconques.

Tableau I: Propriétés des nombres

▪ <b>Identité</b>	▪ <b>A1</b>	Soit $a = b$ ou $a \neq b$
	▪ <b>A2</b>	Si $a = b$ alors $b = a$
	▪ <b>A3</b>	Si $a = b$ et $b = c$ alors $a = c$
▪ <b>Ordre</b>	▪ <b>A4</b>	Si $a > b$ alors $b < a$
	▪ <b>A5</b>	Si $a > b$ et $b > c$ alors $a > c$
▪ <b>Additivité</b>	▪ <b>A6</b>	$a + b = b + a$
	▪ <b>A7</b>	Si $a = c$ et $b = d$ alors $a + b = c + d$
	▪ <b>A8</b>	$(a + b) + c = a + (b + c)$
▪ <b>Neutralité du zéro</b>	▪ <b>A9</b>	$a + 0 = a$
	▪ <b>A10</b>	Si $a = c$ et $b > 0$ alors $a + b > c$

Afin de regrouper ces propriétés selon des règles d'assignation, Stevens (dans McDonald, 1999, p.409) propose quatre échelles de mesure, soient les échelles : nominale, ordinale, d'intervalle et de rapport. Selon le même auteur (McDonald, 1999, p.410), chacune de ces échelles est caractérisée par un sous-ensemble de propriétés pour les nombres.

Ainsi, l'échelle nominale regroupe le sous-ensemble des propriétés d'identité de A1 à A3; l'opération de mesure consiste alors à classer des éléments. En effet, le nombre a le symbole d'un numéro et ne sert qu'à identifier un élément ou un sujet. Un exemple de variable mesurée avec cette échelle est le genre; le nombre 0 pouvant représenter le sexe masculin et le nombre 1 le sexe féminin. Néanmoins, le fait qu'un sujet ait réussi un cours de mise à niveau en mathématique ne constitue pas une variable mesurée sur une échelle nominale car il y a un rapport d'ordre entre la réussite et l'échec. Par contre, un numéro de

téléphone et une adresse internet représentent des variables nominales alors que ce n'est pas le cas pour une adresse postale car il y a aussi une notion d'ordre associée à la position d'une maison sur une rue.

Quant à l'échelle ordinale, elle englobe plutôt le sous-ensemble des propriétés A1 à A5, c'est-à-dire que l'on ajoute les relations d'ordre car ici l'opération de mesure va plutôt consister à ordonner les éléments et le nombre prend le symbole de rang. Par exemple, l'échelle de Likert utilisée pour mesurer l'attitude constitue une échelle ordinale ; on associe souvent cette échelle à une échelle descriptive car chaque échelon est décrit par une expression qui qualifie un énoncé, par exemple: peu satisfait, satisfait, très satisfait.

Le sous-ensemble des propriétés d'additivité A6 à A8 vient compléter les propriétés précédentes pour caractériser l'échelle d'intervalle; la mesure à ce niveau va signifier que l'on peut compter un nombre d'unités de mesure. Cependant, le nombre 0 n'a pas une valeur neutre avec cette échelle; en d'autres mots, une valeur de 0 ne signifie pas qu'il y a absence complète du trait mesuré. L'échelle de température en degrés Celsius constitue un exemple d'échelle d'intervalle ; en effet, une température de 0°C pour un corps ne signifie pas que ce corps ne dégage plus d'énergie, par contre, un ajout d'une même quantité spécifique d'énergie à un corps pour deux températures initiales distinctes cause une même augmentation de la température en °C.

Enfin, l'échelle de rapport propose un modèle de mesure avec toutes les propriétés des nombres, de A1 à A10; en plus de compter des unités de mesure, la référence de l'échelle est le zéro absolu. L'échelle de température en Kelvin représente une telle échelle; ici, une valeur de 0 Kelvin pour un corps signifie qu'il ne dégage plus d'énergie et on dit que l'échelle possède la propriété de neutralité du zéro.

La définition classique de la mesure se restreint aux échelles d'intervalle et de rapport qui permettent de mesurer un attribut en produisant des données

quantitatives. Les échelles nominale et ordinale, quant à elles, permettent d'apprécier des attributs en produisant des données qualitatives que Van der Maren (1996, p.85) définit par:

*«Les données qualitatives regroupent toutes les données non métriques. Elles comportent tout ce qui est texte, images et sons, mots exprimant des catégories et des jugements catégoriels.»*

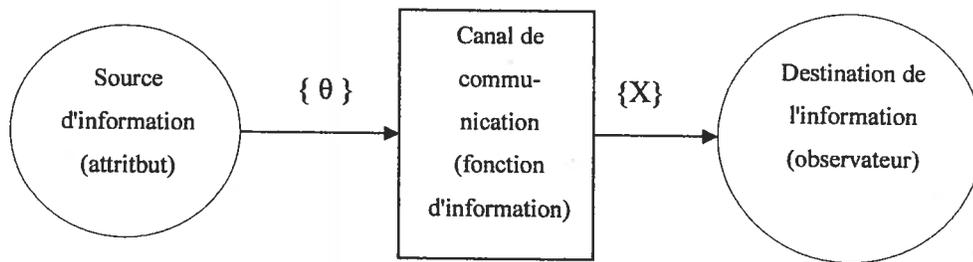
Une analyse des données quantitatives obtenues à partir d'un échantillon de sujets permet d'inférer les résultats à une population alors qu'une analyse des données qualitatives ne peut être que descriptive. En éducation et dans les sciences sociales en général, on ne peut se permettre de se laisser éblouir par l'apparence de supériorité pour la mesure quantitative ; en effet, on doit, au préalable, s'assurer de la fidélité et de la validité des instruments avant d'analyser les données.

- La fidélité de la mesure

Les auteurs Lord et Novick (1968, p.23) affirment qu'un modèle de mesure en psychologie et en éducation est au moins une fonction entre une variable dépendante observable et une variable latente ou non observable. En fait, le schéma de réponses obtenu à partir de l'administration d'un instrument de mesure à un sujet peut constituer la variable observable alors que la variable latente se confond à l'attribut visé par l'opération de mesure.

Cette fonction d'un modèle de mesure prétend expliquer la variable latente à partir de la variable observable. Cependant, dans la pratique il existe toujours une erreur entre la valeur du score observé et la valeur de l'attribut mesuré; plus cette erreur sera petite, plus la mesure reflètera fidèlement l'attribut du monde réel.

Ainsi, une façon de voir la fidélité d'un instrument est de l'aborder sous l'angle de la théorie de l'information énoncée par Shannon (dans Fink et McKenzie, 1975; p.4-2). Le schéma de la figure 2 montre l'interaction entre les concepts pour cette théorie.



**Figure 2:** Le schéma de la communication

On peut comparer la source d'information à l'attribut mesuré par l'instrument. En effet, le canal de communication se confond à l'instrument lui-même où chaque message  $\{\theta\}$  est encodé en un schéma de réponses. De plus, les différentes interférences causées lors de la mesure sont représentées par un signal avec du bruit  $\{X\}$  que reçoit l'observateur associé à la destination de l'information. Enfin, on caractérise la qualité de l'instrument avec une fonction d'information désignée par  $I(\theta)$ .

Le bruit a pour effet d'augmenter l'incertitude, aussi appelée l'entropie, dans le message original et conséquemment, entraîne la perte d'une certaine quantité d'information. C'est ainsi que parlant d'une solution à un problème de mesure, Tourneur (1989, p.59) compare la fidélité à un rapport signal-à-bruit, le signal étant le message observé à la destination. La fidélité est donc cette propriété de pouvoir inférer le message original à partir du signal observé.

En éducation et en sciences sociales, la mesure implique très souvent des personnes et l'instrument est utilisé pour recueillir des données sur un attribut de l'individu ; d'où la possibilité d'avoir plusieurs sources d'erreurs. Par exemple, le sujet peut ne pas toujours être dans une condition physique et mentale adéquate lors de l'administration de l'épreuve, des items peuvent être ambigus ou encore le sujet peut s'adonner au plagiat; alors les données recueillies sont entachées d'erreurs.

La fidélité de la mesure constitue une condition nécessaire mais non suffisante pour s'assurer que la variable observée reflète l'attribut mesuré. En effet, l'instrument peut être fidèle mais mesurer autre chose, alors se pose le problème de la validité. Par exemple, une épreuve qui comprend presque exclusivement des items sur le calcul différentiel mesure d'une façon trop restreinte l'attribut «capacité en mathématique» et à l'inverse, si elle inclut des items de physique et de chimie, elle mesure ce même attribut d'une façon trop large.

- La validité de la mesure

Poser la question de la validité d'un instrument de mesure, c'est se questionner sur son exactitude ou autrement dit, sur son étalonnage. Alors, comment pouvons-nous ancrer dans le monde réel des repères de l'échelle de mesure? Dans le domaine de la physique par exemple, nous avons l'échelle de température en degrés Celsius qui est une échelle d'intervalle. Or, pour étalonner un thermomètre, on peut se référer à la fonte de la glace pure à 0°C et à l'ébullition de l'eau pure à 100°C. Le coefficient d'expansion thermique du mercure nous permettra également de construire un instrument qui nous donne une valeur valide pour une large gamme de températures.

Les instruments que l'on retrouve en éducation et en sciences sociales diffèrent des instruments du domaine de la physique parce qu'on ne peut utiliser une loi pour procéder à leur étalonnage, on a plutôt recours à différents types de validité. Il y a tout d'abord la validité de contenu qui se fonde sur l'analyse rationnelle du contenu de l'instrument. Ensuite, il y a la validité référée à un critère (Allen et Yen, 1979 ; p.97) où les données issues de l'administration de l'instrument sont confrontées aux données provenant d'un critère externe faisant office de repère. Ainsi on qualifie la validité référée à un critère de concomitante lorsqu'il y a une cueillette simultanée des données pour le critère et pour l'attribut mesuré par l'instrument alors qu'on qualifiera la validité de pronostique ou prédictive quand la saisie des données issues du critère s'effectue après l'administration de l'instrument.

Alors que beaucoup d'auteurs appuient ces différentes catégories de validité, Messick (1993) prétend unifier le concept de validité d'un instrument sous le vocable de validité de construit qui comprend deux aspects principaux : la déduction logique et la vérification empirique. Il affirme que pour valider un instrument, il faudra en premier lieu énoncer une hypothèse de construit en effectuant une déduction logique; on peut associer ce premier aspect à la validité de contenu. Par la suite, on vérifie cette hypothèse par l'inférence d'une relation avec un critère constituant une bouée d'ancrage dans le monde réel; nous désignons ce second aspect sous le vocable de validité pronostique ou prédictive.

En éducation, l'interprétation des données provenant d'une opération de mesure sert normalement à l'évaluation. Aussi, pour questionner la validité d'un instrument, il faut connaître au préalable laquelle des fonctions d'évaluation est préconisée parmi l'orientation, la régulation ou la certification. De plus, Cardinet (1986, p.34) distingue la fonction d'orientation de celle de sélection par le fait qu'avec la première tous les sujets sont pris en compte.

Dans le cadre de cette recherche, nous visons à établir la validité pronostique d'une épreuve de mathématique administrée lors de l'admission dans un programme ; la réussite des sujets pour ce même programme constituera le critère de référence. Cette épreuve est administrée depuis plusieurs années et des experts en mathématique ont construit les différentes versions de questionnaire à partir d'une banque de 985 items ; chacune de ces versions contient 60 items à réponses choisies qui se répartissent en six domaines : fonctions élémentaires (13), géométrie (10), trigonométrie (7), matrices et vecteurs (12), calcul différentiel (12) et calcul intégral (6). Par conséquent, nous énonçons le postulat que cette épreuve a déjà une certaine validité de contenu. De plus, la confirmation de la qualité pronostique de cette épreuve vient confirmer sa validité pour mesurer le construit «capacité en mathématique» selon Messick (1993). Enfin, la fonction d'évaluation en est une d'orientation.

Mais comment devons-nous procéder pour établir cette validité? La mise en pratique d'un tel processus de validation doit reposer sur une théorie de la mesure.

- Une théorie de la mesure

Le contexte de la mesure en éducation a toujours suivi celui de la psychologie. En 1980, l'organisme américain «National Council on Measurement in Education» a publié un guide sur l'utilisation des tests dans ces domaines (Brown, 1980). On peut y lire que l'aspect le plus important de la mesure en éducation et en psychologie consiste sans doute dans l'interprétation des scores (Brown, 1980 ; p.71). Ce guide décrit également trois types d'interprétation des scores: l'interprétation normative où la référence est la performance des autres personnes dans un groupe, l'interprétation en référence à un contenu où l'accent est mis sur le niveau de performance par rapport au contenu mesuré par le test et enfin, l'interprétation critériée où les scores sont interprétés en termes de performance à un critère externe (Brown, 1980 ; p.74-75). Conséquemment, le choix d'une théorie de la mesure devrait tenir compte du type d'interprétation que l'on veut faire des scores.

Nous avons assisté à la naissance d'un nouveau courant en psychologie depuis les trente dernières années, soit le cognitivisme. En effet, le discours en éducation a importé de ce courant les concepts de compétence, d'habileté ainsi que ceux de connaissances déclarative et procédurale. Les auteurs Snow et Lohman (1993) montrent comment la recherche dans le domaine de la psychologie cognitive a eu des implications sur la mesure; désormais, il ne faut plus mesurer seulement l'acquisition des connaissances mais aussi la «structure mentale» du sujet. Les nouveaux instruments doivent permettre un diagnostic pour un changement qualitatif ou un changement quantitatif. Or, Oosterhof (1990) affirme que l'on doit faire une interprétation critériée d'une mesure produite par une épreuve diagnostique. Par conséquent, le choix d'une théorie de la mesure devrait tenir compte de cet aspect.

À l'instar de toutes les sciences, nous utilisons en éducation et en sciences sociales des théories pour décrire et expliquer des phénomènes (Legendre, 1988). Si une théorie comprend normalement des concepts et des modèles, elle doit toujours aussi être réfutable pour la qualifier de théorie scientifique. Par conséquent, une théorie de la mesure vise à discerner le vrai du faux dans des résultats.

Bertrand (1988) affirme que l'on retrouve globalement trois théories de la mesure: la théorie classique des tests (TCT), la théorie de réponse à l'item (TRI) et la théorie de la généralisabilité (TG). Quant à Marcoulides (1999, p. 129), il classe les théories de la mesure en deux catégories: les théories dites d'échantillonnage, comprenant la TCT et la TG, et la théorie de réponse à l'item (TRI) dont le modèle émergent à l'heure actuelle est le modèle de Rasch.

Ces théories de la mesure citées précédemment permettent-elles d'inférer la capacité en mathématique des sujets à partir de leurs réponses à l'épreuve et de poser un diagnostic sur leur possibilité de réussite?

## ***2. La théorie classique des tests (TCT)***

Déjà dans la Chine antique (2200 av. J.-C.), on utilisait des tests pour mesurer la compétence des futurs fonctionnaires dans six domaines: l'écriture, l'arithmétique, la musique, le tir à l'arc, l'équitation et les rites cérémoniales (Dubois, 1970). Cependant, il a fallu le début du XX<sup>e</sup> siècle pour avoir un embryon de la théorie classique des tests. En effet, en 1904 Charles Spearman écrit deux articles sur l'analyse des données d'un test (dans McDonald, 1999, p.5). L'un de ces articles montre comment mesurer l'erreur dans les scores au test pour pouvoir inférer le score vrai, concept fondamental de la théorie classique des tests.

- Le score vrai

Le score  $x_a$  d'un sujet, a, est déduit à partir de son schéma de réponses aux items du test et il estime le trait mesuré. Un cas particulier cependant, c'est lorsque le score représente la somme des «points» attribués aux réponses exactes.

$$x_a = \sum x_{ai} \quad (1)$$

Où:  $x_{ai}$  : score observé pour la personne «a» à l'item, i;  
 $x_a$  : score observé au test pour la personne «a».

Avec la théorie classique, l'attribut latent devient le score vrai. En plus, Lord et Novick (1968, p.34) énoncent (voir équation 2) que le score observé pour une personne est égal à son score vrai plus une erreur de mesure.

$$x_a = \tau_a + e_a \quad (2)$$

Où :  $\tau_a$  : score vrai pour la personne «a» ;  
 $e_a$  : erreur pour la personne «a».

On avance le postulat que l'erreur est aléatoire seulement c'est-à-dire, que l'on peut retrouver théoriquement le score vrai si on prend l'espérance mathématique d'une série de mesures équivalentes obtenues dans les mêmes conditions puisque les erreurs s'annulent. Et le score vrai est significatif seulement si on le réfère à une échelle de mesure puisqu'on associe le nombre produit par la mesure à une quantité d'unités standard.

- L'échelle de mesure avec la théorie classique

Si on analyse globalement les résultats d'un test administré à un échantillon d'une population de sujets, on formule le modèle de mesure sous la forme d'une espérance mathématique pour la population. Avec la TCT, nous assumons que l'erreur n'est pas corrélée avec le score vrai, ce qui veut dire que l'erreur est aléatoire entre les sujets. Conséquemment, pour une population de sujets ou de mesures, l'espérance de l'erreur devient nulle,  $E(e)=0$ , et le modèle se réduit à l'équation 3.

$$E(x) = E(\tau) \quad (3)$$

Cela signifie que l'espérance des scores vrais est égale à l'espérance des scores observés et, à la limite, il n'y a pas d'erreur pour la référence de l'échelle de mesure qui se confond à la moyenne de la population,  $\mu$ .

Par contre, la variance des scores observés indique la dispersion des scores par rapport à la tendance centrale de la population et cet indice est toujours positif car il représente une espérance des écarts à la moyenne au carré comme l'indique l'équation 4.

$$\sigma_x^2 = E(x_a - \mu)^2 \quad (4)$$

En prenant la racine carrée de la variance des scores, on obtient l'écart-type,  $\sigma_x$  qui peut constituer la base de l'unité de mesure. Or, cet indice implique la plupart du temps une part de bruit ou d'erreur lorsqu'on fait une interprétation normative des scores.

- La fidélité de la mesure avec la théorie classique

La partie de la variance des scores observés qui n'est pas expliquée par la variance des scores vrais est due au hasard ou à la variance des erreurs. C'est ainsi que Lord et Novick (1968, p.57) définissent le coefficient de fidélité,  $\rho_{xx'}$  comme étant le rapport entre la variance des scores vrais et la variance des scores observés.

$$\rho_{xx'} = \sigma_\tau^2 / \sigma_x^2 \quad (5)$$

En d'autres mots, la fidélité représente la proportion de la variance des scores observés qui est expliquée par la variance des scores vrais. Mais il est possible d'exprimer le coefficient de fidélité en fonction de la variance des erreurs.

$$\rho_{xx'} = 1 - (\sigma_e^2 / \sigma_x^2) \quad (6)$$

Au maximum, ce coefficient vaudrait 1, ce qui signifie que la variance des erreurs est nulle. Au minimum, il vaut 0 et alors la variance des erreurs est égale

à la variance des scores observés ce qui signifie que seulement le hasard, c'est-à-dire la variance des erreurs, explique la variance des scores observés.

À partir de l'équation 6, nous pouvons déduire l'erreur-type de mesure,  $\sigma_e$ , qui nous permet d'établir un intervalle de confiance où doit se situer le score vrai. Ainsi, plus le coefficient de fidélité tend vers 1, plus petite est cet intervalle de confiance et plus grande est la fonction d'information du test. D'ailleurs, Laurencelle (1997, p.37) mentionne que la fonction d'information d'un test est l'inverse de la racine carrée de l'erreur-type de mesure.

$$I(\theta) = \frac{1}{\sqrt{\sigma_e}} \quad (7)$$

En manipulant l'équation 6, on peut déduire l'estimé de l'erreur-type de mesure,  $s_e$ , à partir d'une estimation d'un coefficient de fidélité,  $r_{xx'}$ , et d'une estimation de l'écart-type des scores observés,  $s_x$  (voir éq. 8).

$$s_e = s_x \sqrt{1 - r_{xx'}} \quad (8)$$

Il existe plusieurs façons d'estimer le coefficient de fidélité,  $r_{xx'}$ . Pour De Landsheere (1979, p.126):

*Les principaux indices de fidélité sont : le coefficient de stabilité (corrélation test-retest), le coefficient d'homogénéité (notamment à l'aide des formules 20 et 21 de Kuder-Richardson), le coefficient d'équivalence (corrélation entre les scores obtenus à deux formes parallèles du test).*

De plus, les formules de Kuder-Richardson constituent des cas particuliers de l'équation 9 qui décrit le coefficient alpha,  $\alpha$ , de Cronbach (Lord et Novick, 1968, p.89). Cette formule constitue la façon la plus usuelle d'estimer le coefficient de fidélité en tenant compte des réponses aux items.

$$\rho_{xx'} > \alpha = n / (n - 1) [1 - (\sum \sigma^2(y_i) / \sigma_x^2)] \quad (9)$$

Où :  $\alpha$ : coefficient estimé de la fidélité ;

$n$  : nombre d'items dans le test ;

$y_i$  : score à l'item  $i$  pour un sujet ;

$\sigma^2(y_i)$  : variance des scores à l'item  $i$  pour la population de sujets.

Toutefois, ce coefficient  $\alpha$  sous-estime la fidélité réelle et nous pouvons interpréter cette équation en concluant que plus la variance des scores observés au test,  $\sigma_x^2$ , est expliquée par la somme des variances des items, moins grande est la fidélité de la mesure. Donc, la variance des scores aux items doit être la plus faible possible.

Mais il existe d'autres coefficients que alpha. Par exemple, McDonald (1999, p.89) présente le coefficient de fidélité  $\omega$  de Guttman-Cronbach qui est plus général que le coefficient  $\alpha$ . Ce coefficient  $\omega$  donne un meilleur estimé de la fidélité que le coefficient  $\alpha$  du fait qu'il tient compte en plus de la covariance entre les items pour mesurer un attribut commun.

- La validité pronostique de la mesure avec la théorie classique

Si on administre l'épreuve de capacité en mathématique dans le but de sélectionner les sujets les plus forts pour l'admission, alors les axiomes de la théorie classique des tests prennent toute leur pertinence. En effet, l'estimé de la fidélité nous permet de connaître la marge d'erreur possible lors de la prise de décision. Or, on administre plutôt cette épreuve dans le but de classer tous les sujets en fonction du critère de réussite ; aussi, la marge d'erreur lors de la décision dépend avant tout de la validité pronostique pour l'instrument.

Normalement, on a recours à un modèle de régression statistique linéaire pour déterminer la validité pronostique de l'épreuve. Avec un tel modèle, on prédit le rendement moyen d'un sujet à la fin de ses études à partir du score observé à l'épreuve; plus particulièrement, la variable prédite  $Y$  se confond au rendement alors que la variable pronostique  $X$  s'assimile au score de l'épreuve. Pour ce, le

coefficient de corrélation permet de tracer une droite de régression qui donne la valeur minimum pour les erreurs de prédiction au carré. De plus, le carré de ce coefficient, appelé coefficient de détermination, donne la proportion de variance de la variable prédite Y qui est expliquée par la variable pronostique X (Allen et Yen, 1979).

Toutefois, Violato *et al.* (1992, p.153) affirment qu'il est rare que ce coefficient de corrélation dépasse 0,6. Cela signifie qu'au plus 36% de la variance à la variable prédite Y peut être expliquée par la variance à la variable pronostique X, et son corollaire, que 64% n'est pas expliquée. Dans le même sens, à partir d'études de validité auprès de milliers d'étudiants dans les collèges américains, Zwick (2002, p.90) mentionne que la corrélation entre le test d'admission SAT (*Scholastic Assessment Test*) et la note dans la première année du collège se situe entre 0,3 et 0,4.

De si faibles indices pour la validité pronostique ne nous permettent pas de minimiser l'incertitude lors de la prise de décision parce qu'ici une décision va prendre le sens d'un diagnostic qui requiert une interprétation critériée des scores alors qu'il est plus pertinent d'utiliser la théorie classique des tests lorsqu'on désire faire une interprétation normative des scores. Donc, nous ne croyons pas que la TCT puisse apporter une solution adéquate au problème de la prédiction de la réussite.

### ***3. La théorie de réponse à l'item (TRI)***

La théorie classique des tests a été élaborée pour répondre à des besoins psychométriques. Aussi, la théorie de réponse à l'item vient combler un vide dans la mesure en éducation du fait qu'elle permet d'inférer un niveau d'habileté du trait mesuré pour un sujet à partir des courbes de probabilité de réponse de plusieurs items.

- L'échelle hiérarchique de Guttman

Si avec la théorie classique, le schéma de réponses d'un sujet n'a qu'une utilité limitée, celui-ci prend une grande importance avec la théorie de réponse à l'item. Par exemple, l'échelle hiérarchique de Guttman (De Landsheere, 1979) permet de discriminer le niveau d'habileté d'un sujet en comparant son schéma de réponses à un scalogramme représentant l'ensemble des schémas de réponses plausibles ou considérés comme vrais. Blais et Laurier (1997, p.69) en énoncent les conditions pour obtenir une telle échelle hiérarchique:

*«... lorsque les items sont ordonnés selon leur degré de difficulté, les scores provenant d'un test unidimensionnel devraient permettre d'obtenir une échelle de Guttman.»*

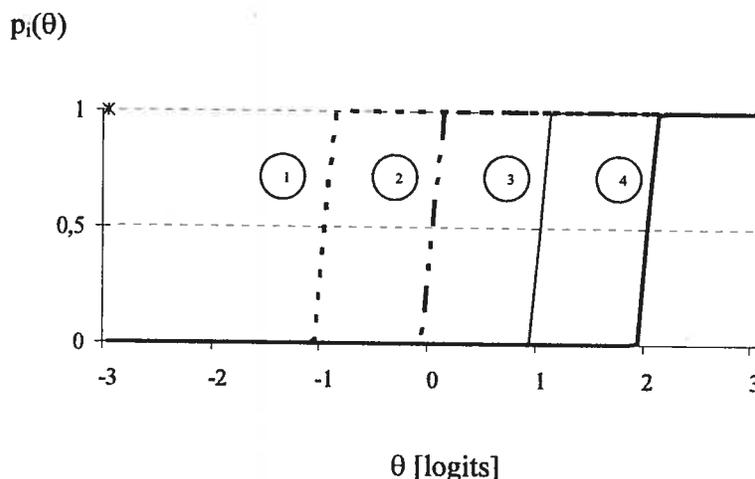
Un exemple illustrant un scalogramme pour un test à quatre items est donné au tableau II.

**Tableau II: Scalogramme d'un test à quatre items**

Sujet	Item #1	Item #2	Item #3	Item #4	Niveau $\theta$
A	0	0	0	0	-2
B	1	0	0	0	-1
C	1	1	0	0	0
D	1	1	1	0	1
E	1	1	1	1	2

Avec un tel test, il doit y avoir une progression dans le degré de difficulté des items. L'item #1 est le plus facile et l'item #4, le plus difficile puisque seul le sujet E réussit l'item #4 alors qu'il y en a quatre qui réussissent l'item #1. D'une façon générale pour un test à n items, on a n+1 schémas de réponses plausibles parmi  $2^n$  schémas possibles. Ainsi, le score au test est la somme des scores aux items et infère autant de niveaux d'habiletés  $\theta$  mais il faut que le trait mesuré soit unidimensionnel, c'est-à-dire que l'ensemble des items ne mesure qu'une seule habileté.

Et le nombre d'items détermine la précision de la mesure comme l'illustre le graphique de la figure 3.



**Figure 3:** Ordonnancement des items

Ce graphique représente la probabilité de réponse exacte d'un sujet aux items d'un test en fonction de son habileté exprimée en logits. En effet, la probabilité de réussite de l'item #1 devient certaine pour le sujet qui a une habileté supérieure à -1 (courbe pointillée). De même la probabilité de réussite de l'item #2 devient certaine lorsque le sujet a une habileté supérieure à 0 (courbe semi-pointillée). Ainsi, on peut relier le score au niveau d'habileté du sujet mais avec une précision limitée puisqu'on ne peut inférer une fraction de logits ; en effet, les scores possibles sont : -1, 0, +1 et +2 logits. En plus, ce modèle idéalise une situation de mesure car il est presque impossible que la probabilité de réussite d'un item puisse passer de 0 à 1 pour une très faible différence d'habileté. Un autre modèle, celui de Rasch, est plus collé à la réalité.

- Le modèle de Rasch

Avec la théorie de réponse à l'item (TRI), nous parlons plutôt de trait latent, c'est-à-dire d'un attribut qui n'est pas directement observable. Selon Embretson (1999, p.9), le niveau du trait est estimé par l'interrogation suivante : *Étant*

*données les propriétés des items, quel est le niveau de trait le plus vraisemblable pour expliquer les réponses de la personne?*

Pour obtenir une réponse à cette question, il faut au préalable caractériser chaque item par un modèle. Aussi, Embretson (1999, p.9) suggère d'utiliser le modèle de Rasch qui épouse la forme d'une fonction logistique (voir éq. 10) entre le niveau d'habileté  $\theta$  et la probabilité de réussite à l'item  $i$ .

$$p_i(\theta) = \frac{1}{1 + e^{-a(\theta - b_i)}} \quad (10)$$

Où :  $p_i(\theta)$  : probabilité de réussite à l'item  $i$  ;

$e$  : nombre égal à 2,718... ;

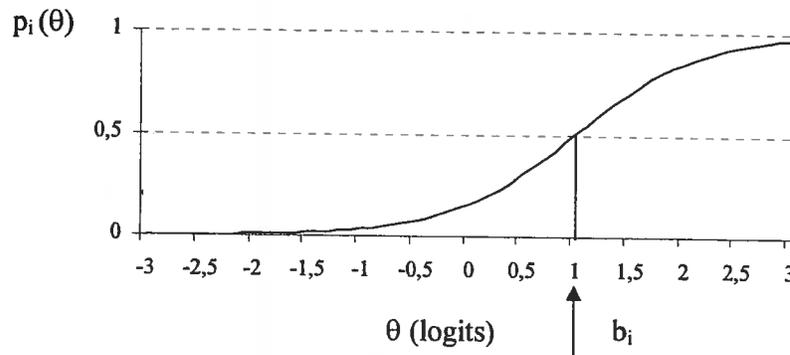
$a$  : constante quelconque qui représente un indice de discrimination uniforme pour tous les items du test;

$b_i$  : indice de difficulté de l'item « $i$ »;

$\theta$  : habileté du sujet.

Dans cette équation, les paramètres ont une même échelle de mesure ; l'unité de mesure de  $b$  et  $\theta$  est le logit nous venant de la contraction de l'expression anglaise «LOGistic unIT». Aussi, la grandeur  $\theta$  indique la valeur de l'attribut latent en logits, cet attribut étant le niveau d'habileté inféré à un sujet. Pour De Landsheere (1979, p.154), la méthode des intervalles égaux, aussi appelée méthode de Thurstone, permet d'établir une échelle d'intervalle en comparant les scores de plusieurs juges à des items différentiels du type «d'accord, pas d'accord». Or, Andrich (1978) démontre que le modèle de Rasch est équivalent à cette méthode. Conséquemment, nous pouvons considérer que l'échelle pour  $\theta$  en est une d'intervalle, la référence à zéro étant arbitraire et correspondant au niveau d'habileté moyen mesuré par le test. De plus, ce modèle se caractérise par une fonction logistique qui comprend deux asymptotes horizontales, l'une à 0 et l'autre à 1. Le graphique de la figure 4 illustre cette fonction logistique où l'on peut voir la probabilité de réussite à un item versus le niveau d'habileté  $\theta$  mesuré

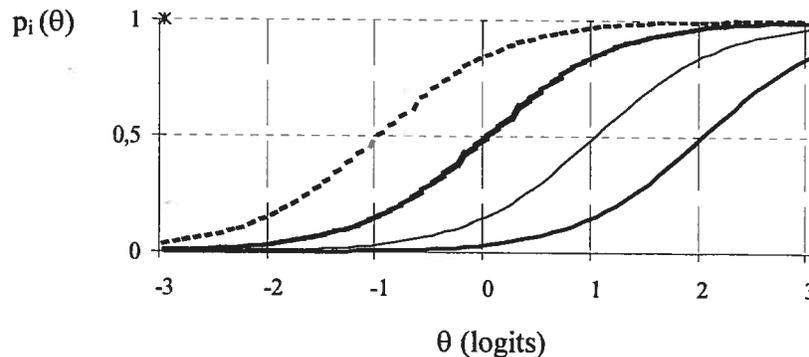
sur une échelle de -3 à +3 logits ; cependant, cette échelle varie normalement de  $-\infty$  à  $+\infty$ .



**Figure 4: Le modèle de Rasch**

Ainsi, pour un item dont l'indice de difficulté est égal à l'habileté du sujet, nous retrouverons un argument nul pour l'exponentielle et c'est ce qui donne une probabilité de réussite de 0,5. Dans le graphique de la figure 4, cet indice doit être légèrement supérieur à +1 logit. La probabilité  $p_i(\theta)$  est la probabilité que le sujet réponde correctement à l'item  $i$ ; néanmoins, on suppose que la correction pour cet item est dichotomique (0 ou 1). Cette probabilité augmente lorsque l'habileté devient supérieure à l'indice de difficulté,  $p_i(\theta) > 0,5$  si  $\theta > b_i$ , et diminue dans le cas contraire,  $p_i(\theta) < 0,5$  si  $\theta < b_i$ .

Afin d'effectuer une mesure sur une échelle d'intervalle, Wright (1999, p.97) propose de construire un test en ordonnant par niveaux de difficulté les caractéristiques des items, le graphique de la figure 5 en illustre une application avec quatre items.



**Figure 5:** Mise en ordre des items selon le modèle de Rasch

Ce graphique (figure 5) représente un test à quatre items avec chacun un indice de difficulté différent. Une inspection rapide nous permet de voir que ces items ont respectivement des indices de difficulté  $b_i$  de -1, 0, +1 et +2 logits. En construisant un test avec des items ayant des niveaux de difficulté différents, nous devons pouvoir inférer une valeur du trait latent correspondant à un niveau d'habileté  $\theta$  pour le sujet.

Cependant, il est possible qu'un sujet ayant un niveau d'habileté  $\theta$  donné puisse répondre correctement à un item plus difficile et incorrectement à un item plus facile. Par exemple selon le graphique de la figure 5, un sujet avec une habileté de +1 logit, aurait une probabilité d'environ 0,95 de répondre à l'item 1, d'environ 0,8 de répondre à l'item 2, de 0,5 de répondre à l'item 3 et d'environ 0,2 de répondre à l'item 4. Or, il est loin d'être certain que le modèle idéal de schémas de réponses plausibles illustré par le scalogramme du tableau II puisse être respecté.

- Les modèles à deux et trois paramètres

Afin de mieux modéliser les données provenant de l'administration d'un test, on peut recourir à des items calibrés avec un modèle à deux paramètres. En plus du paramètre indiquant le niveau de difficulté, nous avons un paramètre  $a_i$  pour indiquer la discrimination. En analysant l'équation 11, nous pouvons voir que l'indice de discrimination,  $a_i$ , vient influencer la pente de la courbe

caractéristique lorsque l'habileté  $\theta$  du sujet est égale à l'indice de difficulté  $b_i$ . De plus, le modèle à trois paramètres vient ajouter l'indice pour le pseudo-hasard,  $c_i$ , il entraîne alors un décalage vers le haut de l'asymptote horizontale qui était à 0.

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta - b_i)}} \quad (11)$$

Où :  $a_i$  : indice de discrimination;

$c_i$  : indice pour le pseudo-hasard.

Ainsi, pour un item à réponse choisie avec cinq choix possibles, nous avons un indice  $c_i$  de 0,2 car le candidat a une chance sur cinq de réussir cet item par le hasard. Toutefois, plus l'indice de discrimination  $a_i$  est élevé, plus nous avons la certitude que l'examiné réponde correctement à l'item si son habileté est supérieure à l'indice de difficulté  $b_i$  et vice-versa. À l'inverse, un item avec un indice de discrimination trop faible doit être rejeté.

- La fonction d'information avec la théorie de réponse à l'item

Avec la TRI, nous ne pouvons pas parler directement d'erreur de mesure. En effet, il faut plutôt parler de fonction d'information et se référer au schéma de la communication de la figure 2 puisqu'un test idéal doit reproduire fidèlement les messages de la source d'information.

Or, pour maintenir une grande précision, il faut choisir les items avec des indices de discrimination et de difficulté qui maximisent la fonction d'information du test. En ce sens, Hashway (1998, p.58) propose une fonction d'information pour un item caractérisé par un modèle à deux paramètres (voir éq. 12).

$$I(u_i, \theta) = a_i^2 p_i(\theta) [1 - p_i(\theta)] \quad (12)$$

La quantité d'information,  $I(u_i, \theta)$ , est fonction de la réponse à l'item,  $u_i$ , et de l'habileté  $\theta$ . Un item modélisé avec deux paramètres contribue pour le maximum d'information lorsque  $p_i(\theta) = 0,5$ . À ce moment celui-ci discrimine à

son maximum et le trait mesuré,  $\theta$ , est égal à l'indice de difficulté,  $b_i$ . La fonction qui maximise l'information pour le test est donc égale à la somme des fonctions d'information pour chaque item (voir éq. 13).

$$I(x^*, \theta) = \sum I(u_i, \theta) \quad (13)$$

Le score idéal,  $x_a$ , peut être obtenu en choisissant pour chaque item un poids optimal qui est égal à l'indice de discrimination  $a_i$ .

Si on augmente jusqu'à l'infini le nombre d'items dans le test pour tendre vers une population alors la distribution théorique, entre le nombre d'items réussis et le niveau d'habileté,  $\theta$ , épousera la forme d'une courbe normale (Allen et Yen, 1979; p.269). Avec cette situation idéale, l'espérance mathématique du trait estimé pour un sujet est la valeur vraie du trait latent,  $\theta$ . Et l'erreur-type est inversement proportionnelle à la racine carrée de la fonction d'information et donc, du nombre d'items.

$$\sigma_e = 1 / \sqrt{I(x^*, \theta)} \quad (14)$$

En conclusion, nous pouvons affirmer que plus la fonction d'information des items est élevée, plus on se rapproche de l'échelle idéale de Guttman.

Avec l'avènement des ordinateurs, il s'est développé une nouvelle forme de mesure, du nom de test adaptatif, où les items sont présentés au sujet en fonction de son habileté. Avec un tel test adaptatif, les réponses du sujet à une première série d'items permettent d'estimer grossièrement son habileté. L'item suivant cette série est sélectionné en fonction de son habileté estimée ; encore une fois, une estimation de l'habileté est effectuée à partir de la réponse et ce processus est répété tant qu'il n'y aura plus de variations dans l'estimation de l'habileté.

- La validité pronostique de la mesure avec la théorie de réponse à l'item

Avec la théorie de réponse à l'item, le test mesure une seule habileté. Si l'on se réfère aux types d'interprétation des scores de Brown (1980, p.84), cette théorie

favorise une interprétation en référence à un contenu, c'est-à-dire que la mesure est indépendante de la population des sujets testés et ne dépend que du contenu disciplinaire. Ainsi, la réussite d'un item peut devenir un pronostic en autant que cet item discrimine bien par rapport à l'habileté mesurée.

Cette habileté ne doit pas comporter plus d'une dimension. Toutefois, certains auteurs remettent en doute le concept d'unidimensionnalité qu'on doit assumer lorsqu'on a recours à la théorie de réponse à l'item (TRI). En effet, pour Mislevy (1993), la vision des habiletés humaines, implicite dans les théories classique et de réponse à l'item, est incompatible avec celle de la psychologie cognitive et il ne faut pas négliger des dimensions cachées mesurées par le test. Carver (1974) affirme que les instruments en général comprennent une dimension psychométrique et une dimension éducatrice ; la dimension psychométrique permet une interprétation des scores en référence à une norme alors que la dimension éducatrice favorise l'interprétation en référence à un critère. Auger et Séguin (1992) ont appliqué la théorie de réponse à l'item dans le but de faire une interprétation critériée; l'expérimentation se faisait dans le cadre d'une évaluation sommative où la fonction est la certification et le test était également adaptatif.

La théorie de réponse à l'item trouve sa pertinence dans une situation où l'on fait une interprétation des scores en référence à un contenu; on rencontre une telle situation lorsque l'évaluation a pour fonction la régulation ou la certification car alors on ne désire mesurer qu'une seule dimension. Par contre, pour une évaluation avec une fonction d'orientation, il faut pouvoir prédire un critère. Or, le pronostic en éducation constitue selon nous un processus trop complexe pour se restreindre qu'à un seul trait unidimensionnel.

L'épreuve diagnostique en mathématique, faisant l'objet de la présente recherche, est de type papier-crayon et comporte des domaines qui ne peuvent être regroupés dans une seule dimension ou habileté dans le sens donné par la théorie de la réponse à l'item (TRI). Rappelons également que la fonction

d'évaluation en est une d'orientation et que par conséquent, les conclusions de Auger et Séguin (1992) ne peuvent être utilisées. Un exemple d'habileté unidimensionnelle avec cette épreuve peut être le calcul différentiel et intégral si on ordonne les items par degré de difficulté en progressant à partir des items sur les fonctions élémentaires et en poursuivant avec ceux sur le calcul différentiel et le calcul intégral. Or, le trait mesuré par cet instrument est la capacité en mathématique qui comprend la dimension de l'habileté pré-citée mais aussi des dimensions dans d'autres sphères telles la géométrie, la trigonométrie, les matrices et les vecteurs. Pour toutes ces raisons, nous ne croyons pas à la pertinence de recourir à la théorie de réponse à l'item pour étudier la validité pronostique de cette épreuve de mathématique.

#### ***4. La théorie de la généralisabilité***

Afin de minimiser l'influence du hasard, on doit tenir compte de plusieurs facettes de la mesure; c'est dans ce sens que Cronbach et *al*, (1972) proposèrent la théorie de la généralisabilité (TG). Pour un test, on retrouve au minimum deux facettes de la mesure : la facette «item» et la facette «personne». En pratique, il existe toujours des facettes, cachées ou non, qui agissent comme sources d'erreurs. Selon Blais et Fournier (1997, p.78) :

*«La théorie de la généralisabilité a mis à profit les développements importants de l'analyse de la variance en statistique pour proposer une vision différente de l'erreur de mesure.»*

Bref, cette théorie vise à expliquer la variance des scores observés,  $\sigma_x^2$ , à partir des variances pour chacune des facettes.

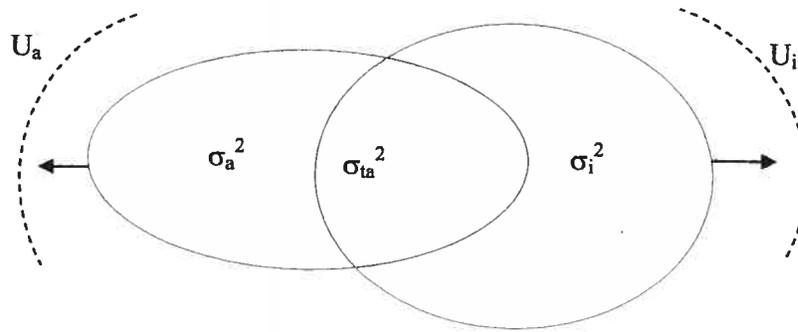
- **Le plan de mesure**

Selon Bertrand (1988, p.7) : «Cette théorie (.) permet de tenir compte de façon très explicite des multiples sources d'erreur de mesure.» Parmi toutes les facettes, il y aura une facette de différenciation, c'est-à-dire une facette correspondant au trait mesuré. Dans ce sens, les autres facettes deviennent des

sources d'erreur. Le plan de mesure vient spécifier quelles sont les facettes en jeu. Pour Tourneur (1989-1, 31) :

*«Les possibilités de même type sont ainsi regroupées en facettes dont les plus courantes sont : les élèves, les questions, les traitements ou les méthodes, les objectifs, les évaluateurs, les moments et les phases de l'apprentissage.»*

Cronbach et al, (1972) ont suggéré d'utiliser un diagramme de Venn (voir figure 6) pour représenter les variances en jeu dans un plan de mesure.



**Figure 6 : Le plan de mesure a x i**

Le plan de mesure «a x i» signifie que l'on croise les personnes «a» avec les items «i». Les arcs en pointillés représentent l'univers des sujets,  $U_a$ , et l'univers d'items,  $U_i$ . D'une part, l'ensemble des examinés, représenté par l'ovale de gauche, est inclus dans un univers de personnes, aussi appelé population. D'autre part, l'ensemble des items, représenté par l'ovale de droite, est inclus dans un univers d'items. Les flèches montrent le sens de la généralisation.

- La facette de différenciation

Cardinet (1986) affirme qu'il faut maximiser la variance d'une facette selon ce que l'on veut mesurer. La variance entre les personnes,  $\sigma_a^2$ , représente la facette de différenciation lorsqu'on veut faire une interprétation en référence à une norme. Avec la TCT, nous disons qu'il s'agit de la variance du score vrai. Avec

la TG, l'expression «score vrai» est plutôt remplacée par «score de l'univers» (Brennan, 1980 ; p.189). Le coefficient d'assurance (Cardinet, 1986 ; p.135) indique la fidélité et dans ce cas le rapport entre la variance des scores de l'univers,  $\sigma_a^2$ , et la variance des scores observés,  $\sigma_x^2$ .

L'avantage de la TG sur la TCT est que l'on y considère en plus la variance des items,  $\sigma_i^2$ , dans le plan de mesure. Et avec la TRI on considère cette variance aux items comme facette de différenciation pour déterminer l'habileté d'un sujet. L'interprétation des scores est alors en référence à un contenu.

Lorsqu'on effectue une analyse de variance (ANOVA) avec deux facettes, il apparaît une variance qui est due à l'interaction des deux facettes et qui produit un effet sur le score observé (Glass et Hopkins, 1984 ; p.402). La variance d'interaction,  $\sigma_{ia}^2$ , est considérée par Brennan (1980, p.197) comme une source d'erreur. D'ailleurs, une valeur importante pour cette variance due à l'interaction signifie que les items ont été construits à la fois pour différencier les individus et les contenus.

- *Le plan de mesure pour l'interprétation en référence à un critère*

Pour Cardinet (1986, p. 97), il faut tenir compte des facettes «traitement» et «moments» lorsque l'évaluation est prédictive. Or, nous ne pouvons considérer ces facettes dans le cadre de notre recherche puisque les scores observés à l'épreuve de capacité se situent à un moment qui précède le traitement que constitue l'enseignement d'appoint.

De plus, il faut effectuer l'analyse de variance en tenant compte d'un score de césure au test, cela constituant la facette de différenciation. Or, Brennan (1980, p.214) démontre que l'indice de fidélité pour une telle application est fonction du nombre d'items et de la distance du score de césure par rapport à la moyenne. Il n'y a donc pas de garantie d'une fidélité élevée.

Cardinet (1986, p.120) utilise la théorie de la généralisabilité pour une évaluation sommative dans un but de certification. Or, il n'y a pas, selon nos sources, d'application de cette théorie pour une mesure pronostique dans un but d'orientation. Conséquemment, la théorie de la généralisabilité ne nous est pas plus utile que les théories précédentes pour solutionner le problème du pronostic sur la réussite.

##### *5. La mesure critériée et la théorie des ensembles*

Nous avons convenu dans le chapitre précédent que la mesure au début du processus d'enseignement joue une fonction d'orientation et qu'elle doit être en référence à un critère pour poser un diagnostic. Or, il y a eu plusieurs tentatives pour expliquer le concept de mesure critériée et la plupart d'entre elles se caractérisent par une décision dichotomique prise à partir d'un score brut. De plus, cette décision dichotomique s'assimile à une opération de classement dans un ensemble et la théorie des ensembles s'attarde justement au classement des éléments. Nous pouvons donc supposer que cette théorie mathématique est utilisée implicitement lorsqu'on fait une interprétation critériée de la mesure, les éléments étant confondus avec les sujets.

Rappelons-le, l'objectif de la recherche est de prédire la réussite d'un sujet afin de l'orienter dans l'une ou l'autre des voies d'enseignement suivantes: l'enseignement régulier ou l'enseignement d'appoint. Il s'agit d'un objectif purement éducatif et les scores à l'épreuve diagnostique doivent être interprétés en référence à un critère car en effet, le diagnostic doit supporter ici une décision de classement. Pour Dassa et Laurier (2003, p.121): «Le classement sert habituellement à situer l'élève par rapport au continuum que représentent une séquence de cours ou un programme.» Situer l'élève signifie: poser un diagnostic. Lohman et Ippel (1993) affirment que le but d'une théorie de la mesure doit être tout d'abord de déterminer une structure de relations entre observations et ensuite d'élaborer une procédure de classement de ces observations. Cela laisse entendre que le diagnostic peut être complexe.

Cependant, la réalité d'une situation de classement ne requiert bien souvent qu'un diagnostic simple (Dassa et Laurier, 2003 ; p.122) :

*«Toutefois, dans la réalité, les milieux scolaires ne permettent guère de raffiner le classement au point de tenir compte d'autres aspects que le niveau général, de sorte qu'un diagnostic détaillé est relativement peu utile dans une décision de classement.»*

Lorsque la mesure s'effectue durant le processus d'enseignement et si on veut supporter une évaluation ayant la fonction de régulation, il est plus approprié de se référer à un diagnostic complexe établi à partir des scores à des sous-tests. Cependant, dans la présente recherche, nous devons nous assurer que la mesure supporte une évaluation ayant la fonction d'orientation ou de classement. Aussi, nous épousons la position de Dassa et Laurier (2003) et le diagnostic sera établi à partir du score total à l'épreuve afin de maximiser sa valeur pronostique.

- L'échelle d'appréciation

Avec la théorie des ensembles, l'échelle de mesure devient une échelle d'appréciation qui prend la forme d'un classement à deux échelons: échec-réussite. Or, lorsqu'on effectue une mesure référée à un critère, on signifie en principe que l'on désire mesurer le produit de l'apprentissage d'un individu et le classer selon ces deux mêmes échelons. Alors, l'objectif d'apprentissage décrit l'attribut «réussite» pour un univers de sujets un peu comme la moyenne et l'écart-type décrivent les caractéristiques d'une population de sujets pour une mesure référée à une norme.

Une simple appréciation échec-réussite contient peu d'information mais toutefois la mesure référée à un critère a surtout pris son essor avec la théorie de l'apprentissage du behaviorisme qui propose un système de classification. Cette théorie, qui a pris son essor à la fin du XIXe et au début du XXe siècle (St-Yves, 1982 ; p.15), se centre sur le comportement. Ainsi, les comportements attendus deviennent des objectifs d'apprentissage qu'il faut classer. Dans ce sens, Bloom *et al*, (1982, p.6) sont les premiers à élaborer un système de classification aussi appelé taxonomie:

*«L'idée de ce système de classification prit naissance à Boston en 1948 lors d'une simple réunion d'examineurs de collège.»*

La base de ce système de classification repose sur les objectifs pédagogiques. Ainsi, le terme «objectif» devient désormais un critère pour la mesure des apprentissages; de là, est né en 1962 (Berk, 1980 ; p.3) le concept de mesure référée à un critère plus communément appelé «mesure critériée» ou «critérielle». Bref, la mesure critériée vise à positionner le sujet par rapport à un objectif pédagogique. Depuis le milieu des années 60, il y a eu une tendance croissante dans l'utilisation de ce type de mesure pour diagnostiquer la maîtrise des apprentissages (Hashway, 1998 ; p.111).

Néanmoins, le fait d'interpréter les scores en référence à un critère va exiger certaines particularités pour l'instrument de mesure, En effet, selon Popham (1981), les items d'un test critérié sont caractérisés par leur congruence avec un objectif pédagogique. Et pour Huitt (1996), chaque habileté associée à un objectif doit être mesurée par au moins quatre items pour obtenir un échantillon adéquat de la performance du sujet et minimiser l'effet du hasard.

Le score brut au test est comparé à un score de césure pour donner une décision de diagnostic positif ( $\mu_X=1$ ) ou de diagnostic négatif ( $\mu_X=0$ ). Ainsi, nous pouvons modéliser une situation de mesure critériée par l'algorithme 15.

$$\begin{aligned} \mu_X(a) &= 0 & \text{si} & & x_a < x_0 & & (15) \\ \mu_X(a) &= 1 & \text{si} & & x_a \geq x_0 & & \end{aligned}$$

Où:  $\mu_X(a)$ : variable diagnostique pour le sujet «a»;

$x_a$ : score brut du sujet «a»;

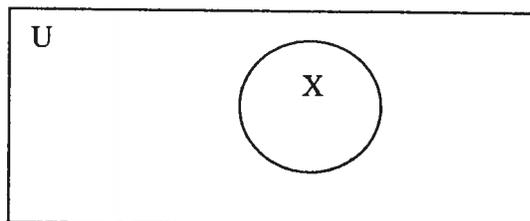
$x_0$ : score de césure.

La variable diagnostique,  $\mu_X(a)$ , constitue une variable de Bernouilli car selon Baillargeon (1989, p.142):

*«Une variable aléatoire discrète qui prend les valeurs 1 et 0 avec des probabilités  $p$  et  $1-p$  est appelée variable de Bernouilli.»*

- La situation de classement

Si on désire représenter une situation de classement, la théorie des ensembles convient davantage. Le diagramme de Venn de la figure 7 illustre le modèle de l'algorithme 15 pour un groupe de sujets.



**Figure 7: Le diagramme de Venn pour une mesure critériée**

L'univers  $U$  représente le groupe de sujets testés et constitue de ce fait l'espace d'échantillonnage alors que l'ensemble  $X$  se constitue de tous ceux ayant eu un diagnostic positif. Plus précisément, chaque sujet est un élément de l'univers et se positionne soit à l'intérieur de l'ensemble  $X$  ou à l'extérieur.

Ainsi, à l'ensemble  $X$  on associe un attribut. Pour ce qui concerne une situation de classement avec une mesure critériée, l'attribut représente un libellé décrivant la maîtrise d'un apprentissage; dans le cadre de la présente recherche, cet attribut est «capacité en mathématique». Donc, le diagnostic,  $\mu_X(a)$ , indique s'il est vrai (=1) ou faux (=0) que le sujet «a» possède cet attribut; cela est vrai si cette même personne «a» appartient à l'ensemble  $X$  et faux dans le cas contraire.

- L'algèbre de Boole

Étant donné que la mesure ne se fait plus sur une échelle d'intervalle, on ne peut plus utiliser les propriétés d'additivité pour les nombres. L'algèbre de Boole<sup>4</sup> vient préciser les règles s'appliquant à la théorie des ensembles (Ralston et Reilly, 1983; p.179) et indirectement à la logique.

---

<sup>4</sup> Ainsi dénommé en l'honneur du mathématicien anglais Georges Boole (1815-1864).

### Les définitions

- Élément:

L'élément est la plus petite entité à considérer. L'appartenance d'un élément à un ensemble est symbolisé par la lettre grecque epsilon,  $\in$ , et la non appartenance par le symbole  $\notin$ . Un sujet est un élément; ainsi, un sujet compétent peut être noté  $a \in X$  et un sujet incompétent par  $a \notin X$ . Nous pouvons aussi associé l'élément à un cas qui représente pour nous un sujet.

- Ensemble:

L'ensemble est un groupe d'éléments ayant un attribut commun; on le symbolise par une lettre majuscule (ex:  $X$ ). On peut définir un ensemble en mettant entre accolades  $\{ \}$  les éléments qui le composent (ex:  $\{a, b, c, d, e\}$ ). Selon Ragin (2000, p.78), les attributs constituent un «espace de propriétés» et chaque combinaison de ceux-ci est à une localisation spécifique de cet espace.

- Famille d'ensembles:

La famille d'ensembles est un groupe d'ensembles ayant un attribut commun; on la symbolise par une lettre majuscule en italique (ex:  $X$ ). Alors on distingue les ensembles d'une famille d'ensembles par un indice. Par exemple, la famille d'ensembles  $I$  regroupera les ensembles des réponses des sujets à chaque item du test tels  $I_1, I_2, \dots, I_k$ ; les réponses à chaque item de l'épreuve constituant un ensemble indicé et les réponses à tous les items formant une famille d'ensembles.

- Univers:

L'univers est un espace-temps d'échantillonnage considéré pour une situation de classement et désigné par la lettre  $U$ . Dans le cadre de cette recherche, l'univers est constitué d'une cohorte de sujets. Aussi chacun de ceux-ci peut être classé dans des ensembles caractérisés par des attributs.

- Complément:

Le complément est une opération qui ne considère que les éléments de l'univers  $U$  qui n'appartiennent pas à l'ensemble considéré. Si  $U \equiv \{a, b, c, d, e\}$  et  $X \equiv \{b, e\}$  alors le complément de l'ensemble  $X$  que nous notons par  $X'$  est égal par définition à  $\{a, c, d\}$ . Du point de vue logique, cela correspond à nier une proposition. Le complément de l'attribut «capacité» est «incapacité» et le complément de l'attribut «réussite» est «échec».

- Intersection:

L'intersection est une opération qui ne considère que les éléments communs à deux ensembles. Si  $X \equiv \{a, b, c\}$  et  $Y \equiv \{b, c, d\}$  alors  $X \cap Y \equiv \{b, c\}$ . Le symbole  $\equiv$  signifie égal par définition tandis que  $\cap$  représente l'intersection. En logique, cela signifie que l'on produit une partie de discours en insérant une conjonction «et» entre deux propositions. À chaque cas étant élément de  $X \cap Y$ , nous pouvons leur rattacher l'attribut «a la capacité et réussit».

- Union:

L'union est une opération qui considère tous les éléments appartenant à deux ensembles. Si  $X \equiv \{a, b, c\}$  et  $Y \equiv \{b, c, d\}$  alors  $X \cup Y \equiv \{a, b, c, d\}$ . Le symbole  $\cup$  représente l'union. En logique, une partie de discours est produite en mettant une conjonction «ou» entre deux propositions.

- Inclusion:

L'inclusion est une opération qui détermine si tous les éléments d'un ensemble sont compris dans un autre. Le résultat donne vrai (=1) ou faux (=0). Si  $X \equiv \{a, b, c\}$  et  $Y \equiv \{b, c, d\}$  alors  $X \subseteq Y$  est faux puisque l'élément «a» appartient à  $X$  et non à  $Y$ . Le symbole  $\subseteq$  représente l'inclusion. En logique, cela se traduit par une relation d'implication entre une proposition conditionnelle et une proposition subordonnée pour donner une partie de discours de la forme: «Si la proposition conditionnelle était vraie alors la proposition subordonnée serait vraie.». Plus précisément pour

notre recherche, nous aurions : «Si le sujet avait la capacité nécessaire en mathématique alors il réussirait.»

### Les partitions

Une partition est une zone de l'univers  $U$  qui ne peut être décomposée. Pour un univers à  $k$  ensembles, il y a  $2^k$  partitions. Chacune d'entre elles est égale par définition à l'intersection des ensembles ou de leur complément qui composent l'univers ; par exemple pour un univers composé de deux ensembles  $X$  et  $Y$ , nous aurions quatre partitions:  $X' \cap Y'$  ;  $X' \cap Y$  ;  $X \cap Y'$  et  $X \cap Y$ . Toutes les partitions sont mutuellement exclusives, c'est-à-dire que l'intersection d'une partition avec une autre donne l'ensemble vide que l'on symbolise par  $\emptyset$ .

Par définition ( $\equiv$ ), l'univers est égal à la réunion de toutes les partitions:

$$U \equiv (X' \cap Y') \cup (X' \cap Y) \cup \dots$$

Ainsi, nous disons que les sujets sont répartis dans l'univers par le fait qu'ils appartiennent à différentes partitions. Par exemple, soit  $X$  être l'ensemble des sujets ayant été diagnostiqués positif et  $Y$  être l'ensemble des sujets ayant réussi, alors nous pouvons représenter l'ensemble des sujets diagnostiqués positif et ayant réussi par la partition  $(X \cap Y)$  et inversement, la partition  $(X' \cap Y')$  représente l'ensemble des sujets diagnostiqués négatif et ayant échoué.

De plus, Kosko (1993, p.30) associe chaque partition à un coin d'un cube à  $k$  dimensions. Ragin (2000, p.183) reprend ce concept d'espace de propriétés sous la forme d'un cube et fait correspondre un attribut à chaque dimension. Contrairement à une situation de mesure modélisée à l'aide de la théorie de réponse à l'item (TRI) où l'on recherche la précision dans une seule dimension, ici on veut davantage acquérir l'information par le nombre de dimensions ou d'attributs malgré l'imprécision pour le classement.

Le cardinal

Le cardinal d'un ensemble est symbolisé par # et représente le nombre d'éléments qui appartiennent à cet ensemble. Ainsi, le nombre de sujets ayant réussi peuvent être représenté par #Y. Il s'agit tout simplement d'un compteur.

• Les propriétés de l'algèbre de Boole

Une algèbre est une partie des mathématiques qui a pour objet la généralisation du calcul des nombres. Le tableau III décrit les principales propriétés de l'algèbre de Boole (Friedman, A.D., 1986. *Logical Design and Switching Theory*. Computer Science Press, Maryland, USA).

**Tableau III: Lois de l'algèbre de Boole**

<i>Loi</i>	<i>Opérations sur les ensembles</i>
Neutralité	$X \cup \emptyset \equiv X$
	$X \cap U \equiv X$
Absorption	$X \cap \emptyset \equiv \emptyset$
	$X \cup U \equiv U$
Idempotence	$X \cap X \equiv X$
	$X \cup X \equiv X$
Involution	$(X)' \equiv X$
Loi du tiers exclus	$X \cup X' \equiv U$
	$X \cap X' \equiv \emptyset$
Commutativité	$X \cup Y \equiv Y \cup X$
	$X \cap Y \equiv Y \cap X$
Associativité	$X \cup Y \cup Z \equiv X \cup (Y \cup Z) \equiv (X \cup Y) \cup Z$
	$X \cap Y \cap Z \equiv X \cap (Y \cap Z) \equiv (X \cap Y) \cap Z$
Distributivité	$X \cup (Y \cap Z) \equiv (X \cup Y) \cap (X \cup Z)$
	$X \cap (Y \cup Z) \equiv (X \cap Y) \cup (X \cap Z)$
Loi de DeMorgan	$(X \cup Y)' \equiv X' \cap Y'$
	$(X \cap Y)' \equiv X' \cup Y'$

L'algèbre de Boole permet d'opérer avec des ensembles sans recourir à la propriété d'additivité. Ainsi, on peut estimer la fidélité et la validité sans passer par le calcul des coefficients de corrélation entre variables.

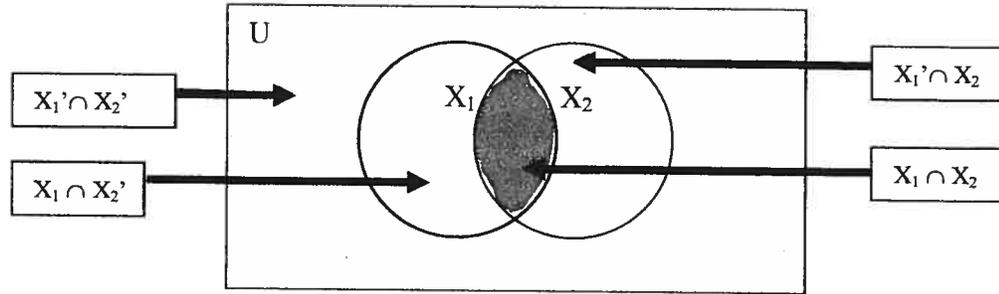
- La fidélité de la mesure critériée

On se rappelle que la fidélité est cette qualité que possède un instrument de mesure de refléter avec le plus de précision possible le trait mesuré ; son corollaire étant l'erreur de mesure. Et une mesure peut être précise sans pour autant être exacte, cela pose alors le problème de la validité. En fait, parler de la fidélité dans le cadre de la mesure critériée signifie qu'il faut se référer aux erreurs de classement qui peuvent être engendrées.

On retrouve quelques coefficients de fidélité pour la mesure critériée. Au plus simple, il y a l'indice  $p_0$  qui estime le rapport entre des partitions pour une situation avec des tests parallèles. Le coefficient kappa tient compte en plus de l'effet du hasard. Enfin, l'indice de Livingston considère l'écart entre le score brut et le score de césure.

Une situation de classement requiert au moins deux juges qui classent les éléments dans des ensembles. Pour une telle situation, Miles et Huberman (1991, p.63) définissent l'indice de fidélité comme étant le rapport entre le nombre d'accords inter-juges et le nombre de classements effectués.

Dans le contexte de la mesure critériée, le classement se fait d'une façon plus automatique en comparant le score brut d'un sujet,  $x_a$ , à un score de césure,  $x_0$ . Aussi, Subkoviak (1980) propose d'établir l'indice de fidélité selon une approche sur la consistance de la décision, c'est-à-dire à partir des résultats pour deux tests parallèles  $X_1$  et  $X_2$  ayant le même score de césure. Cette situation pour un univers de sujets est représentée par le diagramme de Venn de la figure 8.



**Figure 8:** Les diagnostics à des tests parallèles  $X_1$  et  $X_2$

Avec ce diagramme, l'univers se subdivise en quatre partitions puisqu'il y a deux ensembles en cause. Le classement dans l'un ou l'autre de ceux-ci résulte d'un diagnostic engendré par la comparaison du score brut avec le score de césure. Ainsi, la partition  $X_1 \cap X_2$  regroupe les sujets qui ont eu un diagnostic positif aux deux tests ( $\mu_{X_1}(a)=1$ ) et ( $\mu_{X_2}(a)=1$ ). À l'inverse, la partition  $X_1' \cap X_2'$  est plutôt associée aux sujets qui ont eu un diagnostic négatif dans les deux tests ( $\mu_{X_1}(a)=0$ ) et ( $\mu_{X_2}(a)=0$ ). Idéalement pour un test fidèle, nous ne devons retrouver des sujets que dans ces deux partitions que nous nommerons les partitions vraies. Si cette condition est remplie, alors la décision de classement sera consistante et les deux cercles doivent se confondre car il n'y a aucun sujet dans les partitions  $X_1 \cap X_2'$  et  $X_1' \cap X_2$ .

On a également comme indice de fidélité (Subkoviak, 1980 ; p.132) le rapport entre le cardinal de l'union des partitions vraies et le cardinal de l'univers (voir éq. 16).

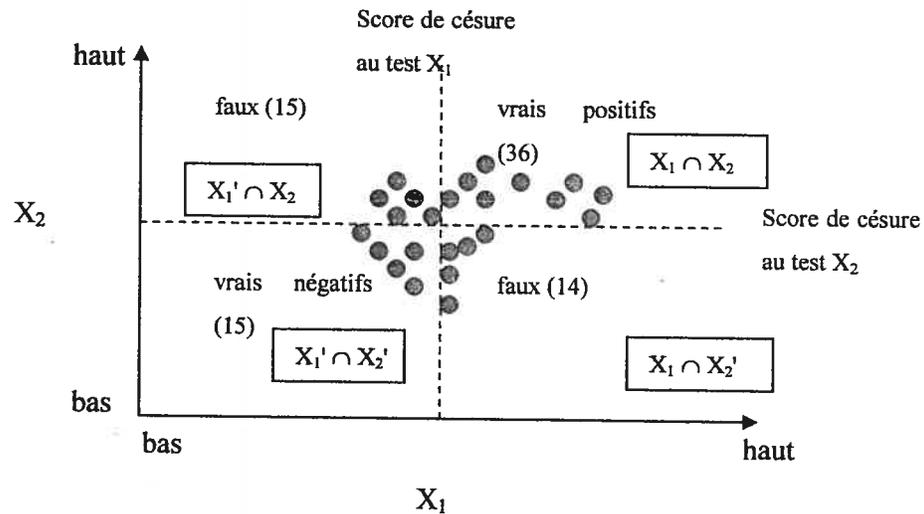
$$p_0 = \frac{\#((X_1' \cap X_2') \cup (X_1 \cap X_2))}{\#U} \quad (16)$$

La valeur de ce coefficient de fidélité se situe entre 0 et 1 mais cependant, il ne tient pas compte de l'effet du hasard. Aussi, le coefficient kappa « $\kappa$ » vient justement ajouter un paramètre pour compenser le hasard. Néanmoins, ces deux coefficients assument que la distribution des scores bruts épousent une forme en U où l'on ne retrouve que très peu de scores bruts près du score de césure. En

effet, on considère avec ces indices que la probabilité d'erreur ne dépend pas du score brut; plus particulièrement, de l'écart entre le score brut et le score de césure. Malgré la popularité de ces deux coefficients, nous ne pouvons les utiliser dans notre recherche car nous croyons que l'incertitude du classement pour un score brut situé près du score de césure engendre une incertitude pour établir le pronostic.

- L'erreur de mesure aléatoire

En pratique, il se peut que des sujets ayant obtenu un score brut près du score de césure se retrouvent dans les deux autres partitions. Bégin (1993) montre, avec un exemple graphique, (voir figure 9) comment on peut expliquer les erreurs avec une mesure critériée.



**Figure 9:** Les erreurs avec une mesure critériée

Chacun des quadrants de ce graphique correspond à une partition. Les erreurs se retrouvent dans deux quadrants et correspondent à deux situations possibles. La situation où le sujet obtient un score dans le test  $X_1$  au-dessous du score de césure et un score dans le test  $X_2$  au-dessus du score de césure, il s'agit de la partition  $X_1' \cap X_2$ . L'autre situation d'erreur est lorsque le sujet obtient un score dans le test  $X_1$  au-dessus du score de césure et un score dans le test  $X_2$  au-

dessous du score de césure, celle-ci constitue la partition  $X_1 \cap X_2'$ . Plus on retrouve de sujets pour ces deux situations moins le test est fidèle. Une situation démontrant une certaine linéarité avec un indice de fidélité élevé fait en sorte que l'on ne retrouve presque pas de cas dans les quadrants faux.

Si le nombre de sujets est approximativement égal pour chacune des deux situations d'erreur, et ce peu importe le nombre de sujets, alors nous pouvons dire que l'erreur est strictement aléatoire. Cependant, lorsqu'il y a plus de sujets pour une situation d'erreur, il y a une partie de l'erreur qui est due à un biais du score de césure et alors nous disons qu'il s'agit d'une erreur systématique ; en pratique, cela signifie que les deux tests ne sont pas parallèles. L'exemple du graphique de la figure 9 peut illustrer une situation avec une erreur aléatoire car il y a 15 sujets pour une situation et 14 pour l'autre, cela donne un indice de fidélité « $p_o$ » de 0,64  $((36+15)/80)$ .

Le problème de fidélité se pose surtout pour les sujets qui ont obtenu un score brut près du score de césure. En effet, on retrouve une zone d'incertitude, c'est-à-dire une zone de scores bruts où le diagnostic serait incertain. Auger et Séguin (1992, p.113) vont dans le même sens lorsqu'ils affirment:

*«Le modèle de décision doit considérer l'existence d'un seuil de réussite et d'une zone d'incertitude correspondante, établis à partir d'un groupe représentatif d'experts dans le domaine mesuré.»*

Selon Oosterhof (1990, p.158), plus le score brut se situe près du score de césure, plus grande est la probabilité de causer une erreur dans le classement. Brennan (1980, p.214) ajoute que la fidélité varie avec la valeur du score de césure et le nombre d'items dans le test si ces items ont un indice de difficulté identique.

- *L'indice de fidélité de Livingston*

Une autre façon de déterminer la fidélité pour un test critérié est de tenir compte de la variance des scores. Cependant, ici la variance n'est pas calculée par

rapport à la moyenne mais plutôt par rapport au score de césure (Hashway, 1998 ; p.138).

$$D^2 = \frac{\sum (x_a - x_0)^2}{N} \quad (17)$$

Où :  $D^2$  : variance ;  
 $x_a$  : score du sujet «a» ;  
 $x_0$  : score de césure ;  
 $N$  : nombre de sujets.

L'indice de covariance entre deux mesures critériées  $X_1$  et  $X_2$  est :

$$D_{X12} = \frac{\sum (x_{1a} - x_0)(x_{2a} - x_0)}{N} \quad (18)$$

Alors l'indice de fidélité devient (Hashway, 1998 ; p.139) :

$$r_{x0} = \frac{D_{X12}}{D_{X1} D_{X2}} \quad (19)$$

Étant donné que cet indice de fidélité est directement relié à la variance des scores bruts autour du score de césure et en postulant que l'erreur aléatoire est distribuée normalement, on peut recourir à l'erreur-type de mesure,  $\sigma_e$  (voir éq. 6), pour établir un intervalle de confiance où doit se situer le score de césure vrai et ce, afin d'établir une zone d'incertitude pour la décision. Ces déductions sont basées sur les axiomes de la TCT vues plus tôt.

Cet indice dit de Livingston va nous servir pour estimer la fidélité de l'épreuve diagnostique en mathématique. Ainsi, avec le calcul de l'erreur-type de mesure et l'intervalle de confiance qui s'en suit, nous allons pouvoir établir trois zones de décision: un diagnostic négatif certain, un diagnostic incertain et un diagnostic positif certain.

- La validité pronostique de la mesure critériée avec la théorie des ensembles

Dans le cas où l'on désire établir la validité pronostique d'un test critérié, on retrouve aussi deux classements: le diagnostic établi en comparant le score brut

avec le score de césure pour classer ou non les sujets dans l'ensemble X et le critère qui permet de classer ou non les mêmes sujets dans l'ensemble Y. De plus, Shaycoft (1979) mentionne que la décision subjective pour déterminer le score de césure cause un décalage de l'échelle de mesure, c'est-à-dire un biais que l'on associe à une erreur systématique et de ce fait, vient influencer grandement la validité. En effet, augmenter la valeur de ce paramètre signifie en langage populaire de mettre la barre plus haute; il y a alors moins de sujets diagnostiqués positivement mais une plus grande certitude que ces sujets réussissent. Cette situation augmente aussi l'incertitude pour prédire l'échec. Donc, avec un mauvais score de césure on ne pose pas le bon diagnostic.

Nous pouvons donc affirmer que le problème de la validité pronostique est directement relié à la valeur du score de césure et à la zone d'incertitude afférente. Alors, comment la valeur de ce score de césure permet de distinguer les sujets qui sont véritablement compétents de ceux qui ne le sont pas? La seule réponse possible sera de savoir si le diagnostic posé en comparant le score brut d'un sujet avec le score de césure permet de prédire sa réussite ou son échec.

Autrement dit, la réussite d'un sujet au test devrait être un bon pronostic de sa réussite au critère et le corollaire, l'échec d'un sujet au test devrait être aussi un bon pronostic de son échec au critère. En ce sens, il nous faut un modèle prédictif qui considère les variables de Bernouilli. Or, les modèles de régression statistiques linéaires utilisés couramment ne tiennent pas compte de ce type de variable. Pour une situation de mesure critériée où l'on désire établir un pronostic sur la réussite ou l'échec futur, il faut plutôt utiliser le modèle prédictif bayésien qui s'applique lorsque deux événements sont observés en deux temps différents, nous avons alors une probabilité *a priori* et une probabilité *a posteriori*. Et ce modèle se fonde sur un paradigme des probabilités subjectives.

Un événement est un fait qui se produit ou ne se produit pas. Plus précisément, un diagnostic et le fait qu'un sujet réussisse un cours ou un programme constituent des événements. Pour établir la validité pronostique d'un test, on

veut prédire la probabilité qu'un sujet réussisse étant donné qu'il a été diagnostiqué positif. Auparavant, on définit la probabilité *a priori* d'être diagnostiqué positif au présent par l'équation 20:

$$p(\mu_X=1) = \frac{\#X}{\#U} \quad (20)$$

Cette probabilité peut être établie à partir des résultats que l'on vient d'observer au test. Toutefois, pour mettre en évidence la relation entre deux événements, on utilise la probabilité *conjointe* établie par l'application de l'équation 21 sur des données du passé:

$$p(\mu_X \wedge \mu_Y=1) = \frac{\#(X \cap Y)}{\#U} \quad (21)$$

L'ensemble Y représente l'ensemble des sujets de l'univers qui ont réussi. L'expression  $p(\mu_X \wedge \mu_Y=1)$  représente la probabilité observée dans le passé qu'un sujet «a» quelconque ayant été diagnostiqué positif ait réussi (le symbole  $\wedge$  représente la conjonction et). Pour obtenir un indice de validité pronostique, nous pouvons recourir à la probabilité *a posteriori* estimée par l'équation 22 (Chou, 1975; p.410):

$$p(Y|X) = \frac{p(\mu_X \wedge \mu_Y=1)}{p(\mu_X=1)} = \frac{\#(X \cap Y)}{\#X} \quad (22)$$

L'expression  $p(Y|X)$  représente la probabilité de réussite d'un sujet étant donné qu'il a été diagnostiqué positif. Cette probabilité *a posteriori* pour le futur est établie en prenant le rapport entre la probabilité *conjointe* obtenue avec des données du passé et la probabilité *a priori* estimée avec l'administration actuelle du test. Nous disons que cette probabilité *a posteriori* est une probabilité subjective car elle se base sur la croyance que le processus reprendra dans le futur le même comportement que dans le passé.

Pour notre recherche, le processus sera le processus d'enseignement échelonné sur plusieurs années. L'établissement d'un pronostic devient plus difficile par le fait que l'observation du critère a lieu beaucoup plus tard après le diagnostic. Avec le modèle bayésien, le diagnostic ne peut prendre que deux valeurs: 0 ou 1; il n'y a pas de place pour l'incertitude dans la décision. Conséquemment, on ne peut appliquer directement ce modèle à notre situation de recherche car il ne tient pas compte d'une fidélité imparfaite pour l'épreuve diagnostique qui pourrait entraîner une plus faible valeur de la probabilité *a posteriori*. Toutefois, certains auteurs prétendent que les diagnostics locaux sont plus efficaces que les diagnostics globaux.

- *L'approche avec des diagnostics locaux*

À l'instar du médecin qui pose un diagnostic sur le taux de glycérides, le taux de mauvais cholestérol et le taux de bon cholestérol avant de poser un diagnostic global, l'éducateur peut poser des diagnostics locaux avant d'émettre un diagnostic global. Gitomer et Rock (1993) ont suggéré d'utiliser des blocs d'items ou sous-tests qui ont une cohérence interne avec un processus cognitif. Pour établir des diagnostics locaux, Oosterhof (1990, p.367) prétend qu'un test doit produire une série de scores dont chacun mesure une habileté spécifique.

Tatsuoka (1995), dans un article sur la structure des connaissances et le diagnostic cognitif, propose une approche fondée sur les attributs psychométriques et également une matrice d'incidence **Q** qui met en relation les réponses aux items du test avec des attributs psychométriques non observables.

Afin d'obtenir la validité prédictive d'un tel test, on voit apparaître la notion de réseau bayésien avec un article de Jameson (1996). Desmarais et Pu (2004) reprennent ce concept pour l'appliquer au testing adaptatif par ordinateur. À notre connaissance, aucune étude ne démontre la validité d'une telle approche pour établir un pronostic envers un critère éloigné dans le temps et lorsque l'instrument est utilisé pour une évaluation ayant une fonction d'orientation au début du processus.

## 6. *La théorie des ensembles flous, un changement de paradigme*

Tout comme pour la théorie des ensembles classiques, la théorie des ensembles flous nous provient des mathématiques. Par contre, cette dernière peut produire un discours en logique floue à partir de données expérimentales.

La logique est un domaine de la philosophie et des mathématiques qui a pris son essor dans l'Antiquité avec la civilisation grecque. Le plus connu de tous les philosophes logiciens grecs est sans nul doute Aristote qui établit le syllogisme pour obtenir une déduction d'une règle particulière à partir d'une prémisse générale. Au XIX<sup>e</sup> siècle, le mathématicien anglais Georges Boole réussit à modéliser des énoncés logiques sous la forme d'équations, mais avec des variables qui ne peuvent prendre que deux valeurs : 0 ou 1, faux ou vrai; ce sont ses travaux qui permirent de créer les premiers ordinateurs. Au début du XX<sup>e</sup> siècle, le mathématicien polonais Lukasiewicz décrit une logique à trois niveaux (Brule, 1992) et le logicien anglais Bertrand Russell s'intéresse à la possibilité d'une logique à niveaux multiples. En 1937, le scientifique Max Black énonce le principe d'un ensemble vague (Kosko, 1993; p.136). Cependant, ce n'est seulement qu'en 1965 que l'ingénieur Lofti Zadeh (1965) publie un article dans lequel il énonce les axiomes de la théorie des ensembles flous. Cette théorie intègre la notion d'incertitude dans l'appartenance à un ensemble et donne la possibilité d'assigner des valeurs entre le complètement faux et le complètement vrai.

- *La modélisation de la réalité*

Pour tout modèle, le défi est de tendre le plus possible vers la vérité, c'est-à-dire être le plus réaliste possible. Selon Kosko (1993, p.82), les philosophes distinguent la vérité logique de la vérité factuelle. La vérité logique vaut pour des énoncés décrivant des relations logiques ou mathématiques. Quant à la vérité factuelle, celle-ci vaut pour des énoncés décrivant des faits ou événements possibles, des sous-ensembles de continuums espace-temps possibles. La première vérité fait appel à la déduction alors que la seconde est plutôt affaire

d'induction. Cependant, il ne peut y avoir qu'une seule vérité, qu'elle soit logique ou factuelle.

Aussi, on parle de vérité logique à la suite d'une déduction logique. À partir d'une prémisse générale et d'une condition particulière on déduit une conclusion particulière qui ne peut qu'être vraie ou fausse. On dira donc qu'un énoncé est vrai du point de vue de la logique s'il répond au critère de la cohérence. La vérité factuelle, quant à elle, s'applique lorsqu'il s'agit d'induire un énoncé à partir de données, faits ou événements. Nous dirons alors qu'un énoncé correspondant au monde réel répond au critère de vérité factuelle. Par ailleurs, il subsiste toujours une incertitude entre le monde réel et la théorie.

Albert Einstein (cité dans Kosko, 1993; p.85) nous fait prendre conscience des difficultés inhérentes à la modélisation du monde réel :

*«Les preuves logiques diffèrent des tests scientifiques ou empiriques. Si vous prouvez un énoncé comme étant 100% vrai, il ne décrit pas le monde. S'il décrit le monde, vous ne pouvez pas le prouver.»<sup>5</sup>*

Einstein allègue que la vérité ne peut jamais être atteinte, on ne peut que s'en approcher ; cela va dans le même sens que la relativité épistémologique amenée par le philosophe Karl Popper (Debays, 1998).

Même Charlot (1995, p.33) renchérit:

*«La pratique n'est pas l'application d'une règle mais une réponse en situation qui comporte une part de flou et de vague.»*

Mais comme l'atteste Einstein, il n'y aura jamais certitude de la correspondance des résultats d'une épreuve avec une hypothèse cohérente. Traditionnellement, les chercheurs ont utilisé des paradigmes issus des théories des probabilités pour

---

<sup>5</sup> Traduction libre de : *Logical proof differs from empirical or scientific test. If you prove a statement 100% true, it does not describe the world. If it describe the world, you can't prove it.*

établir la vérité factuelle à partir des résultats expérimentaux. Ainsi, le degré de confiance est souvent mentionné dans une recherche pour indiquer les limites de la véracité de la conclusion. Mais ces paradigmes ignorent la part de flou et de vague que l'on retrouve dans une réponse en situation pratique.

- Les paradigmes issus des théories des probabilités

Smithson (1989) note que depuis le milieu du XIXe siècle, la science s'en est tenue à des paradigmes fondés sur des théories des probabilités pour établir la vérité factuelle. Cependant, on note une évolution qui passe de la théorie classique des probabilités, à la théorie de la probabilité fréquentielle et enfin à la théorie des probabilités subjectives.

La théorie classique des probabilités se réfère à des événements. Smithson (1989, p.51) décrit cette théorie en termes de nombres de possibilités qu'a un événement de se produire ou de ne pas se produire. La théorie des ensembles classiques n'a pas de difficultés avec la probabilité qu'un événement se produise, premièrement parce que sa définition est établie en termes du nombre de voies possibles à la production d'un événement et non pas au fait qu'il se soit produit ou se produira après plusieurs répétitions.

Selon Smithson (1989, p.52), le paradigme des probabilités fréquentielles se réfère à des choses que l'on peut compter et non plus à seulement un événement qui se produit ou pas. D'ailleurs, le paradigme moderne de probabilité dominant dans plusieurs champs disciplinaires est souvent référé comme fréquentiel. Par exemple, l'interprétation d'un score de 68 à un test de QI, caractérisé par une moyenne de 100 et un écart-type de 16 pour une population, nous permet d'affirmer que le sujet est lent intellectuellement. Cet exemple représente une application de la théorie des probabilités fréquentielles puisque la probabilité qu'une personne de la population obtienne un score inférieur à 68 n'est que de 2,28% selon le modèle de Gauss pour la distribution normale des fréquences de scores.

Smithson (1989, p. 59) décrit la théorie des probabilités subjectives en termes de degré de croyance qui est réfléchi dans la probabilité *a posteriori*. D'ailleurs, comme nous l'avons vu plus tôt, cette théorie peut être utile avec une interprétation critériée des scores pour prédire la performance à un critère observé plus tard.

Nous remarquons que l'évolution chronologique des théories des probabilités, du classique au subjectif, a toujours eu pour but de faire coller un modèle de plus en plus complexe pour se rapprocher de plus en plus de la réalité. Il faut voir la théorie des ensembles flous comme étant un effort de plus pour modéliser cette même réalité, pour rechercher la vérité factuelle. On confond souvent la théorie des ensembles flous avec les théories des probabilités, et davantage avec la théorie subjective ou bayésienne. Cette confusion provient possiblement du fait que les deux théories évaluent les variables sur l'intervalle de nombres réels de 0 à 1 [0, 1]. Avec la théorie bayésienne des probabilités, on observe à des temps différents si des éléments appartiennent ou n'appartiennent pas à des ensembles; le fait qu'un élément appartienne ou non à un ensemble est appelé un événement et la probabilité est déterminée en effectuant le rapport entre le nombre d'événements favorables (par ex: appartenance à un ensemble) et le nombre de possibilités. Quant à la théorie des ensembles flous, elle s'intéresse à l'incertitude du classement des éléments qui sera notée par une valeur entre 0 et 1.

- *Le paradigme de la théorie des ensembles flous*

La plupart des modèles que nous retrouvons dans les théories des probabilités filtrent une partie de la réalité car ils ont été élaborés à partir d'une logique binaire telle : un événement se produit ou ne se produit pas. Par exemple, la loi au Québec spécifie qu'une personne est considérée être un adulte lorsqu'elle a un âge de 18 ans ou plus ; cependant, dans la réalité l'attribut «adulte» pourrait être décerné à des personnes plus jeunes qui ont développé une grande maturité alors qu'on devrait peut-être l'enlever à des personnes immatures et plus âgées. En plus, il y a des comportements considérés matures et d'autres non, et ce peu

importe l'âge. La preuve que ce critère de passage à l'âge adulte est flou c'est qu'il diffère d'une province à l'autre.

La théorie des ensembles flous a pris naissance avec le besoin d'un plus grand consensus pour la modélisation de la réalité. Même si cela fait plus de 40 ans que Zadeh (1965) a écrit son premier article, ce n'est seulement qu'au cours des deux dernières décennies que l'on a pu mettre cette théorie en application dans plusieurs domaines. Selon le mathématicien Bart Kosko (1993), la logique floue s'est implantée plus facilement dans les pays asiatiques avec l'influence de la religion Bouddhiste ; nous, Occidentaux, avec l'influence de la logique bivalente d'Aristote depuis plus de 2000 ans, sommes habitués à voir les choses en noir et blanc alors que les Orientaux sont plus attachés à l'équilibre entre les oppositions (l'équilibre entre le Yin et le Yang avec la philosophie du Tao).

Pour Klir et Yuan (1995, p.30), le paradigme de la logique floue émerge du besoin de combler l'écart entre les modèles mathématiques et leurs interprétations empiriques. Cet écart tend à devenir de plus en plus perturbateur, spécialement dans les sciences biologiques, cognitives et sociales. Les mêmes auteurs (Klir et Yuan 1995, p.32) identifient quatre aspects qui font que le paradigme de la théorie des ensembles flous est supérieur à celui des théories des probabilités.

- Le nouveau paradigme permet d'exprimer l'observation et les incertitudes de classement dans leurs manifestations variées et font que ces incertitudes soient intrinsèques aux données empiriques.
- L'expérience générale vient nous dire que plus le problème est complexe, plus grande sera la supériorité des méthodes floues.
- Le nouveau paradigme a la capacité d'opérer en termes mathématiques avec des problèmes qui requièrent l'utilisation d'un langage naturel.
- Le nouveau paradigme offre plus de possibilités pour capturer le sens commun du raisonnement humain, la prise de décision et d'autres aspects de la cognition humaine.

- *La justification de l'utilisation de la théorie des ensembles flous*

L'utilisation de la théorie des ensembles flous peut être justifiée par le fait qu'elle permet de quantifier et d'analyser certains types de concepts vagues et ambigus (Smithson, 1989; p.93). Ni les théories des probabilités, ni les théories de la mesure sont capables de représenter la signification de propositions avec des prédicats flous (ex : petit, jeune, sûre, tôt), des quantificateurs flous (quelque, plus, plusieurs, souvent), des probabilités floues (vraisemblablement, conditionnellement, longuement), des valeurs vraies floues (très vrai, pas vrai, un peu vrai), et des modificateurs (très, plutôt, peu, quelquefois). En conséquence, les théories des probabilités ne permettent pas de faire des inférences à partir de prémisses floues et de ce fait, doivent laisser de côté les problèmes qui ont des composantes floues.

Même si nous retrouvons quelques cas d'ensembles précis dans les sciences de la nature, il reste que les systèmes plus complexes que nous retrouvons en sciences sociales se caractérisent plutôt par des ensembles flous. D'où l'importance d'une théorie des ensembles flous qui s'arrime mieux à la réalité en permettant à un élément de n'appartenir que partiellement à un ensemble

(Smithson, 1989, p.109). L'ouverture pour motiver l'utilisation des ensembles flous est l'observation que plusieurs ensembles et catégories naturelles ne sont pas précis dans le sens qu'un élément appartient à l'ensemble ou en est exclus. Aussi, là où la théorie des ensembles classique ne permet seulement de dire qu'une chose est rouge ou non-rouge, le langage ordinaire alloue des éléments à n'appartenir que partiellement aux choses rouges, comme le reflètent le mot rougeâtre ou l'expression très rouge.

La théorie des ensembles flous doit permettre l'élaboration de modèles plus collés à la réalité dans les domaines plus mous comme les sciences sociales. En ce sens, Ragin (2000, p.14) catégorise la plupart des recherches menées en sciences sociales selon deux types : les recherches orientées vers les variables ou dites quantitatives et les recherches orientées vers les cas aussi appelées qualitatives. Il propose un troisième type qu'il nomme la «recherche orientée vers la diversité». Il rattache deux aspects à la diversité (p.149) : le premier est la distinction entre les catégories et le second est le degré d'appartenance d'un cas à chacune de ces catégories. La recherche orientée vers les cas ne tient compte que du premier aspect alors que la recherche orientée vers les variables se limite presque exclusivement au second. Or, les concepts centraux dans les théories des sciences sociales sont souvent interprétés comme étant des ensembles et non des variables (p.167). Le même auteur utilise les ensembles flous pour lier l'argumentation théorique et l'analyse empirique car selon lui ils combinent l'évaluation qualitative et quantitative dans un simple instrument (p.8).

- La théorie des ensembles flous et la loi du tiers exclus

Une façon de distinguer un ensemble ordinaire d'un ensemble flou est de vérifier si le classement d'une donnée obéit ou non à la loi du tiers exclus. En effet, un ensemble ordinaire est toujours défini dans un ensemble de référence aussi appelé univers, il en va de même pour un ensemble flou. La seule distinction entre un ensemble ordinaire (en anglais *crisp set*) et un ensemble flou, demeure l'incertitude dans l'appartenance pour un élément. Aussi, nous verrons qu'un ensemble ordinaire constitue un cas particulier d'un ensemble flou et pour qu'il

n'y ait pas d'équivoque possible ou d'ambiguïté dans l'interprétation d'un ensemble flou, il faut le définir dans un univers précis.

Ainsi, si  $X$  est un ensemble ordinaire défini dans l'univers  $U$ , on peut affirmer que pour tout élément de l'univers  $U$  il est possible de connaître avec certitude si cet élément appartient à l'ensemble  $X$  ou non. Lorsqu'un élément n'appartient pas à  $X$ , alors il appartient obligatoirement à son complément  $X'$ .

Comme la loi du tiers exclus stipule qu'un élément ne peut appartenir à la fois à un ensemble et à son complément ; on ne peut être physiquement à la fois homme et femme. Le tableau III dans la section 5 du présent chapitre illustre sous la forme mathématique que l'intersection d'un ensemble ordinaire avec son complément donne toujours l'ensemble vide. De même l'union d'un ensemble ordinaire avec son complément donne toujours l'univers. Par exemple, d'un point de vue légal au Québec, si on est un humain on est soit de sexe masculin ou de sexe féminin ; par contre, d'un point de vue psychologique un garçon peut posséder en partie des caractéristiques féminines et une fille des caractéristiques masculines.

Un ensemble devient un ensemble flou lorsqu'il viole la loi du tiers exclus. En effet, si nous ne pouvons affirmer avec certitude qu'un élément appartient ou n'appartient pas à un ensemble, alors nous devons considérer cet ensemble comme étant un ensemble flou.

Toutes les propriétés de l'algèbre de Boole que l'on retrouve au tableau III dans la section 5 sont applicables aux ensembles flous à l'exception bien entendu de la loi du tiers exclus ; Ragin (2000) nomme ce nouvel algèbre, l'algèbre flou. De ce fait, nous pouvons affirmer qu'un ensemble ordinaire constitue un cas particulier d'ensemble flou.

Afin de représenter l'incertitude pour un élément d'appartenir à un ensemble flou, nous utiliserons le concept d'appartenance. Dans le cas d'une mesure des

apprentissages, Samejima (1995, p.398) propose d'utiliser l'expression «degré de maîtrise».

- Les applications de la théorie des ensembles flous dans d'autres domaines

Cette théorie des ensembles flous trouve ses premières applications dans le domaine du génie électrique et plus particulièrement en automatisation. D'ailleurs, l'organisme international «Institute of Electrical and Electronics Engineers» (IEEE) y consacre une revue, la revue «Transactions on Fuzzy Systems».

Peu à peu, cette théorie trouve des applications dans les autres domaines des sciences de la nature comme en informatique pour concevoir des systèmes experts (Kandel, 1992). On se sert également des ensembles flous pour faire des prédictions météorologiques (Hansen, 2000). Au Québec, Sangali (1998) compare la même logique floue à un réseau de neurones.

Ensuite, les sciences administratives et humaines telles la gestion, l'environnement et la géographie commencent à utiliser cette théorie. Dans le domaine de la gestion par exemple, Von Altrock (2000) publie un livre sur les applications de la logique floue en finance et dans les affaires. En environnement, un groupe de chercheurs français (Freissinet et al, 1998) utilise les ensembles flous pour modéliser la composition des eaux souterraines. En géographie, Grasland (1997) propose les ensembles flous pour l'étude des maillages territoriaux, c'est-à-dire plus particulièrement pour décrire les caractéristiques des populations situées près des frontières physiques.

- Les applications de la théorie des ensembles flous en éducation

La théorie des ensembles flous trouve aussi des applications dans le domaine des sciences de l'éducation.

Ammar et al, (2000) utilisent cette théorie pour identifier les écoles publiques élémentaires de la Cité de NewYork ayant une faible performance. Plus particulièrement, ils utilisent les résultats obtenus par les élèves à des tests

standardisés en lecture et en mathématique pour classer les écoles selon diverses méthodes traditionnelles et la méthode fondée sur les ensembles flous. Or, la méthode avec des ensembles flous permet de nuancer les informations et correspond davantage au jugement des experts.

Aussi, Crowther et Batchelder (1995) développent un modèle de perception en logique floue qui comprend trois niveaux: l'évaluation du trait, l'intégration du trait dans un schéma et la classification des schémas. Pour l'évaluation du trait, on utilise des valeurs vraies floues telles 0,5 pour représenter une situation complètement ambiguë et 0,7 pour représenter un cas plus vrai que faux. Lors de l'intégration du trait, les valeurs vraies floues obtenues par l'évaluation sont alors combinées ensemble et comparées à un schéma prototype. Les auteurs démontrent que la probabilité de classer correctement un sujet «j» en fonction d'un stimulus «i» avec ce modèle de perception en logique floue sera équivalente à la probabilité de répondre correctement à un item «i» pour un sujet ayant une habileté « $\theta_j$ » selon le modèle de Rasch (p.404). Selon eux, cette équivalence ouvre la porte à l'utilisation de logiciels spécialisés pour effectuer des inférences statistiques. Dans notre recherche, nous voulons classer un sujet dans un ensemble flou en fonction de sa capacité en mathématique et le modèle de Rasch pourra aussi servir de référence.

Heiser et Groenen (1997) utilisent pour leur part la théorie des ensembles flous pour établir une stratégie d'approximation successive. Grâce à cette stratégie, ils peuvent éliminer les minima locaux lors de l'estimation de la valeur des données psychométriques pour une grappe de sujets.

Katz et Lesgold (1992) utilisent plutôt une technique de modélisation floue pour représenter un environnement d'instruction pratique portant sur le dépannage en électronique. Cet environnement sert pour développer un tuteur informatisé faisant appel à l'intelligence artificielle.

Quant à Pryor et Hesketh (1989), ils développent une échelle d'appréciation floue pour les tests d'orientation professionnelle. Dans un premier temps, le répondant indique un choix ponctuel comme on le fait traditionnellement, mais par la suite il ajoute la tolérance qu'il donne autour de ce choix. L'incertitude ainsi obtenue sera modélisée par la logique floue.

Hawkes *et al.*, (1992) développent un système expert flou qui fait office de tuteur intelligent informatisé. Lors de la résolution d'un problème, la démarche de l'étudiant se compare à celle d'un expert et une rétroaction lui est donnée. Les auteurs prétendent démontrer l'efficacité d'un tel système lors de l'entraînement à résoudre des problèmes d'arithmétique chez les élèves du primaire. Ils assument que ce système pourrait aussi être utilisé pour des adolescents de niveau secondaire.

Et Yi *et al.*, (1993) utilisent la théorie des ensembles flous pour représenter une matrice d'incidence  $Q$  sur les attributs psychométriques dans une application de testing par ordinateur. Ils arrivent à la conclusion que pour un ajustement fin des scores de césure, l'approche avec les ensembles flous peut être préférable à la méthode bayésienne.

De même Fourali (1997) recourt à la logique floue pour représenter l'évaluation des apprentissages à l'aide d'un portfolio. Dans son article, il présente une échelle d'appréciation floue de la compétence d'un sujet qui gradue de définitivement incompetent à définitivement compétent en passant par une position mitoyenne qui représente l'incertitude de l'évaluateur.

Et Jameson (1996), dans un article sur la gestion de l'incertitude numérique dans la modélisation de l'étudiant et de l'utilisateur, compare la théorie des ensembles flous (TEF) aux réseaux bayésiens et à la théorie de l'évidence Dempster-Shafer. Il conclut que la TEF permet de décrire un raisonnement en termes naturels et de l'expliquer avec des opérateurs et des règles par opposition

aux principes et règles mathématiques qui n'utilisent que des concepts précis. La connaissance d'un expert peut ainsi être modélisée.

Également Lopes *et al.*, (1997) utilisent la théorie des ensembles flous pour l'évaluation de la performance des départements à l'intérieur d'une université. À partir de données quantitatives sur les produits de départements (publication d'articles, nombre de diplômes décernés, etc), ils développent des indicateurs de qualité. Ils trouvent que la productivité semble une condition nécessaire mais non suffisante pour obtenir une performance de qualité. De plus, une procédure d'évaluation globale est adoptée et constitue désormais une partie formelle du processus d'allocation des ressources. Enfin, la grande sensibilité du modèle aide à prendre des décisions de gestion pour améliorer la performance des pires cas.

De même dans sa thèse de doctorat, Rasmani (2002) utilise la théorie des ensembles flous pour évaluer la performance des étudiants dans un cours universitaire. À partir de données quantitatives pour trois attributs, il établit des règles en logique floue pour inférer une appréciation finale à trois niveaux: pauvre, moyen, bon. En plus d'être équivalente à la moyenne statistique des cotes, l'auteur prétend que cette nouvelle approche peut tenir compte des informations linguistiques et de ce fait, facilite la prise de décisions pour des cas limites.

Tous les écrits de ces auteurs cités plus haut illustrent bien la pertinence de cette modélisation pour la science en général et pour l'éducation en particulier. Maintenant, voyons comment cette théorie peut contribuer à l'élaboration d'un modèle pronostique pour une mesure référée à un critère.

### ***7. La théorie des ensembles flous : une théorie de la mesure critériée***

Lors de la classification d'un élément, nous éprouvons quelquefois des doutes sur notre choix; c'est justement cette incertitude que le degré d'appartenance à un ensemble flou veut modéliser. Par exemple, il est assez facile de déterminer le sexe d'une personne à partir de ses caractéristiques alors que le jugement de la

compétence en mathématiques de cette même personne est une tâche plus complexe.

- La définition de l'univers

Avant de déterminer l'appartenance d'éléments à l'univers, il serait bon de définir ce dernier. Comme nous l'avons dit plutôt, l'univers doit être un ensemble précis en ce sens que l'on ne doute pas de l'appartenance de ses éléments. Une erreur courante lorsqu'on définit un univers c'est de le confondre avec une population et de n'avoir qu'une vision statique. Au contraire, le concept d'univers peut aussi s'étendre à une perspective dynamique et représenter un continuum espace-temps. Ainsi une cohorte de sujets observée sur plusieurs années peut constituer un tel continuum espace-temps.

Nous disons alors qu'une cohorte d'étudiants correspond à un univers restreint alors que plusieurs cohortes ayant les mêmes caractéristiques forment un univers étendu pouvant être associé à une population. Dans cet univers restreint, nous retrouverions: un ensemble flou X pour représenter le diagnostic sur la capacité en mathématique, un ensemble flou Y pour représenter le critère de la réussite mais aussi d'autres ensembles pour représenter les attributs des sujets.

Rappelons que Ragin (2000, p.63) définit un univers composé d'ensembles comme étant un espace de propriétés. En effet, chaque propriété illustre un attribut associé à un ensemble. Il ajoute que ces espaces de propriétés sont des hypothèses de travail pouvant être révisées à n'importe quelle étape du processus de recherche.

- Le degré d'appartenance à un ensemble flou

Plusieurs auteurs contribuent à l'élaboration de la théorie des ensembles flous. Huang (1999, p.99) émet l'opinion que la théorie des ensembles flous doit permettre l'élaboration de modèles sur l'imprécision du monde réel. Zadeh (1965, p.339) propose d'assigner des valeurs d'appartenance aux éléments, ces valeurs se situent entre 0 et 1, 0 représentant la certitude de ne pas appartenir à un ensemble et 1 la certitude d'y appartenir. Klir et Yuan (1995, p.11) précisent

cet axiome en parlant d'une fonction d'appartenance,  $\mu_A(a)$ , qui associe à chaque élément «a» de l'univers U une valeur d'appartenance à un ensemble quelconque A; ceci se note par l'expression suivante:

$$\mu_A(a) : U \rightarrow [0, 1] \quad (23)$$

Une valeur d'appartenance de 0,5 signifiant qu'il y a une incertitude quant à savoir si un élément «a» appartient ou non à l'ensemble A. Une fonction d'appartenance assigne aux éléments «a» de l'univers U, qui est toujours un ensemble précis, des nombres réels de l'intervalle [0, 1]. Étant donné qu'un élément «a» appartient toujours certainement à l'univers U puisque ce dernier est un ensemble précis, nous pouvons conclure que la somme des degrés d'appartenance d'un élément à un ensemble et à son complément est toujours égale à 1.

$$\mu_A(a) + \mu_{A^c}(a) = 1 \quad (24)$$

Par exemple, le jugement porté par rapport à un item à réponse construite dans une épreuve pourrait être traduit en un degré d'appartenance à un ensemble flou. Une valeur de 1 voudrait dire que le sujet a répondu entièrement à la question mais une réponse partielle recevrait un score compris entre 0 et 1.

- L'échelle de mesure ordinale avec la théorie des ensembles flous

Pour Crowther et Batchelder (1995, p.399), l'appréciation des valeurs d'appartenance à des ensembles flous ne constitue pas une mesure sur une échelle d'intervalle; en effet, même si les valeurs d'appartenance sont représentées par des nombres réels, il demeure qu'elles sont confinées à l'échelle unitaire [0,1] et ne peuvent pas accepter des transformations d'échelle. Donc, déterminer le degré d'appartenance d'un élément à un ensemble flou signifie qu'on mesure cet élément sur une échelle ordinale et de ce fait, qu'il n'y a pas d'unité de mesure. De toute façon, on ne cherche pas à effectuer des opérations additives avec les données.

En éducation, beaucoup de chercheurs tentent d'élaborer des échelles ordinales pour mesurer le comportement humain. En effet, depuis les années 50, plusieurs auteurs tentent de catégoriser les apprentissages par niveau de tâches, c'est ce que l'on a appelé les taxonomies. Tout d'abord, il y a Bloom (1982) et ses collaborateurs élaborant une taxonomie des objectifs d'apprentissage dans le domaine cognitif en six niveaux. Par la suite, Gagné (St-Yves, 1982; p.38) nous propose trois niveaux de capacité.

En fait, ces taxonomies constituent des échelles de mesure ordinales avec des repères pour chaque échelon; ces repères constituant des bouées d'ancrage dans le monde réel. Voici comment Dassa et Laurier (2003, p.120-121) perçoivent le rôle de l'échelle ordinale en éducation:

*«Les débats relativement au bulletin descriptif au Québec confirment la difficulté que pose le choix entre un résultat global dont l'univocité n'est souvent qu'apparente et une information diagnostique qui présente les forces et les faiblesses de l'élève. L'application d'échelles descriptives dans le cadre de programmes construits selon une logique de compétences apparaît comme une solution avantageuse. Chaque compétence est alors jugée en se référant à une échelle dont chaque échelon correspond à un portrait type des manifestations de la compétence à un certain degré de son développement.»*

De plus, Ragin (2000, p.109) mentionne qu'il faut associer des expressions qui qualifient les degrés d'appartenance, qui pour nous seront des prédicats. Il préconise une méthode pour étalonner une fonction d'appartenance (p.158) :

«La première tâche (.) serait de spécifier trois ancres qualitatifs importants : le point auquel il y aurait certitude d'appartenance, le point auquel il y aurait certitude de non appartenance et le point où il y aurait le maximum d'ambiguïté.»<sup>6</sup>

On donne alors une valeur de 1,0 au premier point d'ancrage, 0,0 au second et 0,5 au troisième.

- La fonction d'appartenance discrète

Afin de pouvoir classer des éléments dans un ensemble flou, Huang (1999, p.106) propose une description qualitative des degrés d'appartenance en cinq échelons. En effet, comme l'indique le tableau IV, il fait correspondre une étiquette qualitative, aussi appelée prédicat, à des intervalles de degré d'appartenance.

**Tableau IV: L'appartenance à un ensemble flou**

Étiquette	Prédicat	$\mu_A(a)$
A	Certainement vrai	De 0,8 à 1,0
B	Possiblement vrai	De 0,6 à 0,8
C	Incertain	De 0,4 à 0,6
D	Possiblement faux	De 0,2 à 0,4
E	Certainement faux	De 0,0 à 0,2

À partir d'une situation de classement et de données numériques discrètes, nous pouvons déduire, en s'aidant des appréciations qualitatives du tableau IV, un barème de degrés d'appartenance pour un ensemble flou. Tout d'abord, il faut savoir qu'il y a trois ancres de repères pour l'étalonnage. Un premier point pour la certitude d'appartenance ( $\mu = 1$ ), un second pour la certitude de non

---

<sup>6</sup> Traduction libre de : *The first task (.) would be to specify three important qualitative anchors : the point at which full membership is reached, the point at which full non-membership is reached, and the point of maximum ambiguity.*

appartenance ( $\mu = 0$ ), un troisième pour le maximum d'incertitude dans l'appartenance ( $\mu = 0,5$ ).

Ragin (2000, p.156) propose, quant à lui, cinq types d'échelle pour une situation de classement : l'ensemble ordinaire binaire, l'ensemble flou à trois échelons, l'ensemble flou à cinq échelons, l'ensemble flou à sept échelons, l'ensemble flou continu. Le tableau V illustre ces échelles avec le prédicat pour la proposition «le cas est inclus dans l'ensemble».

**Tableau V : Les échelles pour une situation de classement (Ragin, 2000)**

2 éch.	3 éch.	5 éch.	7 éch.	continue	Prédicat
1,0	1,0	1,0	1,0	1,0	Certainement vrai
			0,83	scores	Semble vrai
		0,75		numériques	Plus vrai que faux
			0,67	>0,5 et <1	Plus ou moins vrai
	0,5	0,5	0,5	0,5	Ni vrai, ni faux
			0,33	scores	Plus ou moins faux
		0,25		numériques	Plus faux que vrai
			0,17	>0 et <0,5	Semble faux
0,0	0,0	0,0	0,0	0,0	Certainement faux

On a ainsi un choix plus grand d'échelles pour les différentes situations de classement susceptibles d'être rencontrées.

Par exemple, pour déterminer l'échelle de la réussite dans un programme, on doit au préalable établir quelles données correspondent aux trois points d'ancrage que voici: la réussite certaine, l'échec certain et le maximum d'ambiguïté. Par la suite on choisit le nombre d'échelons en fonction des données disponibles.

- La fonction d'appartenance continue comme modèle

Nous ne pouvons utiliser une fonction d'appartenance discrète pour représenter une situation de testing où nous avons une possibilité de plusieurs scores distincts; dans ce cas, il est préférable d'utiliser une fonction d'appartenance continue. Et un sujet qui réussit tous les items d'un test maîtrise l'objectif mesuré par ce même test alors que celui qui n'en réussit aucun ne le maîtrise pas du tout ; cependant, nous ne pouvons rien affirmer pour un sujet obtenant un score intermédiaire. Conséquemment, il faut établir un modèle qui permette de transformer un score en un degré de maîtrise. Par exemple, pour l'épreuve diagnostique en mathématique, le score sera traduit en un degré de capacité qui correspond au degré d'appartenance d'un sujet à l'ensemble flou X.

Dans le contexte de poser un diagnostic cognitif global à partir d'une multitude de diagnostics locaux, comme avec l'approche fondée sur les attributs psychométriques, Samejima (1995, p.398) suggère d'adopter une fonction logistique entre le degré de maîtrise et le trait mesuré. Cela va dans le sens de la suggestion de Smithson (1987, p.79) qui mentionne qu'une fonction d'appartenance requiert une forme en S, car la valeur d'appartenance doit tendre vers 1 lorsque le score observé,  $x$ , tend vers le maximum, cette même valeur d'appartenance doit tendre vers 0 pour un score qui tend vers le minimum et une fonction continue et différentiable détermine cette même valeur d'appartenance pour les scores intermédiaires..

Cette fonction est dérivée du postulat additionnel qu'un accroissement dans la croyance individuelle que le sujet «a» ayant obtenu un score spécifique,  $x_a$ , soit élément de A,  $a \in A$ , est proportionnel au produit des degrés de croyance que selon son score observé, il est élément de A,  $a \in A$  ( $\mu_A(x)$ ), et n'est pas élément de A,  $a \notin A$  ( $1 - \mu_A(x)$ ). Si  $d\mu_A(x)/dx$  représente la dérivée par rapport à «x» de l'appartenance et «c» être une constante de proportionnalité quelconque, alors la fonction logistique de l'équation 26 est la solution à l'équation différentielle 25:

$$d\mu_A(x)/dx = c \mu_A(x)(1-\mu_A(x)) \quad (25)$$

$$\mu_A(x) = \frac{1}{1 + e^{-\alpha(x-\beta)}} \quad (26)$$

- Où: e: constante valant 2,718....  
 x: score observé;  
 $\alpha$ : paramètre déterminant la largeur de la zone d'incertitude;  
 $\beta$ : paramètre déterminant le point de césure;  
 $\mu_A(x)$ : appartenance à l'ensemble A selon le score d'un sujet.

Cette fonction est identique à celle utilisée par le modèle de Rasch à une différence près, dans le modèle de Rasch on utilise la variable latente  $\theta$  alors qu'ici on utilise la variable manifeste X qui est continue. Les trois points d'ancrage sont:

$x_{1/2} = +\beta$ , alors $\mu_A(x) = 1 / (1 + e^0) = 0,5$	<i>point de césure</i>
$x_0 = +\beta - 3/\alpha$ , alors $\mu_A(x) = 1 / (1 + e^{+3}) = 0,05$	<i>certitude que a <math>\notin</math> A</i>
$x_1 = +\beta + 3/\alpha$ , alors $\mu_A(x) = 1 / (1 + e^{-3}) = 0,95$	<i>certitude que a <math>\in</math> A</i>

Ici, le niveau de certitude est de 95%. Le point de césure,  $x_{1/2}$ , correspond à un maximum d'incertitude à savoir si l'élément appartient ou non à l'ensemble. Le point  $x_1$  correspond à un standard de performance, c'est-à-dire le score qu'un sujet doit atteindre pour que l'on soit certain qu'il a la capacité d'entreprendre des études d'ingénieur. Le point  $x_0$  correspond plutôt à un standard de contre-performance; en d'autres mots, on est certain que le sujet n'a pas la capacité d'entreprendre des études d'ingénieur s'il obtient un score inférieur à ce standard.

- La fidélité de la mesure avec la théorie des ensembles flous

Avec la théorie des ensembles flous, on fait une interprétation critériée de la mesure et on indique en plus le degré de certitude du diagnostic. En effet, le modèle, déterminé par la fonction d'appartenance continue, reflète la caractéristique du test à discriminer les sujets par rapport au critère mais indique également le degré de certitude pour cette discrimination. Dans le même ordre d'idées, Laurencelle (1997, p.29) affirme:

*«La capacité discriminante (.) est essentiellement basée sur la fidélité de l'instrument de mesure.»*

Cette capacité discriminante pour un test critérié dépend en grande partie des intervalles de confiance pour les diagnostics positif et négatif. Plus on retrouve de sujets avec un score brut près du score de césure, plus forte est la probabilité de poser un diagnostic erroné; il y a donc une zone d'incertitude autour de ce point. D'ailleurs, Allen et Yen (1979, p.228) affirment qu'idéalement la distribution des scores à un test critérié épouse une forme en U de manière à ce qu'il y ait le moins de scores possibles près du score de césure. Par conséquent, les intervalles de confiance sont les zones complémentaires à cette zone d'incertitude.

Le paramètre  $\alpha$  de la fonction caractéristique décrite par l'équation 26 peut modéliser cette incertitude dans le diagnostic et indirectement, détermine les intervalles de confiance. Aussi, le graphique de la figure 10 illustre le modèle pour une mesure référée à un critère et est identique à celui utilisé en électronique pour la communication numérique sauf que l'échelle de l'abscisse est alors le potentiel du signal en volts ; par exemple, la norme TTL (transistor-transistor-logic) établit qu'un potentiel inférieur à 0,8 V représente un «0» et un potentiel supérieur à 2,0 V signifie un «1», tout signal à un potentiel entre ces limites est considéré incertain et est susceptible d'engendrer du bruit dans la communication.

### Mesure critériée floue

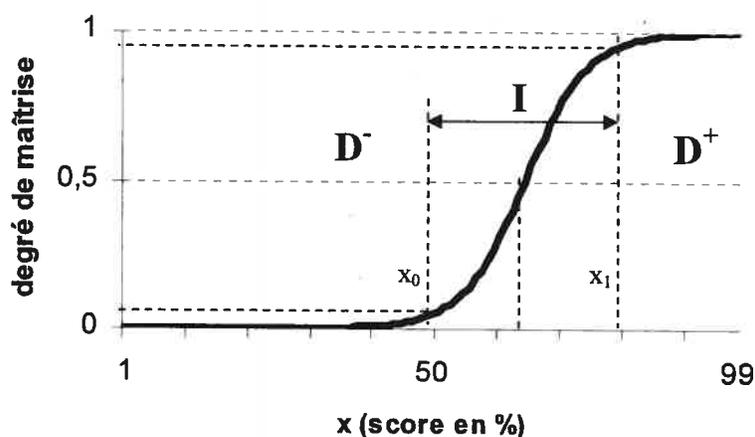


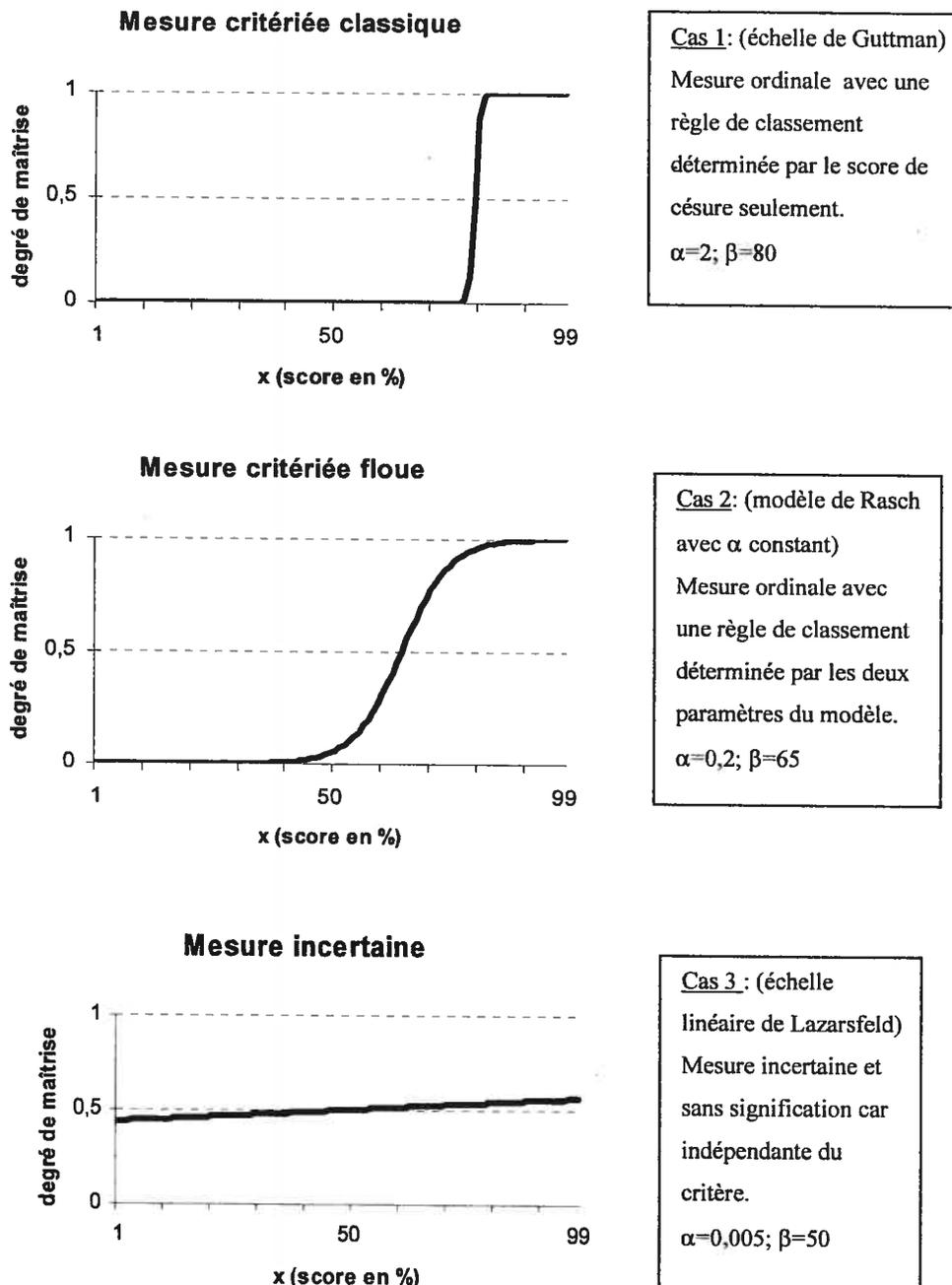
Figure 10 : Le modèle d'une mesure critériée floue

La zone I sur le graphique représente la zone des scores où il y a de l'incertitude pour le diagnostic et est délimitée par le standard de performance,  $x_1$ , et le standard de contre-performance,  $x_0$ . La zone  $D^+$  correspond à l'intervalle de confiance pour poser un diagnostic positif, c'est-à-dire tous les scores supérieurs au standard de performance,  $x_1$ . Inversement, la zone  $D^-$  est associée à l'intervalle de confiance pour poser un diagnostic négatif et elle comprend tous les scores inférieurs au standard de contre-performance,  $x_0$ .

Ce modèle possède un score de césure,  $\beta$ , à 65% et un paramètre  $\alpha$  de 0,2. Dans ce cas-ci, le standard de performance est donc fixé à 80% et le standard de contre-performance à 50%.

- Les variations du modèle d'une mesure critériée

Les graphiques de la figure 11 illustrent trois cas pour le modèle.



**Figure 11: La fonction caractéristique d'une mesure critériée**

Le cas 1 représente une mesure critériée classique consistant en une classification dans un ensemble précis, la zone d'incertitude est réduite à néant et

donc le paramètre  $\alpha$  est grand (2 dans ce cas) ; ici, on ne tient pas compte de la capacité discriminante du test. Le score de césure,  $\beta$ , est égal au standard de performance et il n'y a pas de standard de contre-performance ; dans ce cas-ci, ce paramètre vaut 80%. Cela ressemble à l'échelle de Guttman (voir figure 3) qui représente une situation de mesure idéale.

Quant au cas 2, il est associé à une situation de mesure critériée floue où l'incertitude du diagnostic est indiquée, alors le paramètre  $\alpha$  est moyen (0,2 dans ce cas). La valeur de ce paramètre dépend de la fidélité du test et de ce fait, de sa capacité discriminante. Il représente le cas pratique où un test n'a pas une fidélité parfaite. C'est un peu comme le modèle de Rasch (voir figure 4) qui est plus collé à la réalité que l'échelle de Guttman.

Enfin, le cas 3 représente une mesure indépendante du degré de maîtrise, car le diagnostic est toujours incertain quelque soit le score du sujet; le paramètre  $\alpha$  est alors minime (0,005 dans ce cas).

À noter qu'une valeur négative pour le paramètre  $\alpha$  vient signifier une relation inverse entre le degré de maîtrise et le score observé. Par exemple, dans le cas d'une dictée, le score représente le nombre de fautes et plus il est élevé moins le sujet est compétent. Néanmoins, nous ne retrouvons pas une valeur négative de  $\alpha$  avec le modèle de Rasch car la fonction est monotone et croissante.

Plus le test discrimine facilement les sujets par rapport au critère, plus le paramètre  $\alpha$  est élevé et à la limite le modèle ressemble au cas 1. À l'inverse, moins le test discrimine les sujets, plus le paramètre  $\alpha$  est faible et à la limite il n'y a pas de discrimination du tout comme dans le cas 3. Ce paramètre doit donc être déterminé dans le but d'indiquer les erreurs possibles de diagnostic.

### 8. *Les concepts de la TEF*

Il ne faut pas confondre le modèle pour une mesure critériée avec celui que nous retrouvons dans la théorie de réponse à l'item (TRI). Ici la variable dépendante en ordonnée représente le degré de vérité dans l'affirmation «le sujet est compétent» alors que pour le modèle de la TRI, elle représente la probabilité que le sujet réponde correctement à un item donné. Avec la TRI, nous avons un modèle pour chaque item alors qu'ici il n'y a qu'un seul modèle pour le test. Le but recherché avec la TRI est d'inférer une habileté sur une échelle d'intervalle pour un sujet quelconque et ce, à partir de la réussite à plusieurs items qui ont des modèles de mesure distincts. Ici, le but visé est de poser un diagnostic sur une échelle ordinale et d'indiquer le degré de certitude afférent ; ceci se fait à partir du score à un test caractérisé par un seul modèle de mesure.

L'avantage d'obtenir le degré de vérité est surtout utile lors de la prise de décision suivant un diagnostic. Ainsi un diagnostic certain doit entraîner irrémédiablement une prescription alors qu'une décision en rapport avec un diagnostic incertain doit plutôt prendre la forme d'une information donnée au sujet. De plus, les opérateurs de la logique floue permettent d'obtenir différents indicateurs dont l'inclusion d'un ensemble dans un autre qui facilite l'analyse de causalité nécessaire pour établir la validité pronostique d'un test.

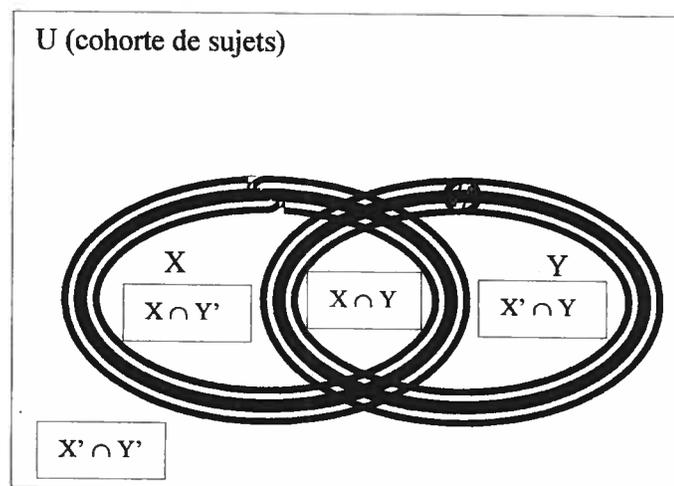
- *L'analyse de la causalité*

Il faut voir la validité comme étant une relation entre la mesure d'un trait et un critère ; auparavant, on postule qu'il existe une relation de cause-à-effet entre ce même trait et ce même critère. Selon De Landsheere (1979, p.17), une cause est :

*«Un antécédent nécessaire, c'est-à-dire une condition qui précède toujours l'apparition d'un phénomène donné et en l'absence de laquelle le phénomène ne se produit jamais.»*

Par exemple, dans le cadre de notre recherche l'appartenance à l'ensemble X dont l'attribut est «capacité en mathématique» doit être une condition nécessaire

pour l'appartenance à l'ensemble  $Y$  ayant comme attribut «réussite». Toutefois, pour que  $X$  soit la cause de l'effet  $Y$ , il faut aussi que la non appartenance à  $X$  (appartenance à  $X'$ ) soit une condition nécessaire pour la non appartenance à  $Y$  (appartenance à  $Y'$ ). Ragin (2000) propose une méthode pour vérifier la nécessité et la suffisance d'une condition en prenant en considération d'autres attributs intervenants. Le diagramme de Venn de la figure 12 représente la situation de mesure avec une cohorte de sujets. Les ensembles  $X$  et  $Y$  sont des ensembles flous et représentent respectivement les sujets diagnostiqués positifs et les sujets qui ont réussi; ils sont symbolisés avec un trait épais à trois bandes pour signifier justement que leur frontière est floue.



**Figure 12:** Le diagramme de Venn pour la validité pronostique

Cet univers comprend  $2^2$  ou 4 partitions. À noter que le nombre de partitions pour des ensembles flous est identique à celui pour des ensembles ordinaires. Cependant, les frontières floues des ensembles signifient qu'un sujet peut appartenir à plus d'une partition, ce qui n'est pas le cas pour des ensembles ordinaires.

Rappelons que pour notre recherche la fonction d'appartenance de l'ensemble  $X$  est continue et épouse la forme d'une fonction logistique à deux paramètres.

Une variation du paramètre  $\beta$  produit un déplacement du score de césure et de ce fait, un décalage de la fonction caractéristique. Or rappelons-le, une mauvaise valeur pour ce paramètre peut entraîner une erreur systématique de la mesure. Autrement dit, il se peut que le score de césure soit trop faible et que l'on diagnostique positivement un groupe de sujets qui sont en réalité inaptes ou l'inverse, un score de césure trop élevé fait en sorte que l'on diagnostique négativement des sujets aptes. L'épreuve est valide si les paramètres  $\alpha$  et  $\beta$  du modèle sont adéquats. Afin de calibrer ce modèle avec les bons paramètres, il faut s'ancrer dans la réalité en établissant le pronostic du critère de la réussite par la méthode de l'analyse causale.

Pour que l'épreuve soit valide, il faut que le sujet ayant été diagnostiqué compétent réussisse mais aussi que le sujet ayant été diagnostiqué incompetent échoue; cette situation correspond respectivement aux partitions  $X \cap Y$  et  $X' \cap Y'$ . Autrement dit, il faut que la très grande majorité des sujets appartiennent à ces deux partitions.

Les opérateurs de la logique floue vont donc nous permettre d'opérer avec ces concepts de partition et d'intersection.

- Les opérateurs de la logique floue

Les propositions reliées entre elles par des opérateurs logiques produisent un discours. Ces opérateurs sont: la négation, la conjonction «et», la conjonction «ou». Il y a eu des tentatives (Klir et Yuan, 1995) de représenter les conjonctions par des opérateurs autres que ceux proposés par Zadeh (1965); cependant, ces tentatives n'ont pas été reprises puisque la plupart des applications de la théorie des ensembles flous utilisent les opérateurs originaux.

#### La négation

Il s'agit ici de la négation d'un énoncé ou tout simplement du degré d'appartenance à l'ensemble complémentaire. Du fait que l'univers est un ensemble précis, nous avons déduit précédemment qu'il est toujours vrai qu'un

élément appartient à un ensemble ou à son complément. Cela s'est traduit par l'équation 24 pour un ensemble quelconque A; de là, Zadeh (1965, p.340) déduit que la négation floue est représentée par l'équation suivante:

$$\mu_{A'}(a) = 1 - \mu_A(a) \quad (27)$$

La propriété d'involution est maintenue, c'est-à-dire que si on nie deux fois une affirmation cela équivaut à ne pas la nier du tout. Par exemple, si  $\mu_A(a)=0,3$  alors  $\mu_{A'}(a)=0,7$  et  $\mu_{(A')'}(a)=0,3$ . D'ailleurs, une démonstration peut être faite pour démontrer tous les cas possibles; en effet, il s'agit tout simplement de faire:

$$\mu_{A''}(a) = 1 - \mu_{A'}(a) = 1 - (1 - \mu_A(a)) = \mu_A(a)$$

Il ne faut pas oublier qu'avec un ensemble flou chaque élément peut appartenir à la fois à un ensemble A et à la fois à son complément A'.

#### La conjonction et

Deux propositions simples reliées par une conjonction «et» doivent toutes les deux être vraies pour que l'élément de discours résultant soit vrai. Pour ce cas, Zadeh (1965, p.341) a proposé de prendre le minimum entre les deux degrés d'appartenance.

$$\mu_{A \cap B}(a) = \text{Min}(\mu_A(a), \mu_B(a)) \quad (28)$$

Par exemple, prenons la situation suivante: «le sujet <a> a été diagnostiqué avec un degré de capacité de 0,67 suite à la transformation de son score obtenu à l'épreuve diagnostique». Or, si ce même sujet «a» a obtenu un degré de réussite de 0,25 alors on évaluera la valeur vraie de l'élément de discours: «le sujet a la capacité et a réussi» en opérant avec  $\mu_X(a)=0,67$  et  $\mu_Y(a)=0,25$  :

$$\mu_{X \cap Y}(a) = \text{Min}(\mu_X(a), \mu_Y(a))$$

$$\mu_{X \cap Y}(a) = \text{Min}(0,67, 0,25) = 0,25 \text{ plus faux que vrai}$$

(voir Tableau V)

Selon (Ragin (2000, p.189), l'utilisation du minimum dans l'algèbre floue peut être vu comme une extension des principes de l'algèbre de Boole.

La conjonction ou

À l'inverse, deux propositions simples reliées par une conjonction «ou» doivent toutes les deux être fausses pour que l'élément de discours résultant soit faux. Zadeh (1965, p.340) propose que l'appartenance à l'union de deux ensembles soit le maximum des degrés d'appartenance pour chacun des ensembles.

$$\mu_{A \cup B}(a) = \text{Max}(\mu_A(a), \mu_B(a)) \quad (29)$$

Dans l'exemple précédent, nous pouvons avoir l'élément de discours: «le sujet a la capacité ou a réussi». Or, cet élément de discours est plus ou moins vrai car:

$$\mu_{X \cup Y}(a) = \text{Max}(\mu_X(a), \mu_Y(a))$$

$$\mu_{X \cup Y}(a) = \text{Max}(0,67, 0,25) = 0,67 \quad \textit{plus ou moins vrai}$$

(voir Tableau V)

- Les indicateurs pour décrire les ensembles flous

Le cardinal scalaire est un concept qui permet d'introduire deux indicateurs pour décrire un ensemble flou: le taux d'appartenance et l'entropie.

Le cardinal scalaire

Rappelons-le, le cardinal d'un ensemble ordinaire représente le nombre d'éléments que contient cet ensemble. Pour maintenir le caractère général d'un ensemble flou, il faut donc que le cardinal tienne compte de cette particularité. Klir et Yuan (1995, p.28) ont énoncé que pour un ensemble quelconque A défini dans un univers U, son cardinal scalaire noté |A|, est la somme des degrés d'appartenance pour tous les éléments «a» de l'univers U décrite par la formule suivante:

$$|A| = \sum \mu_A(a) \quad (30)$$

On compte chaque élément en les pondérant de leur degré d'appartenance. Le tableau VI représente un exemple de situation avec dix sujets. Le degré de capacité,  $\mu_X$ , est déduit à partir du modèle vu précédemment à la figure 10, c'est-à-dire avec  $\alpha = 0,2$  (zone d'incertitude) et  $\beta = 65$  (score de césure).

Tableau VI: Un exemple de calcul pour la cardinalité scalaire

Élément	Score (%)	$\mu_x$	$\mu_y$
A	56	0,14	0,30
B	73	0,83	0,50
C	80	0,95	0,90
D	75	0,88	0,80
E	55	0,12	0,90
F	64	0,45	0,60
G	45	0,02	0,40
H	90	0,99	0,80
I	50	0,05	0,45
J	67	0,60	0,30

$$|X| = 5,03 \text{ et } |Y| = 5,95$$

#### Le taux d'appartenance

Un indicateur d'efficacité très utilisé en éducation est le taux d'appartenance. Ce dernier est tout simplement le cardinal scalaire relatif, c'est-à-dire le rapport entre le cardinal scalaire d'un ensemble et le cardinal de l'univers.

Prenons par exemple l'ensemble critère Y de la réussite. On obtient le taux de réussite en effectuant la somme des degrés de réussite pour tous les sujets et en la divisant par le nombre de sujets de la cohorte, soit 10. Avec l'exemple numérique du tableau VI, on a un cardinal scalaire pour Y de 5,95, ce qui nous donne un taux de réussite de 0,595. Ce concept est plus flou que le taux de diplomation largement utilisé dans le milieu scolaire mais par contre, il est plus réaliste. D'ailleurs, pour fournir une information la plus complète possible on est souvent obligé de produire plusieurs taux de diplomation tels la diplomation en 4 ans, la diplomation en 5 ans, etc. Le taux de réussite flou vient condenser cette information dans un seul indicateur.

L'entropie

Afin de décrire la distribution des sujets pour un ensemble donné, Kosko (1993, p.133) propose de définir le concept d'entropie  $E(A)$  et suggère comme indicateur (voir équation 31) le rapport des expressions pour la loi du tiers exclus.

$$E(A) = \frac{|A \cap A'|}{|A \cup A'|} = \frac{\sum \text{Min}(\mu_A, 1-\mu_A)}{\sum \text{Max}(\mu_A, 1-\mu_A)} \quad (31)$$

Cet indicateur nous informe jusqu'à quel degré l'ensemble  $A$  est flou et nous pouvons retrouver ses limites en tenant compte des extrêmes. D'une part, si l'ensemble  $A$  est un ensemble ordinaire (cas 1 dans la figure 11), alors la loi du tiers exclus s'applique et le numérateur deviendra nul; ceci donne une entropie de 0; autrement dit, cette valeur signifie que nous sommes certains que tous les éléments «a» de l'univers sont classés soit dans  $A$  ou dans  $A'$ . D'autre part, si nous avons pour tous les éléments de l'univers le maximum d'incertitude quant au classement dans un ensemble quelconque  $A$  (cas 3 dans la figure 11), alors leur degré d'appartenance vaudrait 0,5 et le numérateur de l'expression 31 sera égal au dénominateur; ceci donne une entropie de 1. Aussi, nous pouvons utiliser l'entropie pour indiquer l'incertitude envers les diagnostics posés pour un groupe de sujets.

Le tableau VII reprend l'exemple du tableau VI pour l'ensemble  $X$  mais dans le but de calculer l'entropie.

Tableau VII: Un exemple de calcul de l'entropie

Élément	$\mu_X$	$\mu_{X \cap X'}$	$\mu_{X \cup X'}$
A	0,14	0,14	0,86
B	0,83	0,17	0,83
C	0,95	0,05	0,95
D	0,88	0,12	0,88
E	0,12	0,12	0,88
F	0,45	0,45	0,55
G	0,02	0,02	0,98
H	0,99	0,01	0,99
I	0,05	0,05	0,95
J	0,60	0,40	0,60

$$E(X) = \frac{1,53}{8,47} = 0,18$$

Une distribution en U où nous pouvons distinguer une séparation entre les sujets forts et les sujets faibles donne une entropie faible et tout près de 0. À l'inverse, une distribution normale centrée à 0,5 et avec une faible dispersion amène une entropie forte et tout près de 1. Conséquemment, toute autre distribution située entre ces deux extrêmes génère une entropie située entre 0 et 1. Le résultat obtenu de 0,18 laisse entrevoir une distribution légèrement en U ; en effet, une distribution uniforme aurait une entropie de 0,33 (voir Annexe 1 pour démonstration).

L'entropie et le taux d'appartenance représentent deux indicateurs qui vont nous être utiles pour décrire un ensemble flou; ils sont d'autant plus pertinents si cet ensemble contient beaucoup d'éléments. En fait, le taux remplace la moyenne et l'entropie l'écart-type; les deux statistiques que nous avons avec la TCT.

- L'inclusion et l'implication logique

Pour établir la validité pronostique de l'épreuve de capacité en mathématique, il faut prédire la réussite à l'aide d'une règle de prédiction par raisonnement. Or, cette règle de prédiction prend la forme d'une implication logique.

L'implication logique est reliée directement au fait qu'un ensemble soit inclus dans un autre. En effet, si un élément appartient à l'ensemble X et que cet ensemble X est inclus dans l'ensemble Y, alors l'appartenance de cet élément à l'ensemble X implique qu'il appartienne aussi à l'ensemble Y. L'implication représente une relation de cause-à-effet entre deux variables et ressemble au coefficient de corrélation que nous avons entre deux variables mesurées sur une échelle d'intervalle avec la TCT, toutefois le coefficient de corrélation indique le degré de relation sans préciser qu'il y a un lien de cause-à-effet.

Klir et Yuan (1995, p.28) définissent le degré d'inclusion du sous-ensemble A dans l'ensemble B par la formule de l'équation 32. Smithson (2005, p.443) ajoute : «L'idée générale est d'évaluer les degrés d'appartenance de x dans l'intersection des ensembles flous A et B (...). Ces degrés d'appartenance sont alors comparés avec les degrés d'appartenance pour l'ensemble qui est inclus.»<sup>7</sup>

$$\|\mu_A \Rightarrow \mu_B\| \Leftrightarrow \|A \subseteq B\| = \frac{|A \cap B|}{|A|} \quad (32)$$

Le symbole  $\|\text{expression}\|$  veut dire la valeur vraie pour l'expression; de ce fait, l'expression  $\|A \subseteq B\|$  signifie que la valeur vraie de l'inclusion de A dans B est équivalente à la valeur vraie que  $\mu_A$  implique  $\mu_B$ . L'équation 32 vient annoncer que si le cardinal scalaire de l'intersection des deux ensembles A et B est égal au cardinal scalaire de l'ensemble A, alors ce dernier est totalement inclus dans l'ensemble B et l'évaluation de l'expression donne 1. En ce cas-là il est

---

<sup>7</sup> Traduction libre de : *The general idea is to evaluate the membership of x in the intersection of fuzzy sets A and B (.). These membership values are then compared with membership in the including set.*

certainement vrai que l'appartenance d'un élément à l'ensemble A implique qu'il appartienne aussi à l'ensemble B; ceci se note par:  $\|\mu_A \Rightarrow \mu_B\|$  et le symbole  $\Leftrightarrow$  signifie que la relation de cause-à-effet qu'est l'implication logique est équivalente à l'inclusion entre deux ensembles.

À l'opposé, si chaque élément appartenant à A n'appartient pas du tout à B, alors l'implication est inverse et la valeur de  $\|\mu_A \Rightarrow \mu_B\|$  donne 0. Par contre, la valeur de  $\|\mu_A \Rightarrow \mu_B\|$  donne 1. Si les éléments de A appartiennent autant à l'ensemble B qu'à son complémentaire B'; autrement dit, chaque élément de A a une appartenance de 0,5 à l'ensemble B, en ce cas-là l'implication donne aussi une valeur vraie de 0,5 et les deux attributs sont complètement indépendants les uns les autres; en logique, on note cela par l'expression  $\mu_A \perp \mu_B$ .

- Interprétation de la valeur vraie d'une implication logique

La valeur vraie d'une implication logique indique la possibilité pour un élément que le degré d'appartenance à l'ensemble impliqué B soit supérieur ou égal au degré d'appartenance à l'ensemble A impliquant. D'ailleurs, Ragin (2000, p.214) va dans le même sens lorsqu'il affirme qu'un ensemble A est un sous-ensemble de B si les degrés d'appartenance des cas dans l'ensemble A sont inférieurs ou égaux aux degrés d'appartenance pour les mêmes cas respectifs dans l'ensemble B. Or, il calcule la valeur vraie de l'implication logique en effectuant le rapport entre le nombre de cas où cette dernière règle est respectée et le nombre de cas concernés.

Pour notre part, nous appliquerons plutôt l'équation 32 pour trouver la valeur vraie de l'implication logique. Pour Smithson (2005, p.443), l'opérateur minimum ne requiert pas de conditions strictes d'utilisation d'une échelle d'intervalle et de ce fait, favorise beaucoup d'applications en sciences sociales. Le tableau VIII reprend l'exemple du tableau VI pour une implication en logique floue.

**Tableau VIII: Exemple de la valeur vraie d'une implication logique**

Élément	$\mu_X$	$\mu_Y$	$\mu_{X \cap Y}$
a	0,14	0,30	0,14
b	0,83	0,50	0,50
c	0,95	0,90	0,90
d	0,88	0,80	0,80
e	0,12	0,90	0,12
f	0,45	0,60	0,45
g	0,02	0,40	0,02
h	0,99	0,80	0,80
i	0,05	0,45	0,05
j	0,60	0,30	0,30

$$|X| = 5,03 \text{ et } |X \cap Y| = 4,08$$

$$\text{Alors: } \|\mu_X \Rightarrow \mu_Y\| \Leftrightarrow \|X \subseteq Y\| = 4,08 / 5,03 = 0,81$$

D'une part, si on applique les prédicats proposés par Huang (1999, p.106) que l'on retrouve au tableau IV à la valeur vraie de 0,81 trouvée, il faut interpréter le résultat comme étant possiblement vrai que l'appartenance à X implique l'appartenance à Y ou en d'autres mots, qu'un diagnostic positif à l'épreuve entraîne possiblement la réussite. D'autre part, si on applique les prédicats proposés par Ragin (2000) que l'on retrouve au tableau V à la même valeur vraie de 0,81 trouvée, il faut plutôt interpréter le résultat comme étant plus vrai que faux qu'un diagnostic positif implique la réussite.

Également, on peut remarquer la ressemblance entre l'équation 32 pour l'implication logique et l'équation 22 pour la probabilité bayésienne *a posteriori*. Cependant, dans le cas d'une probabilité, il s'agit d'événements alors qu'ici il s'agit plutôt d'appartenance à un ensemble.

Si Zadeh (1965, p.340) affirme que la notion d'ensemble flou est complètement non- statistique par nature, Ragin (2000) avance plutôt qu'il est possible d'inférer le résultat à une population en effectuant un test statistique sur les proportions. En effet, on peut considérer l'équation 32 comme étant une proportion d'appartenance de l'ensemble X à l'intersection de X et Y.

Au préalable, il (Ragin, 2000 ; p.111) suggère d'établir un repère (en anglais «bench-mark») associé à un prédicat. Par la suite, on applique un test unilatéral de proportions en calculant, comme à l'équation 33, la variable z pour une distribution normale centrée réduite.

$$z = \frac{(P - p) - 1/2N}{\sqrt{(pq/N)}} \quad (33)$$

- Où P : proportion observée en appliquant l'équation 32 ;  
 p : repère ou «bench-mark» correspondant à un prédicat ;  
 q : complément du repère (1-p) ;  
 N : nombre de cas.

L'utilisation de cette formule n'est valide cependant que s'il y a plus de 30 cas. Pour un seuil de signification à 95% avec un test unilatéral, c'est-à-dire avec 19 chances sur 20 de dire vrai, il faut que la valeur calculée de z dans l'équation 33 soit supérieure ou égale à 1,65. Par exemple, pour un groupe de 100 cas avec un repère de 0,75 et une valeur vraie observée de 0,81, nous obtenons une valeur de 1,27 pour z ; par conséquent, la règle causale n'est pas significative car la valeur de z calculée est inférieure à 1,65. Alors qu'avec un repère de 0,67 et la même valeur vraie de 0,81, nous obtenons une valeur de 2,87 pour z, ainsi la règle causale devient significative. D'une façon générale, il est justifié d'affirmer qu'il est plus ou moins vrai que si le sujet a la capacité alors il réussira.

Aussi, plus il y aura de cas à analyser, plus il y aura de chances de rendre significatif un repère qui sera le plus près possible de la valeur vraie observée. À l'inverse, un petit nombre de cas oblige à rendre plus flou le prédicat.

### ***9. Application de la théorie des ensembles flous à l'élaboration d'un modèle prédictif de la réussite***

Alors que la prédiction se fait par conjecture avec les différentes théories des probabilités, elle se fera par raisonnement avec la théorie des ensembles flous. Les auteurs Flood et Carson (1988, p.209) précisent la séquence des opérations pour obtenir une prédiction qualitative: la mesure, la description par propositions, la description par la logique, l'explication qualitative et enfin la prédiction qualitative. Nous avons vu jusqu'à maintenant comment décrire le diagnostic cognitif et la réussite sous la forme de propositions et d'expressions logiques. L'analyse de la nécessité et de la suffisance va permettre maintenant d'élaborer un moteur d'inférence composé de règles causales afin de prédire la réussite.

- ***L'analyse de la nécessité et de la suffisance***

À prime abord, nous supposons qu'un diagnostic positif est nécessaire pour prédire la réussite mais il est possible que d'autres attributs des sujets soient aussi nécessaires. Ragin (2000, p.272) affirme qu'il faut se questionner pour savoir si l'ensemble critère Y est inclus dans un ou plusieurs ensembles associés aux conditions causales pour déterminer si ces causes sont nécessaires. Un attribut causal sera nécessaire si son absence entraîne la non appartenance à l'attribut critère. En résumé, un attribut nécessaire effectuera une conjonction «et» avec les autres attributs pour former une combinaison causale. Ainsi, nous pouvons dire qu'avoir suivi un cours d'appoint en mathématique est un attribut nécessaire pour réussir si le sujet n'a pas la capacité alors que le fait d'être de sexe féminin n'est en aucun temps un attribut nécessaire.

Ragin (2000, p.245) spécifie ainsi la procédure pour effectuer l'analyse de la suffisance. En premier lieu, il faut déterminer les expressions causales à être testées. Il ajoute également que le nombre d'expressions causales possibles est de  $3^k - 1$ , où k est le nombre d'attributs causals déterminés lors de l'analyse de nécessité. Par la suite, il faut calculer les degrés d'appartenance pour chaque cas

et les comparer au degré d'appartenance du critère. Pour nous, cette dernière étape sera modifiée pour calculer plutôt l'équation 32.

Ainsi, pour deux attributs causals  $X$  et  $M$ , il faudrait tester 8 expressions possibles ( $3^2-1$ ), soient :  $X' \cap M'$ ,  $X' \cap M$ ,  $X \cap M'$ ,  $X \cap M$ ,  $X'$ ,  $X$ ,  $M'$ ,  $M$ . Plus spécifiquement, il faut vérifier si chacune de ces expressions est incluse dans l'ensemble critère  $Y$ . Un attribut suffisant effectue une conjonction «ou» avec les autres expressions et sa présence entraîne l'appartenance à l'ensemble critère.

En bref, pour analyser la suffisance, il faut au préalable évaluer le degré d'appartenance de chaque cas et ce, pour chaque attribut. Par la suite, on dresse un bilan des appartenances à des intersections d'ensembles en prenant le minimum des degrés d'appartenance aux ensembles impliqués. Ainsi, le tableau IX représente la structure d'un bilan des appartenances pour une analyse de suffisance dans un espace de propriétés à trois attributs et avec un critère positif.

**Tableau IX: Le bilan des appartenances**

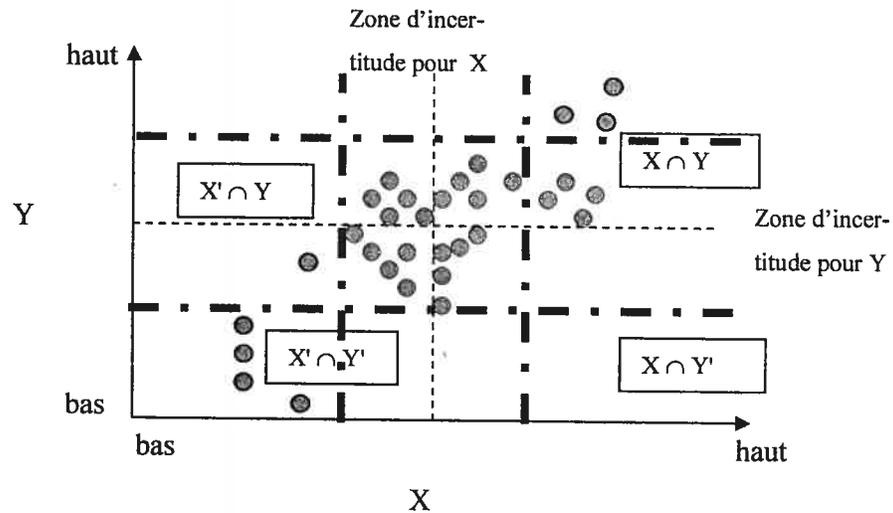
No.	Expression causale	valeur du cardinal (1)	Intersection de la cause et de l'effet	Valeur du cardinal (2)	valeur vraie de l'inclusion (2/1)
1	$X' \cap M'$		$X' \cap M' \cap Y$		
2	$X' \cap M$		$X' \cap M \cap Y$		
3	$X \cap M'$		$X \cap M' \cap Y$		
4	$X \cap M$		$X \cap M \cap Y$		
5	$X'$		$X' \cap Y$		
6	$X$		$X \cap Y$		
7	$M'$		$M' \cap Y$		
8	$M$		$M \cap Y$		

Smithson (2005, p.448) privilégie d'établir au départ un repère pour la valeur vraie de l'inclusion avant de traiter les données. En choisissant au préalable un seuil de signification statistique, nous pourrions établir quelles sont les implications logiques qui ont une valeur vraie significative. Nous dirons alors que l'expression causale associée à chacune de ces implications est suffisante.

- Les fonctions d'appartenance et la valeur vraie pour l'implication logique

Comme nous l'avons vu précédemment, le degré de capacité,  $\mu_x$ , sera déterminé par l'application du modèle au score brut obtenu à l'épreuve diagnostique. Nous avons alors une zone de diagnostic positif où le degré de capacité est tout près de 1.

D'autre part, si la fonction d'appartenance au critère de réussite n'épouse pas la même forme que celle de la capacité, alors nous aurons un biais dans le calcul de la valeur vraie pour l'implication logique. Par exemple, si la fonction d'appartenance du critère ne prévoit pas de zone de certitude, alors nous aurons certainement plusieurs cas où le degré de capacité sera supérieur au degré de réussite,  $\mu_x > \mu_y$ . Cependant, pour trouver la valeur vraie de l'implication logique, l'équation 32 démontre que le minimum entre le degré de capacité et le degré de réussite se confond avec le degré de capacité le plus souvent possible si on veut obtenir une valeur vraie qui tend vers 1. Et le graphique de la figure 13 illustre un exemple de relation entre le score à l'épreuve diagnostique et le score critère.



**Figure 13 : Les zones d'incertitude pour la cause X et l'effet Y**

En l'absence de cette restriction pour la zone d'incertitude de Y, nous aurons une sous-estimation de la valeur vraie de l'implication logique. Par conséquent, si la fonction d'appartenance du critère Y a une zone d'incertitude trop grande, le cardinal de la partition  $X \cap Y$  sera moindre.

- Le moteur d'inférence

Une implication logique est comme une règle. Selon Schneider et Kandel (1992, p.34), une règle comprend deux parties: la prémisse et la conclusion. Et sa forme générale est: *Si la prémisse était vraie alors la conclusion serait vraie aussi.*

Or, la prémisse est souvent complexe et comprend la plupart du temps plusieurs ensembles. Pour notre recherche, l'ensemble X représentera l'ensemble flou des sujets qui ont été diagnostiqués positivement, c'est-à-dire qu'ils ont la capacité en mathématique pour entreprendre des études d'ingénieur. Il y aura aussi d'autres attributs dont il faudra tenir compte et qui nous donnera d'autres ensembles faisant parties des expressions causales. La prémisse d'une règle sera pour nous associée à l'union des expressions causales qui ont été déclarées suffisantes.

L'ensemble Y représente l'ensemble flou des sujets ayant réussi ; il s'agit pour nous du critère ou de la conclusion d'une règle. À partir des données empiriques du bilan des appartenances, il sera possible d'inférer des règles d'implication logique, d'où le nom de «règle d'inférence». L'expression «moteur d'inférence» désigne la partie d'un système expert qui regroupe toutes les règles d'inférence ; on peut aussi utiliser l'expression «base de connaissances» qui représente la même entité. Dans notre recherche, nous avons deux règles d'inférence possibles : l'une avec une conclusion positive (Y) et l'autre avec une conclusion négative (Y').

Toutefois, il se peut que deux règles donne une contradiction. Selon Ragin (1987), une contradiction existe lorsqu'une même prémisse entraîne une conclusion et son contraire.

Prenons un exemple pour mieux illustrer cette notion de moteur d'inférence. Supposons qu'une valeur vraie significative de l'inclusion a été obtenue, pour un repère de 0,67, avec les expressions causales 4 et 6 du tableau IX pour un critère positif et avec l'expression 1 pour le critère négatif. Sous la forme algébrique, le moteur d'inférence serait :

Règles :

$$1 : (\mu_X \wedge \mu_M) \vee \mu_X \rightarrow \mu_Y$$

$$2 : \neg\mu_X \wedge \neg\mu_M \rightarrow \neg\mu_Y$$

Et les symboles  $\neg$ ,  $\wedge$ ,  $\vee$  représentent respectivement les opérateurs logiques de négation, de conjonction «et» et de disjonction «ou». Il n'y a pas ici de contradiction car les prémisses des deux règles ne sont pas identiques. Cependant, l'attribut X de la règle 1 serait à la fois nécessaire et suffisant pour que l'attribut M ait un effet dans la prédiction du critère, donc il y aurait possibilité de simplifier cette même règle pour ne s'en tenir qu'à X:

Règle :

$$1 : \mu_X \rightarrow \mu_Y$$

En utilisant les prédicats de Ragin (voir tableau V), la forme verbale du moteur d'inférence sera représentée par:

Règles :

1. *Si le sujet avait été diagnostiqué positivement à l'épreuve de math alors il serait plus ou moins vrai qu'il puisse réussir.*
2. *Si le sujet avait été diagnostiqué négativement à l'épreuve de math et s'il avait échoué au cours de mise à niveau alors il serait plus ou moins vrai qu'il puisse échouer.*

À noter que si la valeur vraie d'une inclusion est nettement supérieure au repère, il sera pertinent de tester la signification avec un repère plus élevé. Ragin (2000, p.284) affirme que la suffisance d'une cause c'est lorsqu'elle peut expliquer à elle seule l'effet. Étant donné qu'une cause peut être l'intersection de l'attribut diagnostic avec des attributs intervenants, on contrôle ainsi ces derniers lors de la détermination du pronostic. C'est un peu comme l'analyse de régression multiple qui cherche à expliquer la variable dépendante à partir du rapport entre sa variance et les variances pour plusieurs variables indépendantes mais aussi en tenant compte des covariances entre toutes les variables. Par contre, avec l'analyse causale que permet le moteur d'inférence, on n'a pas besoin d'assumer que les variables sont mesurées sur une échelle d'intervalle.

- *Le moteur d'inférence pour l'épreuve diagnostique*

Pour mieux comprendre le moteur d'inférence comme modèle prédictif, revenons à notre comparaison d'une situation en médecine. Nous disions qu'un taux de cholestérol élevé était un pronostic pour un accident cardio-vasculaire. Cependant, le médecin tiendra compte aussi de d'autres facteurs avant d'établir son diagnostic comme, le genre, les habitudes de vie du patient (fumeur ou non-fumeur), le traitement, etc.

Dans notre recherche, le score à l'épreuve diagnostique agit un peu comme le taux de cholestérol et les autres facteurs se confondent avec les attributs tels le genre et le traitement. Toutefois, il faut vérifier si l'information provenant de l'épreuve diagnostique est significative pour prédire la réussite. Alors on pourra

déterminer la validité prédictive de cet instrument s'il y a une valeur vraie d'implication logique d'au moins 0,75 (plus vrai que faux) entre la cause principale et l'effet. D'une manière plus pratique, on se demandera si l'ensemble des sujets diagnostiqués positivement ont réussi et on se questionnera aussi sur le corollaire, c'est-à-dire si l'ensemble des sujets diagnostiqués négativement ont échoué. Ce questionnement peut se faire par les relations d'inclusion 34.

$$\begin{aligned} \|X \subseteq Y\| &\geq 0,75 & (34) \\ \|X' \subseteq Y'\| &\geq 0,75 \end{aligned}$$

Ainsi la validité pronostique de l'épreuve peut être vérifiée par rapport à la réussite d'un cours de première année mais aussi par rapport à la réussite dans un programme; la réussite d'un cours de première année étant à son tour un pronostic de la réussite dans le programme. Selon Zwick (2002, p.91), la corrélation entre le score au test d'admission SAT et le rendement dans un cours ne varie pas significativement si ce cours se situe en première année ou dans les années subséquentes du programme. Cependant, rappelons-nous que cette corrélation est déjà très basse et se situe entre 0,35 et 0,45.

L'avantage d'avoir un moteur d'inférence avec des règles en logique floue pour prédire la réussite est que chacune des règles peut être traduite sous une forme verbale dans un langage en sens commun. De plus, une inspection rapide des données peut confirmer la validité des règles.

Ce moteur d'inférence peut faciliter le travail de l'éducateur qui est responsable du plan de réussite surtout en ce qui concerne les décisions. À partir des caractéristiques d'un sujet et de son score à l'épreuve diagnostique, il pourra décider de l'encadrement à donner de manière à assurer la réussite du sujet. Une telle décision doit être prise à partir d'un diagnostic. Lossin (2005) mentionne : «un plus grand degré de flou coïncide avec un plus grand degré d'indécision.»<sup>8</sup> Autrement dit, si le diagnostic est flou, la décision devra l'être aussi.

---

<sup>8</sup> Traduction libre de : *a higher degree of fuzziness coincides with a higher degree of indecision.*

En résumé, la théorie des ensembles flous permet de faire une interprétation critériée d'un score à un test qui n'a pas une fidélité parfaite en proposant un modèle épousant la forme d'une fonction logistique. Avec ce modèle, on traduit les scores en degré d'appartenance à un ensemble flou  $X$  qui constitue le fondement pour établir un diagnostic. De plus, la valeur vraie de l'inclusion de l'ensemble diagnostic  $X$  dans l'ensemble critère  $Y$  est un indicateur de la validité pronostique de l'instrument et indirectement du modèle pronostique qui y est rattaché pour interpréter les scores.

## Chapitre troisième

### La méthodologie

Dans le chapitre précédent, nous avons énuméré les raisons de la pertinence de la théorie des ensembles flous pour élaborer un modèle servant à prédire la réussite. Dans cette partie, nous verrons la manière d'élaborer un modèle s'inspirant de la théorie des ensembles flous et utilisant les données de la présente recherche. Nous expliquerons dans un premier temps nos choix méthodologiques et nous parlerons de la provenance des données avant d'aborder les hypothèses de recherche qui découlent du cadre conceptuel. Enfin, nous expliquerons les méthodes d'élaboration et de validation du modèle en spécifiant les limites à la généralisation des résultats.

#### *1. Le choix méthodologique*

Rappelons-le, l'objectif de cette recherche est d'élaborer un modèle pour prédire adéquatement la réussite des sujets dans leur programme. Or, Flood et Carson (1990, p.208) nous proposent une méthodologie pour la modélisation. Plus précisément, ils évaluent les différentes méthodes d'élaboration des modèles selon l'intention visée par la recherche, à savoir: la description, la prédiction ou l'explication. Notre intention est de viser la prédiction. Et la méthode qu'on choisira devra aussi permettre de généraliser les résultats.

- *Une méthode mixte*

Selon ces mêmes auteurs, les méthodes statistiques ou quantitatives constituent un choix judicieux dans une situation de recherche où l'intention est la prédiction; cependant, il faut que les conditions expérimentales soient maintenues stables. Avec ces méthodes, une hypothèse est déduite d'une théorie que les résultats de l'expérience viennent confirmer ou infirmer et les conclusions de la recherche peuvent être généralisées à une population entière. Toutefois, en éducation et dans les sciences sociales, la pratique fait en sorte que bien souvent les conditions expérimentales ne peuvent être satisfaites.

Les méthodes logiques ou qualitatives sont aussi un choix judicieux pour une situation de prédiction selon Flood et Carson (1990, p.208) car elles constituent une approche pratique et fondée sur l'étude de quelques cas. Or, ces méthodes logiques ou qualitatives ne mettent en cause que peu de sujets et pose le problème de la généralisation des résultats. Autrement dit, nous pouvons expliquer la réussite de quelques sujets mais nous ne pouvons pas reproduire tous les cas possibles. D'ailleurs, selon Laperrière (1997, p.383): «...la recherche de reproductibilité n'a pas vraiment de sens dans les méthodes qualitatives».

Miles et Huberman (1991) parlent plutôt d'une approche mixte sans pour cela la définir. Johnson et Onwuegbuzie (2004) vont plus loin et propose un nouveau paradigme qui mixe les deux méthodes dans une recherche. Pour Ragin (2000, p.63), une méthode quantitative favorise l'étude en largeur d'un problème avec quelques variables sur plusieurs cas alors qu'une méthode qualitative se centre sur l'étude en profondeur de plusieurs variables avec quelques cas. Il propose plutôt une recherche orientée vers la diversité qu'il définit sous deux aspects (Ragin, 2000 ; p.149) :

*«Le premier aspect de la diversité est capté par la notion de différences entre catégories (.). Le second aspect de la diversité est capté par la notion de degré d'appartenance.»<sup>9</sup>*

Normalement, une donnée qualitative est classée dans une catégorie alors qu'une donnée quantitative telle un score ne peut l'être directement. Toutefois, la notion de catégorie réfère à un ensemble, la notion de degré d'appartenance vient préciser qu'il s'agit d'un ensemble flou puisqu'on ajoute avec quel degré de certitude une donnée appartient à une catégorie. Bassey (1998) ajoute que la théorie des ensembles flous (TEF) suggère une voie pour représenter la connaissance générée par la recherche empirique. En effet, il est possible d'appliquer une donnée quantitative à un modèle pour la traduire en degré

---

<sup>9</sup> Traduction libre de : *The first aspect of diversity is captured by the notion of differences in kind(.). The second aspect of diversity is captured by the notion of degree of memberships.*

d'appartenance avec la TEF et ainsi de traiter autant les données qualitatives que quantitatives dans une recherche orientée vers la diversité. Une généralisation floue remplace alors la certitude de la généralisation probabiliste (il est vrai que...dans 95 % des cas) par l'incertitude d'un prédicat (il est parfois vrai que...). Ainsi, on retire l'avantage de l'étude en largeur que l'on avait avec une méthode quantitative et de l'étude en profondeur avec une méthode qualitative.

- La généralisation

Les méthodes d'analyse autant avec des données quantitatives que qualitatives utilisent l'induction, c'est-à-dire qu'à partir d'un ensemble de faits particuliers on énonce un principe général. Avec une méthode quantitative, on mesure quelques variables auprès d'un échantillon nombreux et la relation obtenue peut être inférée à une population. À l'opposé, avec une méthode qualitative on met l'accent sur la quantité d'information recueillie par l'observation de plusieurs variables non métriques pour établir des relations entre elles. Cependant, dans ce dernier cas il n'existe pas de modèles de distribution préalables comme pour une méthode quantitative où l'on assume un modèle de distribution normale et la généralisation est presque impossible.

Rappelons notre point de vue épistémologique, et celui du philosophe Karl Popper (Debays, 1998 ; p.7) qui propose : «de renverser l'inductivisme et son critère vérificationniste.» Pour Popper (Debays, 1998 ; p.8) :

*«La science authentique se distingue également des autres activités humaines (.) par sa recherche constante de la vérité (.). La vérité doit être entendue au sens de correspondance avec les faits.»*

Avec la théorie des ensembles flous (TEF), on est aussi dans une recherche constante de la vérité en indiquant à quel degré les faits correspondent à la réalité. Plus particulièrement, il s'agit d'établir un modèle qui fait correspondre le score d'un sujet à l'épreuve de mathématique au diagnostic posé. Ce diagnostic sera vrai si le pronostic sur la réussite de ce même sujet est également vrai mais il peut aussi n'être que partiellement vrai. La généralisation devient possible seulement si on applique ce modèle avec des données d'une autre

cohorte. De plus, avec la pensée de Popper, la généralisation est temporelle en ce sens que le modèle peut être ajusté à travers le temps d'une cohorte à l'autre.

## *2. La provenance des données*

L'ensemble des étudiants qui fréquentent les programmes de génie de l'École Polytechnique de Montréal sera la population étudiée dans la présente recherche. La cohorte de 1997 constituera l'échantillon principal ; plus spécifiquement, seuls seront considérés les sujets qui auront passé l'épreuve de mathématique de juin 1997. L'épreuve diagnostique en mathématique est obligatoire pour les quelques centaines d'étudiants admis chaque année à l'École Polytechnique et il y a trois séances de passation; toutefois, c'est à la séance de juin où l'on retrouve le plus grand nombre de sujets. Quelques uns d'entre eux suivront librement un cours d'appoint en mathématique et ceci pourra avoir un effet déstabilisateur pour le pronostic de la réussite.

### *• Confidentialité, disponibilité et traitement des données*

Des chercheurs de la Faculté des sciences de l'éducation de l'Université de Montréal ont demandé et obtenu de la Commission d'accès à l'information la permission d'utiliser les données relatives aux sujets de plusieurs cohortes qui ont passé les épreuves de mathématique administrées lors de l'admission. Bien entendu, cette permission est conditionnelle à la conservation du caractère confidentiel de ces données. Par conséquent, les données ne seront en aucun temps identifiées ni par le nom, ni par le matricule de la personne à qui elles sont rattachées.

Une série de rencontres eut lieu entre les chercheurs intéressés par ces données et des personnes-ressources de l'École polytechnique. Ces rencontres avaient pour but de clarifier la requête d'utilisation des données qui avait été autorisée par la Commission. Aussi, un devis fut transmis en juin 2005 aux autorités compétentes de l'École Polytechnique (voir Annexe II). Il faut dire qu'auparavant les données relatives aux épreuves furent disponibles vers le début de 2004.

Les données relatives à la cohorte de 1997 furent transmises par M. Jean-Guy Deschesnes (consultant pour les données institutionnelles de l'École Polytechnique) en date du 18 novembre 2005 par un fichier Excel attaché à un courrier électronique et les données relatives à la cohorte de 1998 nous sont parvenues sous le même pli en date du 23 novembre de la même année 2005. Par contre, il n'y avait aucune donnée relativement à l'inscription des sujets à un cours d'appoint en mathématique et il n'y avait pas non plus de données relativement à l'épreuve diagnostique de juin 1998.

Devant cet état de fait, une nouvelle requête fut acheminée pour obtenir les données sur l'épreuve de juin 1999 et de la cohorte de l'automne 1999. Ces données furent reçues en janvier 2006.

Donc, pour chaque sujet qui a passé l'épreuve de mathématique en juin 1997, on aura comme attributs ses réponses à chacun des 60 items. En plus, on aura son inscription à chaque trimestre du programme, la date d'obtention de son diplôme s'il y a lieu, sa cote pour le cours «Calcul I», et le fait qu'il détienne un DEC en sciences de la nature ou non.

Toutes les opérations mathématiques nécessaires à l'élaboration du modèle pronostique seront effectuées avec le logiciel Excel.

- Les observations

Pour cueillir toutes ces données, on effectuera des observations à différents moments du processus.

Premièrement, l'observation  $O_1$  représentera les scores obtenus selon le modèle de mesure de la TCT à l'épreuve de mathématique pour donner la variable pronostique  $X$ . Cette épreuve est subdivisée en six domaines, soient : les fonctions élémentaires, la géométrie, la trigonométrie, les matrices et vecteurs, le calcul différentiel et le calcul intégral. Mais nous utiliserons le score total comme donnée à être traitée par le modèle pour obtenir le degré d'appartenance

à l'ensemble  $X$ . La zone d'incertitude de ce modèle viendra compenser une fidélité imparfaite pour cette épreuve.

Deuxièmement, l'observation  $O_{2c}$  correspondra aux cotes pour le cours de «Calcul I», cette observation inférera la variable prédite  $Y$  associée à la réussite dans un cours. En plus, ce cours est offert à l'ensemble des étudiants au premier trimestre pour le cheminement classique de la plupart des programmes. Le contenu de ce cours est la suite du cours de calcul différentiel et intégral abordé au niveau collégial et il serait plausible qu'un sujet ayant été diagnostiqué positif à l'épreuve diagnostique puisse le réussir. Cependant, la fidélité de cette donnée n'est pas assurée car elle est établie à partir d'appréciations des enseignants.

Troisièmement, l'observation  $O_{2p}$  sur les données d'inscription et de diplomation des étudiants s'échelonne tout au long du curriculum et même jusqu'en juin 2003 afin de pouvoir constituer la variable prédite  $Y$  de la réussite dans un programme. Nous utiliserons les prédicats qualitatifs de Ragin (2000) décrits par le tableau V dans la section 7 du cadre conceptuel pour traduire la donnée en un degré d'appartenance à l'ensemble flou  $Y$ .

Quatrièmement, l'observation  $O_3$  constituera le score à l'épreuve diagnostique de juin 1999. On appliquera cette donnée pour chaque sujet au modèle pronostique élaboré avec les données de la cohorte de 1997.

Cinquièmement, l'observation  $O_4$  représentera la cote au cours «Calcul I» pour la cohorte de 1999. Toutes ces données devraient servir à valider le modèle par le fait que les pronostics devraient être constamment exacts.

- Les attributs intervenants

D'autres attributs peuvent être nécessaires pour établir le pronostic sur la réussite d'un sujet. Lorsque ces attributs constituent des variables qualitatives, on les représente par des ensembles ordinaires, sinon, on les associe à des ensembles flous.

Dans une étude sur le cheminement des étudiants de la cohorte 2001, le ministère de l'Éducation du Québec (MEQ, 2004-2 ; p.23) a établi que 72% des nouveaux inscrits possédaient un diplôme d'études collégiales (DEC), 21% étaient citoyens canadiens sans DEC et 7 % des étrangers. Le taux de persévérance après un an était de 91,8% pour ceux qui avaient obtenu récemment un DEC, 79,6% pour les citoyens canadiens sans DEC et de 81,6% pour les étudiants étrangers. Le diplôme qui appuie l'admission du sujet constitue donc un attribut significatif. D'ailleurs, le dépliant explicatif de la structure des programmes de l'École Polytechnique de Montréal (2004) mentionne:

*«Pour entreprendre des études à l'École Polytechnique, l'ingénieur en herbe doit posséder certaines qualités. Il doit d'abord et avant tout démontrer une aptitude et un goût marqués pour les mathématiques et les sciences, ces disciplines constituent le fondement des programmes de génie.»*

L'ensemble ordinaire D des sujets détenant un diplôme du collégial (DEC) en sciences de la nature sera considéré pour la cohorte de 1997. Comme il n'y avait pas de données, pour la cohorte de l'automne 1999 sur l'attribut D, nous avons défini un nouvel attribut désigné par C qui indique si le sujet provenait d'un collège public québécois (cégep) ou non.

Pour prédire la réussite dans un programme de génie de l'École Polytechnique de Montréal, un attribut supplémentaire sera pris en compte : la nature du programme. En effet, il serait très plausible que le taux de diplomation du programme de génie électrique soit plus faible que celui du programme de génie mécanique ; les notions étant plus abstraites en électricité qu'en mécanique. Un ensemble flou P sera utilisé pour représenter cet attribut et sa fonction d'appartenance sera déterminée à partir des taux de diplomation calculés à partir des données.

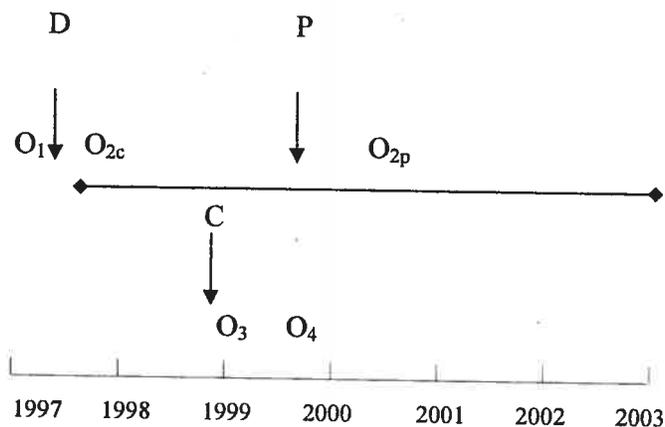
Mais il y a d'autres attributs pour les sujets qu'on estime superflus pour établir un pronostic de réussite. Ainsi, nous n'avons pas tenu compte de l'attribut «genre» puisque dans le cadre d'une étude sur la réussite scolaire des garçons, le MEQ (2004-3 ; p.23) affirmait:

*«Par ailleurs, il ressort clairement de l'ensemble des études consultées que les difficultés scolaires des garçons se manifestent principalement sur le plan de la langue d'enseignement.»*

La maîtrise de la langue d'enseignement a un impact beaucoup moins important dans des études en génie, l'accent est surtout mis sur les mathématiques et les sciences.

- Le plan d'analyse

Le plan d'analyse de la figure 14 vient résumer le déroulement de la recherche. Ce plan intègre les attributs intervenants D et C en plus de l'attribut P qui constitue la particularité du programme de génie en regard de la réussite académique.



**Figure 14: Le plan d'analyse**

Plusieurs cours de l'École Polytechnique ont été modifiés depuis quelques années comme ce cours «Calcul I» qui nous intéresse particulièrement. Le sigle

de ce cours a été changé mais une grille d'équivalence a été produite à l'automne 2005 (voir Annexe III). On pourra remarquer aussi que ce cours équivaut à un même nombre de crédits pour les années 1997 à 2000 et ce n'est qu'en 2001 qu'il est passé à 2 crédits mais cette différence, quant au nombre de crédits, n'a aucun impact pour notre recherche puisque les observations  $O_{2c}$  et  $O_4$  se situent respectivement en 1997 et 1999.

Le modèle pronostique sera établi à partir des observations  $O_1$  et  $O_{2c}$  pour la cohorte de 1997. Et une validation externe sera effectuée en utilisant les observations  $O_{2p}$ ,  $O_3$  et  $O_4$ .

### ***3. Les hypothèses de recherche***

La recherche vise en premier lieu à répondre à la question : «Est-ce que le modèle pronostique appliqué aux données d'une cohorte de l'École Polytechnique de Montréal permet de prédire la réussite des sujets?» Pour cela, il ne faut pas une seule hypothèse mais un corpus d'hypothèses. La première hypothèse sera donc relative à la validité interne de l'épreuve en ce sens qu'elle vérifie les qualités pronostiques de l'instrument pour les données de la cohorte 1997. L'hypothèse 2 consistera à vérifier la validité pronostique à long terme. Enfin, l'hypothèse 3 servira à confirmer la validité externe du modèle pronostique dans le sens que ce modèle pronostique pourra être appliqué à une autre cohorte.

- *L'hypothèse de la validité interne du modèle pronostique*

La première hypothèse veut que les données provenant des observations  $O_1$  et  $O_{2c}$  peuvent servir à élaborer un modèle avec une zone d'incertitude limitée à l'intérieur de l'étendue pour l'échelle des scores; le corollaire étant qu'il existe deux intervalles de confiance pour les diagnostics positif et négatif. En d'autres mots, le modèle produit devra ressembler à la fonction logistique de la figure 10 dans la section 7 du cadre conceptuel. La valeur vraie de l'implication logique viendra confirmer la qualité pronostique de l'épreuve.

$H_1: \quad \|X \subseteq Y\| \geq 0,75 \quad \text{Il est plus vrai que faux que le diagnostic}$   
 $\|X' \subseteq Y'\| \geq 0,75 \quad \text{posé à l'aide du modèle pronostique est}$   
*une condition suffisante pour prédire la*  
*réussite du cours «Calcul I».*

- L'hypothèse de l'exactitude du pronostic de la réussite dans un programme

Les données provenant des observations  $O_1$  et  $O_{2p}$  seront utilisées pour vérifier les qualités pronostiques de l'épreuve en regard de la réussite dans un programme.

$H_2: \quad \|X \subseteq Y\| \geq 0,75 \quad \text{Il est plus vrai que faux que le diagnostic}$   
 $\|X' \subseteq Y'\| \geq 0,75 \quad \text{posé à l'aide du modèle pronostique est}$   
*une condition suffisante pour prédire la*  
*réussite dans le programme auquel est*  
*inscrit le sujet.*

Ainsi, nous pourrions voir si l'épreuve diagnostique possède une validité pronostique à long terme.

- L'hypothèse de la généralisation du modèle à d'autres cohortes

Si cette dernière hypothèse était vérifiée alors il faudrait chercher à généraliser les résultats à d'autres cohortes. Pour cela, les indices d'entropie et de taux de d'appartenance seraient calculés à partir des observations  $O_1$  et  $O_3$ . La description des ensembles flous à partir de ces indices viendra indiquer si les cohortes et les épreuves sont équivalentes d'une année à l'autre. Le modèle pronostique établi à partir des données de la cohorte de 1997 sera alors appliqué aux données de l'observation  $O_3$  de la cohorte de 1999. Ainsi, l'hypothèse serait :

$H_3: \quad \|X \subseteq Y\| \geq 0,75 \quad \text{Il est plus vrai que faux que le diagnostic}$   
 $\|X' \subseteq Y'\| \geq 0,75 \quad \text{posé à l'aide du modèle pronostique est}$   
*une condition suffisante pour prédire la*  
*réussite en tout temps du cours «Calcul I».*

La confirmation de cette dernière hypothèse nous permettra de généraliser le modèle pronostique à d'autres cohortes. La confirmation des trois hypothèses constituera la preuve de validité du modèle pronostique.

Notons que les hypothèses prennent pour assises que le diagnostic est une condition suffisante afin d'établir le pronostic. Toutefois, il serait possible que le pronostic ne s'explique pas seulement par le diagnostic mais soit influencé par les autres attributs intervenants.

#### *4. La méthode d'élaboration du modèle pronostique*

Rappelons-le, le modèle pronostique comprend deux paramètres : le paramètre  $\alpha$  qui établit la zone d'incertitude et le paramètre  $\beta$  qui constitue le score de césure. Nous allons utiliser la méthode du modèle des groupes contraires pour inférer le paramètre  $\beta$ . Pour trouver la valeur du paramètre  $\alpha$ , l'indice de fidélité de Livingston sera calculé et la zone d'incertitude sera établie à partir de l'erreur-type de mesure.

- *La méthode du modèle des groupes contraires*

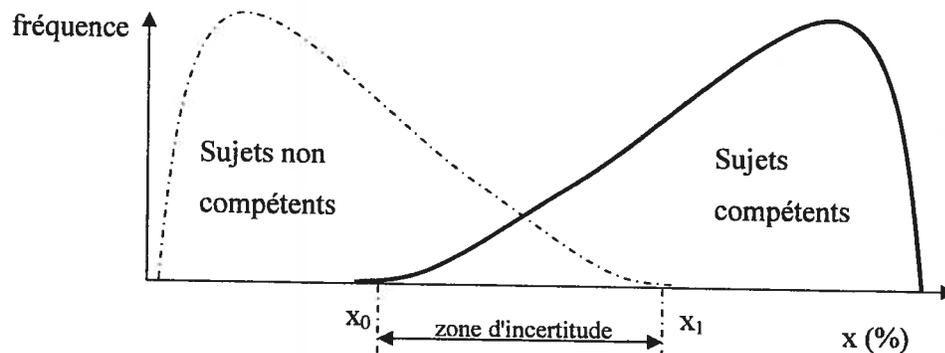
Nedelsky (1954) propose la méthode du modèle des groupes contraires pour établir un standard de performance c'est-à-dire, le score de césure à partir duquel nous sommes assurés qu'un sujet sera jugé compétent:

- Identification des juges qui sont familiers avec la population de sujets.
- Discussion des juges pour déterminer la performance minimale acceptable.
- Identification par les juges des sujets qui maîtrisent et ceux qui ne maîtrisent pas la compétence mesurée. Les cas incertains sont rejetés.
- Administration du test aux deux groupes.
- Juxtaposition des courbes de distribution des scores pour les deux groupes.
- Ajustement du standard de performance basé sur l'intersection de ces deux courbes.

Le plus souvent, l'étendue de la distribution des scores au test, pour les sujets compétents, chevauche l'étendue de la distribution des scores pour les sujets non

compétents (voir figure 15). D'une part, si on établit le standard de performance au score  $x_1$ , c'est-à-dire au-dessus duquel on ne retrouve strictement que des sujets compétents, alors on annulera les cas faussement positifs. D'autre part, si on établit le standard de performance au score  $x_0$ , c'est-à-dire au-dessous duquel on ne retrouve strictement que des sujets non compétents, alors on annulera les cas faussement négatifs.

Les deux distributions théoriques sont illustrées dans le graphique de la figure 15. La variable  $x$  représente le score au test et il devient évident qu'il existe une zone d'incertitude pour le score de césure.



**Figure 15: Distributions des scores pour des groupes contraires**

Dans le cadre de notre recherche, nous allons appliquer une variante de cette méthode pour déterminer le score de césure constituant le paramètre  $\beta$  du modèle:

1. Le critère de réussite pour le cours «Calcul I» remplacera l'équipe de juges. De Landsheere (1988, p.165) indique que l'on peut utiliser une méthode expérimentale qui utilise un critère externe pour déterminer le score de césure.
2. La performance minimale pour réussir ce cours sera la cote  $C$  qui constitue très souvent le seuil de réussite pour des études universitaires de premier cycle.

3. Afin d'éliminer les cas incertains, nous allons former un groupe fort composé des sujets qui ont obtenu une cote égale ou supérieure à C+ pour le cours «Calcul I» et un groupe faible formé de ceux qui ont obtenu une cote égale ou inférieure à D+. Il devrait y avoir assez de sujets dans chaque groupe pour rendre significatif les calculs.
4. Parmi tous les sujets qui ont à la fois passé l'épreuve diagnostique et suivi le cours «Calcul I», on ne retient que les sujets des groupes faible et fort.
5. On ajuste le score de césure au centre de la zone d'incertitude afin d'éliminer l'erreur systématique.

Les deux distributions des scores asymétriques de la figure 15 représentent une situation idéale où la plupart des sujets jugés compétents réussissent le test et la plupart des sujets jugés incompetents l'échouent. Or, en pratique nous assumons plutôt que les deux distributions sont symétriques et ont le même écart-type ; alors, on peut estimer la valeur du point de césure,  $\beta$ , en prenant la moyenne des moyennes pour chacune des distributions. Par la suite, on répète la méthode en restreignant encore plus les critères de formation des groupes faible et fort. En effet, on appliquera la troisième étape de la méthode avec des cotes supérieures ou égales à B pour le groupe fort et avec des cotes inférieures ou égales à D pour le groupe faible. Dans le troisième temps, les critères de formation des groupes seront de B+ et F pour respectivement les groupes fort et faible. La moyenne des trois paramètres  $\beta$  trouvés sera l'estimé pour le paramètre du modèle.

- L'erreur-type de mesure

Une fois le score de césure trouvé, il est possible de calculer l'indice de fidélité de Livingston (voir équation 35) afin de déterminer l'erreur-type de mesure qui nous servira à établir la zone d'incertitude et par le fait même, à spécifier le paramètre  $\alpha$ . Notons que nous avons retenu l'indice de fidélité de Livingston parce qu'il tient compte de l'écart entre les scores bruts et le score de césure. Donc, pour estimer la fidélité avec des sous-tests parallèles, on subdivise l'épreuve en deux moitiés : une moitié avec les items pairs et l'autre moitié avec les items impairs. L'indice de fidélité est calculé à partir des équations 17, 18 et

19 de la section 5 du chapitre second que nous reproduisons ici. À noter que ces énoncés sont fondés sur la théorie classique des tests (TCT).

$$r_{x\theta} = \frac{D_{Xip}}{D_{Xi} D_{Xp}} \quad (35)$$

$$\text{Et : } D_{Xi} = \frac{\sqrt{(\sum (x_{ia} - \beta)^2)}}{\sqrt{N}}$$

$$D_{Xp} = \frac{\sqrt{(\sum (x_{pa} - \beta)^2)}}{\sqrt{N}}$$

$$D_{Xip} = \frac{(\sum (x_{ia} - \beta) (x_{pa} - \beta))}{N}$$

Où :  $r_{x\theta}$  : indice de fidélité ;

$D_{Xip}$  : covariance entre les items pairs et impairs ;

$D_{Xi}$  : écart-type pour le sous-test des items impaires ;

$D_{Xp}$  : écart-type pour le sous-test des items pairs ;

$x_{ia}$  : score du sujet «a» au sous-test des items impaires ;

$x_{pa}$  : score du sujet «a» au sous-test des items paires ;

$\beta$  : score de césure ;

$N$  : nombre de sujets ayant passé l'épreuve diagnostique de juin 1997.

Au préalable, on calcule l'écart-type,  $D_X$ , des scores bruts par rapport au score de césure et à partir de l'équation 6 dans la section 2 du cadre conceptuel, on isole l'erreur-type de mesure pour donner l'équation 36.

$$\sigma_e = D_X \sqrt{(1 - r_{x\theta})} \quad (36)$$

Étant donné que le score de césure a été choisi pour éliminer l'erreur systématique, il est permis de postuler que l'erreur de mesure est distribuée normalement autour du score de césure. Conséquemment, on obtient un intervalle avec un niveau de confiance de 95 % en prenant  $\beta \pm 1,95 \sigma_e$ . La moitié de cette zone d'incertitude sera donc de  $1,95 \sigma_e$  et elle sera aussi égale à

$3/\alpha$  (voir équation 26 dans la section 7 du cadre conceptuel). Donc, ce dernier paramètre pourra être calculé en appliquant l'équation 37 :

$$\alpha = 3/(1,95 \sigma_e) \quad (37)$$

Les paramètres  $\alpha$  et  $\beta$  déterminent entièrement le modèle pronostique.

### ***5. La méthode de validation du modèle pronostique***

C'est le moteur d'inférence qui peut nous informer sur l'influence des variables intervenantes. Auparavant, il faudra définir chacun des attributs pour que l'on puisse les évaluer à partir des données recueillies. Par la suite, le bilan des appartenances nous permettra d'inférer les règles. Deux moteurs d'inférence constitués d'un groupe de deux règles seront élaborés : l'un pour prédire la réussite au cours «Calcul I», l'autre pour prédire la réussite dans le programme auquel est inscrit le sujet.

- ***L'évaluation des attributs***

Pour chaque sujet de la cohorte de 1997 qui a suivi le cours «Calcul I» et qui a passé l'épreuve diagnostique, on établira son degré de capacité en appliquant son score à l'épreuve au modèle pronostique ; ainsi, on obtiendra l'ensemble flou X. On fera de même pour chaque sujet de la cohorte de 1999 mais l'ensemble ne sera pas le même.

L'évaluation des apprentissages à l'École Polytechnique de Montréal est régie par le document «Règlements généraux et particuliers des études supérieures» (ÉPM, 2005). Plus particulièrement, l'article 8.2 précise que l'appréciation de la performance d'un étudiant pour un cours doit être sous la forme de cotes littérales que l'on retrouve dans le tableau X. Ce dernier sera utilisé pour convertir les cotes du cours «Calcul I» en degrés de réussite, et par le fait même en degrés d'appartenance à l'ensemble flou Y.

**Tableau X: Les degrés de réussite au cours «Calcul I»**

Cote	degré de réussite	Prédicat
A*	1,0	Certainement vrai
A	1,0	Certainement vrai
B+	1,0	Certainement vrai
B	0,83	Semble vrai
C+	0,67	Plus ou moins vrai
C	0,5	Ni vrai, ni faux
D+	0,33	Plus ou moins faux
D	0,17	Semble faux
F	0,0	Certainement faux

Ce tableau représente une appréciation à sept échelons et a été construit à partir des trois points d'ancrage suivants : la cote F équivaut à un échec certain (0,0), la cote B+ représente une réussite certaine (1,0) et la cote C est associée au maximum d'ambiguïté quant à la réussite (0,5). Les autres valeurs ont été déduites en appliquant une échelle à 7 échelons.

À partir du modèle pronostique et du tableau X, il devient possible de traduire les données en degrés d'appartenance aux ensembles flous X et Y ; et par le fait même, de vérifier les hypothèses  $H_1$  et  $H_3$ .

On contrôlera les attributs intervenants D et C, c'est-à-dire respectivement les faits que le sujet ait appuyé sa candidature à l'admission avec un diplôme d'études collégiales (DEC) en sciences de la nature ou que le sujet provienne d'un cégep. Ces attributs forment des ensembles ordinaires dans ce sens que l'on ne peut attribuer qu'une valeur de 0 ou 1 à chaque sujet ; 0 représentant la valeur fautive et 1 la valeur vraie.

Pour vérifier l'hypothèse  $H_2$ , il faut évaluer le degré de réussite dans un programme défini en considérant les trois points d'ancrage suivants : le sujet réussit lorsqu'il obtient un diplôme ( $\mu_Y=1$ ), il échoue si le dernier trimestre où il est inscrit est moindre que la durée prescrite, soit le huitième ( $\mu_Y=0$ ), et l'ambiguïté est maximale s'il a suivi le programme durant cinq ans et si le dernier trimestre où il est inscrit est le dixième sans obtenir de diplôme ( $\mu_Y=0,5$ ). Nous postulons également que le sujet suit des cours pour deux trimestres par année. Les autres degrés de réussite seront déterminés en utilisant une échelle à sept échelons comme décrite au tableau V de la section 7 du cadre conceptuel.

Le tableau XI illustre le degré de réussite dans un programme pour les sujets de la cohorte de 1997 en appliquant les préceptes vus plus haut. Le premier trimestre est «Automne 1997» et le dernier est «Hiver 2003».

**Tableau XI : Le degré de réussite dans un programme pour la cohorte 1997**

Donnée		degré de réussite	prédictat
Diplôme obtenu	Dernier trimestre inscrit		
Oui	-----	1,0	certainement vrai
Non	Hiver 2003	0,83	surtout vrai
Non	Automne 2002	0,67	plus ou moins vrai
Non	Hiver 2002	0,5	ni vrai, ni faux
Non	Automne 2001	0,33	plus ou moins faux
Non	Hiver 2001	0,17	surtout faux
Non	Automne 2000	0,0	certainement faux
Non	Hiver 2000	0,0	certainement faux
Non	Automne 1999	0,0	certainement faux
Non	Hiver 1999	0,0	certainement faux
Non	Automne 1998	0,0	certainement faux
Non	Hiver 1998	0,0	certainement faux
Non	Automne 1997	0,0	certainement faux

Selon le site internet de l'École Polytechnique de Montréal (2004), il existe onze (11) programmes de génie distincts auxquels l'étudiant peut s'inscrire. À partir des taux de diplomation pour chacun de ces programmes, on établira un attribut P qui variera de 0 à 1.

Ainsi, on attribuera la valeur 1 pour le programme ayant le taux de diplomation le plus élevé et la valeur 0 pour celui qui a le taux le moins élevé. Pour les programmes ayant des taux de diplomation intermédiaires, on affectera une valeur déterminée par la transformation linéaire de l'équation 38.

$$P = \frac{T_D - T_{D\text{MIN}}}{T_{D\text{MAX}} - T_{D\text{MIN}}} \quad (38)$$

- Où P : attribut intervenant pour pondérer le programme ;  
 $T_D$  : taux de diplomation en % du programme considéré ;  
 $T_{D\text{MIN}}$  : taux de diplomation en % le moins élevé ;  
 $T_{D\text{MAX}}$  : taux de diplomation en % le plus élevé.

L'attribut P pourra être ainsi intégré dans un moteur d'inférence afin que l'on puisse observer son effet.

- Le moteur d'inférence pour prédire la réussite dans un cours

La première étape pour élaborer un moteur d'inférence est de déterminer l'univers. Ici, cet univers aura tout d'abord comme référence l'ensemble des sujets de la cohorte de 1997 afin de vérifier l'hypothèse  $H_1$  et cet univers sera constitué : de l'ensemble X pour le diagnostic, de l'ensemble D pour le diplôme à l'admission et de l'ensemble Y pour la réussite. Il y aura donc une possibilité de  $3^2 - 1$  ou 8 expressions possibles. Par la suite, on considérera une situation semblable pour les sujets de la cohorte de 1999 afin de vérifier l'hypothèse  $H_3$ .

La construction du moteur d'inférence a été décrite dans la section 9 du cadre conceptuel et la première étape de la construction consiste à établir le bilan des appartenances. Pour chaque sujet «a» on déterminera : son degré de capacité,  $\mu_X(a)$ , son degré de réussite,  $\mu_Y(a)$  et son degré d'appartenance aux attributs intervenants.

Afin de vérifier l'effet d'une condition causale sur le critère, on n'a qu'à déterminer le rapport entre le cardinal de l'intersection de la cause et de l'effet puis le cardinal de la cause (voir équation 39). Une condition causale sera significative si la valeur vraie de l'implication logique dépasse significativement le repère de 0,75 et alors une règle sera inférée.

$$\begin{aligned} \text{cause} \cap \text{effet} : & \quad \frac{|X \cap M \cap D \cap Y|}{|X \cap M \cap D|} = \text{valeur vraie} \quad (39) \\ \text{cause} : & \quad |X \cap M \cap D| \end{aligned}$$

Nous testerons chacune des expressions causales au seuil de signification statistique de 0,05. Du moteur d'inférence ainsi généré, nous devrions retrouver les conditions suffisantes de l'hypothèse  $H_1$ . Les autres expressions des prémisses vont nous servir pour la discussion des résultats.

Étant donné que les attributs associés aux variables intervenantes constituent des ensembles ordinaires, il est permis d'utiliser la loi du tiers exclus en plus des autres lois de l'algèbre de Boole pour simplifier une expression (voir le tableau III dans la section 5 du cadre conceptuel). À chacune des deux conclusions possibles ( $Y$  et  $Y'$ ), on fera correspondre une prémisse constituée de l'une des expressions causales. Or, ces prémisses pourront être simplifiées afin d'éliminer s'il y a lieu un attribut intervenant.

Par exemple, si une prémisse est constituée des expressions suivantes :  $X \cap D$  et  $X \cap D'$ ; alors on pourrait simplifier de la manière suivante :

$$\begin{aligned} X \cap D \cup X \cap D' & \\ X \cap (D \cup D') & \quad \textit{distributivité} \\ X \cap U & \quad \textit{loi du tiers exclus} \\ X & \quad \textit{neutralité} \end{aligned}$$

Dans cet exemple, on a éliminé l'attribut  $D$  comme variable intervenante. En d'autres mots, cela signifierait que le fait de détenir un DEC en sciences de la nature n'est pas une condition nécessaire pour réussir.

D'une façon générale, on peut dire qu'un attribut intervenant n'aura pas d'effet si ses valeurs affirmative,  $D$  et négative,  $D'$  sont nécessaires à une même expression causale. Cela ne vaut que si l'attribut intervenant est caractérisé par un ensemble ordinaire parce que cette simplification nous vient de la loi du tiers exclus.

On pourra déterminer qu'une condition est suffisante si on retrouve un seul attribut dans une expression. Par exemple dans l'expression suivante, l'attribut X est suffisant :

$$X \cup X \cap D' = X$$

$$\text{Preuve : } \mu_{X \cup X \cap D'} = \text{Max} (\mu_X, \text{Min} (\mu_X, 1 - \mu_D))$$

$$\text{mais } \text{Min} (\mu_X, 1 - \mu_D) \leq \mu_X$$

$$\text{et conséquemment } \text{Max} (\mu_X, \text{Min} (\mu_X, 1 - \mu_D)) = \mu_X$$

La suffisance d'un attribut vient nous signifier que l'on n'a pas besoin de combiner ce même attribut avec d'autres pour expliquer la cause d'un effet observé.

- *Le moteur d'inférence pour le pronostic de la réussite dans un programme*

Ce moteur d'inférence servira à vérifier l'hypothèse  $H_2$ . Et l'univers serait constitué d'un ensemble supplémentaire, l'ensemble P pour la nature du programme où le sujet est inscrit. Les attributs D et C ne seront considérés que s'ils ont démontré leur nécessité pour le pronostic de la réussite du cours «Calcul I».

Pour construire le moteur d'inférence, on effectuera le bilan des appartenances mais cette fois-ci, il y aura une possibilité de  $3^3 - 1$  ou 26 expressions. Cependant l'ensemble critère Y ne serait pas le même puisqu'il s'agit maintenant de la réussite dans le programme. Les données sur l'inscription à chaque trimestre et la diplomation seraient traduites en degrés d'appartenance à l'ensemble flou Y selon le tableau XI. Ici, une expression causale sera retenue si la valeur vraie de l'implication logique est significativement plus élevée que le repère de 0,75. La signification statistique sera de 0,05 et la valeur de z calculée selon l'équation 33 dans la section 8 du cadre conceptuel devra être supérieure ou égale à 1,65.

## ***6. Les limites à la généralisation des résultats***

Rappelons-nous le problème que nous cherchons à résoudre. Nous voulons prédire la réussite des étudiants admis afin de leur prescrire un enseignement

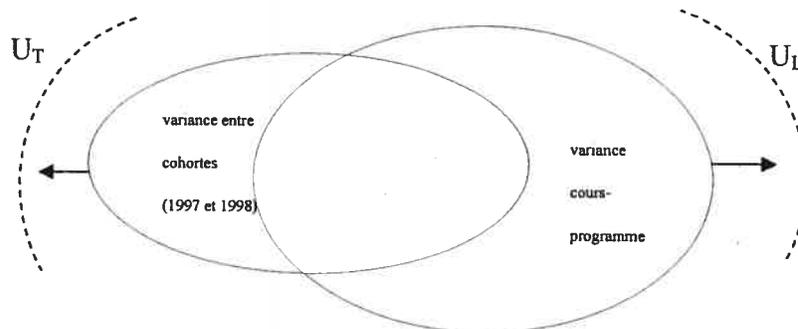
d'appoint si nécessaire et ce, dans la perspective d'améliorer la performance des processus d'enseignement au sens de la Loi sur les établissements d'enseignement de niveau universitaire (L.R.Q., c. E-14.1; a.1). Or, cette amélioration de la performance ne pourra se faire que si le modèle pronostique peut être généralisé.

- Les facettes de la mesure

La généralisation de ce modèle pronostique ne se fera que par rapport à l'efficacité du modèle pronostique pour prédire la réussite. Ainsi, un pronostic sera exact lorsque le sujet ayant été diagnostiqué avec certitude, a une réussite observée identique à celle qu'il aurait dû avoir. Un diagnostic sera certain lorsque le degré de capacité du sujet,  $\mu_x$ , est supérieur à 0,95 ou inférieur à 0,05. Un sujet sera considéré avoir réussi, si son degré de réussite,  $\mu_y$ , est supérieur à 0,95.

Étant donnée que nous pourrions classer chaque cas comme pronostic exact ou pronostic inexact, il devient possible de recourir aux techniques statistiques non-paramétriques. Un seuil de signification de 0,05 sera utilisé pour décider du rejet ou de l'acceptation de l'hypothèse nulle.

Il y aura deux facettes pour la généralisation du modèle pronostique : une facette longitudinale et une facette transversale. La facette longitudinale cherchera à généraliser l'efficacité du pronostic de réussite à travers le temps pour un même individu. Quant à la facette transversale, elle cherchera à généraliser l'efficacité du pronostic de réussite à une population d'individus. Le plan pour la généralisation est illustré à la figure 16.



**Figure 16 :** Plan pour la généralisation du modèle pronostique

À la limite, la facette longitudinale nous permettra d'inférer la réussite professionnelle future,  $U_L$ , d'un individu. De même à la limite, la facette transversale nous permettra d'inférer la validité du modèle pronostique à la population des étudiants admis à l'École Polytechnique de Montréal,  $U_T$ .

- La généralisation de la réussite

Le test statistique non-paramétrique de McNemar (Siegel, 1956 ; p.63) sera utilisé pour généraliser le pronostic de réussite selon la facette longitudinale. L'hypothèse nulle  $H_0$  énonce une faible variance cours-programme en ce qui concerne l'exactitude du pronostic et c'est sa confirmation qui nous permettra de généraliser.

$H_0$ : *Il n'y a pas de différences significatives entre l'exactitude du pronostic de réussite pour un cours et celle pour un programme.*

$H_1$ : *Il y a une différence significative entre l'exactitude du pronostic de réussite pour un cours et celle pour un programme.*

Afin de vérifier cette hypothèse, nous utiliserons un tableau de contingence de 2 par 2, c'est-à-dire avec un degré de liberté ( $df=1$ ). Les colonnes vont différencier les cas inexacts des cas exacts alors que les lignes serviront à différencier le cours du programme. Le calcul du  $\chi^2$  avec les résultats pour la

cohorte de 1997 devrait être égal ou inférieur à 3,84 pour accepter l'hypothèse nulle.

- La généralisation à la population

D'une année à l'autre, l'épreuve de mathématique est élaborée à partir d'une banque de questions d'au-delà 900 items. Beaucoup de ces items ont été validés par les précédentes administrations de l'instrument. Et le même tableau de spécification est utilisé pour construire les épreuves. Aussi, nous assumons que les instruments sont équivalents pour une cohorte à l'autre. Encore une fois, selon l'hypothèse nulle  $H_0$  la variance entre les cohortes devrait être faible.

Le test statistique non-paramétrique du  $\chi^2$  pour des échantillons indépendants (Siegel, 1956 ; p.104) sera utilisé pour vérifier l'hypothèse nulle que voici:

$H_0$ : *Il n'y a pas de différences significatives entre l'exactitude du pronostic de la réussite du cours «Calcul I» pour la cohorte de 1997 et celle pour la cohorte de 1999.*

$H_1$ : *Il y a des différences significatives entre l'exactitude du pronostic de la réussite du cours «Calcul I» pour la cohorte de 1997 et celle pour la cohorte de 1999.*

Encore ici, un tableau de contingence de 2 par 2 sera utilisé. Les colonnes vont différencier les cas inexacts des cas exacts alors que les lignes spécifieront les cohortes. Pour l'acceptation de l'hypothèse nulle il faudrait que le calcul du  $\chi^2$  donne une valeur égale ou inférieure à 3,84.

- La puissance statistique

La puissance statistique est rarement spécifiée pour des tests statistiques non-paramétriques et de ce fait, on ne se restreindra pas seulement aux résultats de ces tests pour inférer les résultats. Pour la facette longitudinale, il faudra tenir compte de la distribution des erreurs de prédiction relativement à la nature du programme alors que pour la facette transversale, on vérifiera la correspondance entre les descriptions des ensembles diagnostiques X d'une cohorte à l'autre.

Ces descriptions se feront par des histogrammes et des calculs pour l'entropie et le taux d'appartenance.

En résumé, nous cherchons à élaborer un modèle pronostique qui prédit la réussite sans pour autant recourir aux méthodes traditionnelles dites quantitatives ou qualitatives. En effet, nous utiliserons plutôt une méthode mixte, fondée sur la théorie des ensembles flous, qui permet de traiter à la fois les données qualitatives et quantitatives

## Chapitre quatrième

### Analyse des données

Ayant déterminé la méthode pour élaborer et valider le modèle pronostique, il nous reste maintenant à l'appliquer aux données disponibles de l'École Polytechnique. Auparavant, il faut décrire et classer ces données dans des ensembles flous si on veut ensuite faire des inférences sur des relations causales entre la capacité et la réussite des sujets.

#### *1. La description des données*

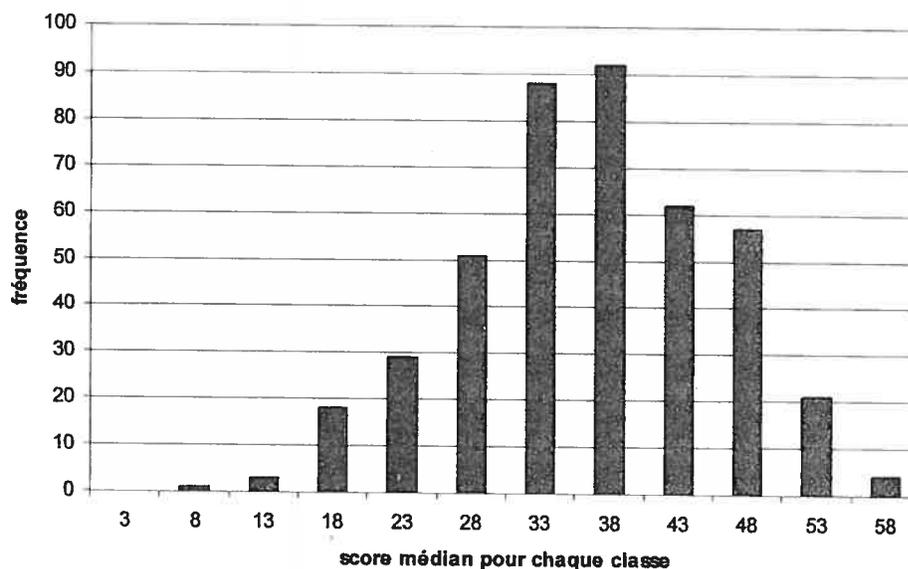
Le plan d'analyse prévoit des cueillettes de données pour les cohortes de 1997 et 1999. Nous allons maintenant décrire ces données.

- *La description des données pour l'observation en  $O_1$*

Les étudiants faisant une demande d'admission à l'École Polytechnique de Montréal doivent être soumis à l'épreuve diagnostique de mathématique. Aussi, trois séances d'administration de cette épreuve sont prévues par année académique: en juin, en août et en janvier. Comme la séance de juin est de loin celle où il y a plus d'étudiants, nous prenons en compte seulement les données relatives à cette épreuve de juin 1997 pour constituer les données de l'observation en  $O_1$ .

Les données furent compilées dans un fichier Excel pour les 426 étudiants passant le test; plus précisément, pour chacun des sujets, on a obtenu ses réponses à chacun des 60 items du test. De plus, chacune des réponses, notée de 1 à 4, correspond au choix du sujet. À partir de la clé de correction, il est possible de traduire ces réponses sous une forme dichotomique 0/1 (échec/réussite) pour chaque item. Ainsi, le score sur 60 se calcule en effectuant la somme de ces réponses dichotomiques. Notons toutefois que pour les fins de la recherche, contrairement aux directives pour la correction, aucun score négatif ou partiel n'est donné au sujet qui échoue un item.

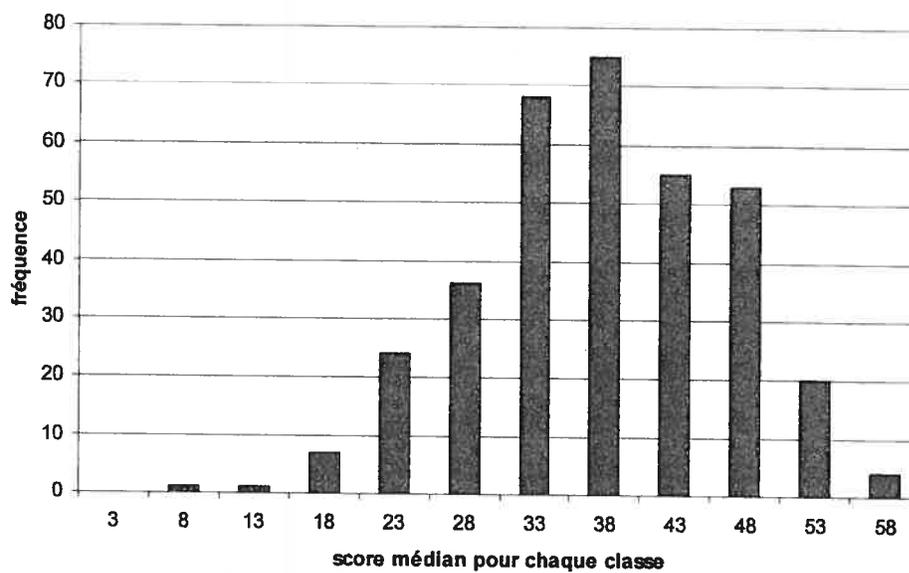
Donc, pour chacun des 426 sujets admis à l'automne 1997 et ayant passé le test de juin de la même année nous retrouvons le score sur 60. La moyenne obtenue fut de 37 et l'écart-type de 9. L'histogramme de la figure 17 représente la distribution des scores pour ces 426 sujets.



**Figure 17: Distribution des scores à l'épreuve de juin 1997 (N=426)**

Dans cet histogramme, il y a douze classes de scores regroupées par intervalle de cinq points. L'abscisse indique le score médian pour chacune de ces classes alors que l'ordonnée indique la fréquence d'observation. Cet histogramme nous montre que le mode est vis-à-vis de la classe dont le score médian est 38. La médiane de la distribution semble aussi se situer vis-à-vis de la classe 38. Étant donné que la moyenne, la médiane et le mode se situent dans la même classe de scores, nous postulons que cette distribution est approximativement normale.

Cependant, ce ne sont pas tous les sujets ayant passé l'épreuve de juin 1997 qui ont suivi le cours «Calcul I» l'année suivante. En effet, seulement 344 sujets ont à la fois passé l'épreuve de juin 1997 et suivi le cours 105 (Calcul I); la distribution des scores pour ces 344 sujets est illustrée par l'histogramme de la figure 18 avec les mêmes classes que précédemment.



**Figure 18 : Distribution de juin 1997 pour les sujets du cours 105 (N=344)**

Les histogrammes des figures 17 et 18 semblent à première vue représenter des distributions identiques. Et le tableau XII vient compléter la comparaison par le calcul d'indices statistiques.

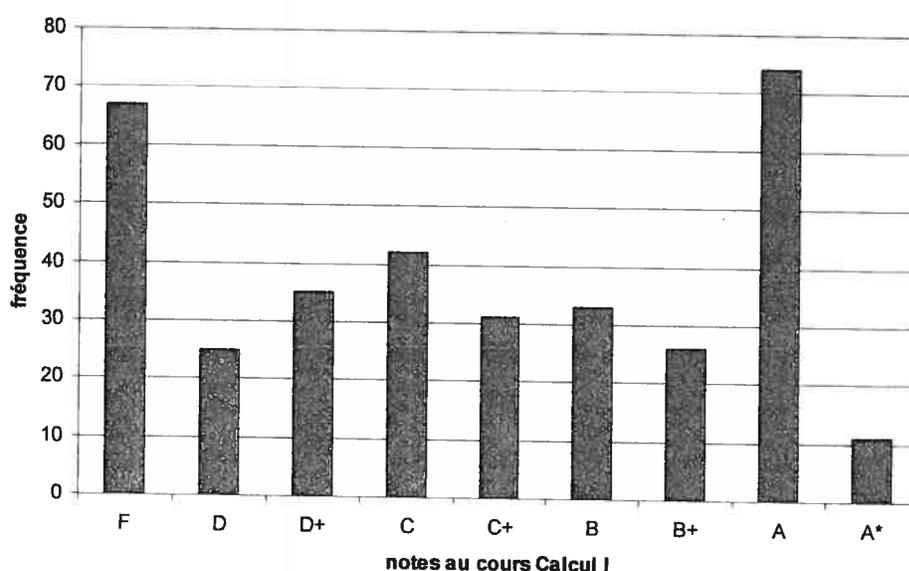
**Tableau XII: Comparaison des distributions précédentes**

Histogramme	Figure 17	Figure 18
N	426	344
Moyenne	37	38
Écart-type	9,00	8,93
Mode	38	38

Les données relatives aux observations en  $O_{2c}$  et  $O_{2p}$  ne concernent que les 344 sujets dont la distribution des scores à l'épreuve diagnostique est représentée par l'histogramme de la figure 18.

- La description des données pour l'observation en  $O_{2c}$

Le cours «Calcul I» ayant à l'époque le sigle 105 (voir Annexe III) était obligatoire pour les étudiants admis dans les onze programmes de génie et était normalement offert dans les premiers trimestres. Aussi, la cote finale de ce cours pour chaque sujet admis à l'automne 1997 constitue la donnée opérationnelle pour l'observation en  $O_{2c}$ . Comme on l'a indiqué dans le tableau X de la section 5 du chapitre précédent sur la méthodologie, cette donnée est une cote inscrite sous forme littérale; en ce sens, l'histogramme de la figure 19 illustre la distribution des cotes pour les sujets de la cohorte de 1997.



**Figure 19 : Distribution des cotes au cours 105 (N=344)**

Cette distribution montre deux modes qui sont aux extrémités; en effet, nous retrouvons plus de 70 sujets qui ont obtenu la cote A alors qu'il y en a un peu moins de 70 qui ont eu une cote F. Par conséquent, la forme de la distribution n'épouse pas celle d'une courbe normale.

- La description des données pour l'observation en  $O_{2p}$

Aucune donnée ne nous fut transmise au sujet des taux de diplomation des programmes d'études. Cependant, des données étaient disponibles sur le fait que

le sujet ait obtenu ou non son diplôme et sur le programme auquel il était inscrit. Il devient alors possible de calculer le taux de diplomation pour chacun de ces programmes en effectuant le rapport entre le nombre de diplômés et le nombre d'étudiants admis. Toutefois, le nombre d'inscriptions dans certains programmes était vraiment très peu élevé, cela pouvait affecter la fidélité de cet indicateur. Aussi nous avons regroupé ces programmes afin de n'en faire qu'un seul.

Parmi les 344 sujets de la cohorte de 1997 ayant passé l'épreuve de juin de la même année et étant inscrits au cours «Calcul I», un cas ne possédait aucune donnée sur le diplôme et sur le dernier trimestre inscrit alors que trois cas étaient dans un programme ne menant pas à un diplôme. Conséquemment, nous avons éliminé ces quatre cas pour ne retenir que 340 sujets lors de l'analyse de données. C'est ce nombre de 340 sujets qui sera considéré être le nombre de sujets admis.

Les données signalent aussi que les sujets étaient inscrits dans 13 programmes différents de génie. À remarquer que le site internet de l'École Polytechnique ne montre que 11 programmes pour 2005. Parmi ces 13 programmes, 6 programmes avaient moins de 10 étudiants d'inscrits ; aussi, nous avons regroupé ces 6 programmes pour n'en faire qu'un seul.

À partir du taux de diplomation pour un programme, il devient possible de calculer le degré d'appartenance à l'attribut P,  $\mu_P$ , pour chaque sujet en recourant à l'équation 38 dans la section 5 du chapitre sur la méthodologie. Pour cela, une transformation linéaire du taux de diplomation a été effectuée en tenant compte des taux obtenus, c'est-à-dire un taux minimum de 65,4% et un taux maximum de 100,0%. Enfin, le tableau XIII illustre les taux de diplomation des programmes pour la cohorte de 1997 et le degré d'appartenance afférent pour l'attribut P ; ce degré d'appartenance,  $\mu_P$ , pouvant s'assimiler à un indice de difficulté du programme.

**Tableau XIII : La diplomation dans les programmes de génie**

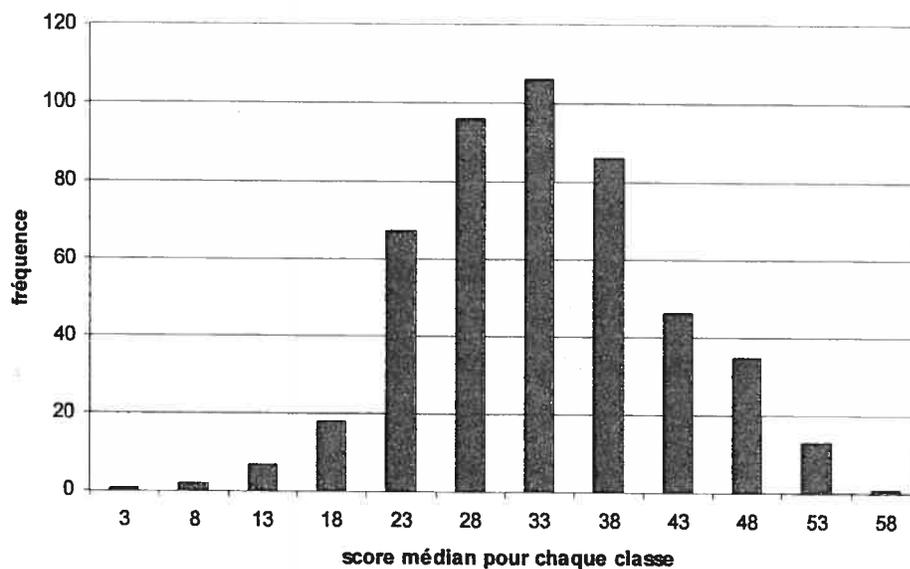
Sigle	Nbre insc.	nbre dipl.	Taux dipl.	$\mu_p$
BIMEC	70	59	84,3%	0,55
BIELE	70	58	82,9%	0,50
BIPHS	14	14	100,0%	1,00
BICHC	15	12	80,0%	0,42
BIINF	102	81	79,4%	0,40
BICHE	19	13	68,4%	0,09
BIIND	24	20	83,3%	0,52
BIREG*	26	17	65,4%	0,00
Total =	340	274	80,6%	---

BIREG\* : regroupement des programmes BIMAC, BIMET, BIGLQ, BIPHC, BIMIN, BICIV

Le préfixe BI signifie Baccalauréat en ingénierie ; ainsi BIMEC représente le Bac. en ingénierie mécanique, BIELE le Bac. en ingénierie électrique, BIINF le Bac. en ingénierie informatique, etc. On peut remarquer que 242 sujets sur une possibilité de 340 étaient inscrits dans trois programmes, soient : en BIMEC, BIELE et BIINF. Ces trois programmes ont un taux de diplomation aux environs de 80% alors que le taux global est de 80,6%.

- La description des données pour les observations en  $O_3$  et en  $O_4$

Il y eut 478 sujets passant l'épreuve de juin 1999, avec une moyenne de 33,2 par rapport à un score maximum de 60 et un écart-type de 8,81. L'histogramme de la distribution des scores est montré à la figure 20.



**Figure 20** : Résultats obtenus à l'épreuve de juin 1999 (N=478)

Comme l'illustre l'histogramme de la figure 20, le mode est à 33, ce qui est très près de la moyenne. On peut remarquer également que la forme de la distribution est symétrique par rapport à l'axe des x.

Rappelons que les données relatives à l'observation en  $O_4$  sont les cotes pour le cours Calcul I de la cohorte de 1999. Or, selon l'annexe III, le cours de Calcul I porte le sigle «ING1001» avec les cheminements scolaires pour les étudiants qui ont été admis à l'automne 1999. Seuls 389 sujets qui avaient passé l'épreuve de juin ont eu une cote pour le cours «ING1001».

## **2. Le classement des données**

Avant tout, les données issues des observations en  $O_1$  et en  $O_{2c}$  vont servir pour déterminer les paramètres  $\alpha$  et  $\beta$  du modèle pronostique. D'une part, ce modèle sert à classer les données provenant des observations en  $O_1$  et en  $O_3$  dans des ensembles flous X. D'autre part, les données pour les observations en  $O_{2c}$ , en  $O_{2p}$  et en  $O_4$  sont classées dans des ensembles flous Y.

- Application de la méthode du modèle des groupes contraires

Afin de déterminer le score de césure,  $\beta$ , selon la méthode du modèle des groupes contraires (voir section 4 de la méthodologie), nous considérons l'ensemble constitué des sujets de la cohorte 1997 ayant obtenu une cote pour le cours 105 (Calcul I) lors de l'observation en  $O_{2c}$ . Aussi, le tableau XIV représente la compilation des résultats obtenus pour trois ensembles de critères de formation des groupes fort et faible ; ces critères furent établis à partir de la cote des sujets et la moyenne calculée, à partir des scores obtenus à l'épreuve diagnostique de juin 1997 pour le sous-groupe formé selon le critère de discrimination (observation en  $O_1$ ).

**Tableau XIV: Groupes contraires pour déterminer la césure**

Discrimination	Petite		Moyenne		Grande	
	Fort	Faible	Fort	Faible	Fort	Faible
Critère	$\geq C+$	$\leq D+$	$\geq B$	$\leq D$	$\geq B+$	$= F$
Nbre de sujets	174	127	143	92	110	67
Moyenne	42,51	32,46	43,79	30,87	45,24	30,54
Césure	37,49		37,33		37,89	
Estimé de $\beta$	37,57					

Pour chacun des 344 sujets de la cohorte de 1997 ayant suivi le cours 105, nous avons formé des groupes fort et faible avec des critères pour obtenir une petite, une moyenne et une grande discrimination. L'estimé pour la césure, représentée par le paramètre  $\beta$ , nous a donné 37,57.

- L'application de la méthode moitié-moitié

Afin de déterminer la valeur du paramètre,  $\alpha$ , qui donne la zone d'incertitude, il faut préalablement calculer l'indice de fidélité de Livingston en appliquant la méthode moitié-moitié. Pour ce faire, nous avons considéré l'épreuve sous l'aspect de deux sous-tests; l'un regroupant les réponses aux items impaires et l'autre, les réponses aux items paires. Aussi, le score à chacun de ces sous-tests

(/30) ainsi qu'un score de césure de 18,785 ( $37,57 \div 2$ ) ont été utilisés pour calculer l'indice de fidélité selon l'équation 35 de la section 4 de la méthodologie. Enfin, le tableau XV compile les données nécessaires au calcul de cet indice de Livingston.

**Tableau XV: Le calcul de l'indice de Livingston (N=426)**

Sous-test	Impair	Pair
Césure	18,785	
Moyenne	20	17
Écart-type	5	5
$D_X^2$	22,17	29,90
$D_{Xip}$	16,41	
$r_{x\theta}$	0,637	

À partir de cet indice de fidélité et de l'écart-type à l'épreuve pour les 426 sujets, l'erreur-type de mesure déduite a donné 5,525 (voir équation 36 dans la section 4 de la méthodologie). Et à partir de l'équation 37 de la même section, nous avons trouvé la valeur de 0,278 pour  $\alpha$ .

En appliquant, les paramètres  $\alpha$  et  $\beta$  trouvés précédemment à l'équation 26 dans la section 7 du cadre conceptuel, nous obtenons le modèle pronostique particulier à notre recherche.

- Le classement des données dans des ensembles flous

Rappelons que le plan d'analyse (voir figure 14 dans la section 2 de la méthodologie) prévoit une cueillette des données à cinq moments : à  $O_1$  en juin 1997, à  $O_{2c}$  pour l'hiver 1998, à  $O_{3p}$  pour tous les trimestres jusqu'en juin 2003, à  $O_3$  en juin 1999 et à  $O_4$  pour l'hiver 2000. Nous obtenons un premier ensemble flou X en appliquant au modèle pronostique trouvé précédemment, le score brut à l'épreuve diagnostique de chacun des 344 sujets de la cohorte de 1997 ayant suivi le cours 105, autrement dit les données observées en  $O_1$ . Afin

d'obtenir un premier ensemble flou Y, nous appliquons la grille établie au tableau X dans la section 5 de la méthodologie aux cotes obtenues par les 344 sujets au cours 105 «Calcul I» (données observées en  $O_{2c}$ ). Enfin, en appliquant la grille du tableau XI dans la section 5 de la méthodologie aux données observées en  $O_{2p}$  auprès des mêmes sujets, nous obtenons un second ensemble flou Y.

Selon l'annexe III, le cours de Calcul I porte le sigle «ING1001» dans les cheminements scolaires des étudiants admis à l'automne 1999. Or, seuls 389 sujets ayant passé l'épreuve de juin ont eu une cote pour ce cours «ING1001». En plus, chaque score des 389 sujets à l'épreuve diagnostique (données observées en  $O_3$ ) fut traduit en un degré de capacité en appliquant le modèle pronostique élaboré précédemment ce qui donna le second ensemble flou X. Enfin, les cotes du cours «ING1001» de ce même groupe de sujets (données observées en  $O_4$ ) ont été traduites sous la forme de degré de réussite en appliquant la grille du tableau X dans la section 5 de la méthodologie ce qui constitua le troisième ensemble flou Y.

Aussi, chaque ensemble flou peut être décrit par deux indicateurs: son taux d'appartenance et son entropie. Plus précisément, le tableau XVI montre la valeur de ces indicateurs décrivant les ensembles flous définis plus haut.

**Tableau XVI: La description des ensembles flous**

Cohorte	Indicateur	X	Y	Y
1997		$O_1$	$O_{2c}$	$O_{2p}$
	Taux d'app.	0,52	0,57	0,81
	Entropie	0,22	0,18	0,01
1999		$O_3$	$O_4$	
	Taux d'app.	0,37	0,52	
	Entropie	0,22	0,11	

Rappelons qu'une entropie de 0 indique que tous les sujets sont classés avec certitude. On peut remarquer une telle situation pour l'ensemble  $Y$  observé à  $O_{2p}$ ; cela signifie que la grande majorité des sujets réussissent ou échouent avec certitude. La description des ensembles flous n'étant pas suffisante pour inférer une implication logique, il faut plutôt établir le bilan des appartenances.

### *3. L'inférence des relations de causalité*

Le bilan des appartenances permet de faire une analyse de données en mettant en évidence l'appartenance à l'intersection de plusieurs ensembles. Ainsi, nous obtenons les valeurs vraies pour les implications logiques représentant les relations de causalité. Chacune des hypothèses sur la validité du modèle pronostique postule une relation de causalité entre la capacité d'un sujet et sa réussite ou son échec. Aussi, pour vérifier chacune de ces hypothèses, nous proposons une démarche commune, soit: le rappel de l'énoncé de l'hypothèse, le test de suffisance pour les critères de réussite et d'échec et le moteur d'inférence.

- *La vérification de l'hypothèse  $H_1$*

L'hypothèse  $H_1$  suppose une relation de causalité entre la capacité d'un sujet de la cohorte de 1997 et sa réussite au cours de mathématique «Calcul I». L'ensemble flou  $X$  représente l'attribut «capacité» alors que l'ensemble flou  $Y$  s'associe au critère de réussite. Selon l'annexe III, le cours «Calcul I» portait le sigle 105 pour cette cohorte. Rappelons aussi que la détention d'un diplôme d'études collégiales en sciences de la nature est l'attribut intervenant dans cette analyse de données selon le plan d'analyse; il s'agit d'un ensemble ordinaire identifié par  $D$ . Parmi les 344 sujets, nous n'avons obtenu de l'information sur cet attribut que pour 218 cas; 192 avaient un DEC en sciences de la nature lors de l'admission alors que 26 ne l'avaient pas.

Enoncé de l'hypothèse

$H_1$  :  $\|X \subseteq Y\| \geq 0,75$       *Il est plus vrai que faux que le diagnostic posé à l'aide du modèle pronostique est une condition suffisante pour prédire la réussite du cours «Calcul I».*  
 $\|X' \subseteq Y'\| \geq 0,75$

Tests de suffisance▪ Critère de réussite

## • Bilan des appartenances

Le tableau XVII représente le bilan des appartenances pour le cours 105 de la cohorte de 1997.

**Tableau XVII : Bilan pour la réussite au cours 105 (N=218)**

Cause	Valeur	Cause $\cap$ effet	Valeur	Inclusion	Valeur
$ X \cap D $	107,1	$ X \cap D \cap Y $	94,0	$\ X \cap D \subseteq Y\ $	0,88
$ X \cap D' $	13,5	$ X \cap D' \cap Y $	10,6	$\ X \cap D' \subseteq Y\ $	0,79
$ X' \cap D $	84,9	$ X' \cap D \cap Y $	49,7	$\ X' \cap D \subseteq Y\ $	0,59
$ X' \cap D' $	12,5	$ X' \cap D' \cap Y $	6,8	$\ X' \cap D' \subseteq Y\ $	0,54
$ X $	176,4	$ X \cap Y $	144,5	$\ X \subseteq Y\ $	0,82
$ X' $	167,6	$ X' \cap Y $	81,4	$\ X' \subseteq Y\ $	0,49
$ D $	192	$ D \cap Y $	125,6	$\ D \subseteq Y\ $	0,65
$ D' $	26	$ D' \cap Y $	15,5	$\ D' \subseteq Y\ $	0,60

Étant donné que la valeur vraie pour  $\|X \subseteq Y\|$  est de 0,82 et dépasse le repère de 0,75, nous allons tester la suffisance de X pour expliquer Y.

- Test statistique unilatéral sur les proportions (voir équation 33 dans la section 8 du chapitre sur le cadre conceptuel):

- seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
- $z = 1,65$  pour un test unilatéral
- $(\|X \subseteq Y\| = 0,82) > 0,75$
- le calcul de  $z$  avec  $N = 176$  donne 2,06 ( $> 1,65$ )

- Décision:

- la partie affirmative de l'hypothèse  $H_1$  est acceptée avec le prédicat «plus vrai que faux» (0,75).

▪ Critère d'échec

- Bilan des appartenances

Le tableau XVIII représente le bilan des appartenances pour le même cours 105 avec le critère d'échec.

**Tableau XVIII : Bilan pour l'échec au cours 105 (N=218)**

Cause	Valeur	Cause $\cap$ effet	Valeur	Inclusion	Valeur
$ X \cap D $	107,1	$ X \cap D \cap Y' $	31,2	$\ X \cap D \subseteq Y'\ $	0,29
$ X \cap D' $	13,5	$ X \cap D' \cap Y' $	4,9	$\ X \cap D' \subseteq Y'\ $	0,36
$ X' \cap D $	84,9	$ X' \cap D \cap Y' $	53,3	$\ X' \cap D \subseteq Y'\ $	0,63
$ X' \cap D' $	12,5	$ X' \cap D' \cap Y' $	7,6	$\ X' \cap D' \subseteq Y'\ $	0,61
$ X $	176,4	$ X \cap Y' $	61,9	$\ X \subseteq Y'\ $	0,35
$ X' $	167,6	$ X' \cap Y' $	116,1	$\ X' \subseteq Y'\ $	0,69
$ D $	192	$ D \cap Y' $	66,4	$\ D \subseteq Y'\ $	0,35
$ D' $	26	$ D' \cap Y' $	10,5	$\ D' \subseteq Y'\ $	0,40

Aucune valeur vraie pour les relations d'inclusion ne dépassent le repère de 0,75.

- Test statistique unilatéral sur les proportions:
  - seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
  - $z = 1,65$  pour un test unilatéral
  - aucune valeur vraie ne dépassant 0,75, nous abaissons le repère à 0,67
  - $(\|X' \subseteq Y\| = 0,69) > 0,67$
  - le calcul de  $z$  avec  $N = 168$  donne 0,45 ( $< 1,65$ )
  
- Décision:
  - la partie négative de l'hypothèse  $H_1$  est rejetée.

### Moteur d'inférence

Le moteur d'inférence ne comprend qu'une seule règle, soit celle d'établir un pronostic sur la réussite:

*Il est plus vrai que faux d'affirmer que si un sujet de la cohorte de 1997 a été diagnostiqué positivement alors il réussira le cours «Calcul I».*

À noter que si le diagnostic n'est pas certain, le prédicat «plus vrai que faux» ne tient pas; il faut plutôt adopter un prédicat moins contraignant en fonction du degré de capacité diagnostiqué.

- La vérification de l'hypothèse  $H_2$

L'hypothèse  $H_2$ , quant à elle, suppose un lien de causalité entre le diagnostic sur la capacité d'un sujet et sa réussite dans le programme. Le diagnostic sur la capacité est représenté par le même ensemble flou  $X$  que précédemment alors que le critère de réussite, associé à l'ensemble flou  $Y$ , diffère. Toutefois, nous n'avons pas retenu l'attribut  $D$  puisque les bilans d'appartenances précédents n'ont pas démontré sa nécessité; ici, c'est plutôt l'attribut  $P$  qui agit comme attribut intervenant. Le degré d'appartenance de cet attribut  $P$  est déterminé en appliquant la donnée  $\mu_P$  pour le programme auquel est inscrit le sujet (voir la grille du tableau XIII dans la section 2 de ce chapitre).



- Test statistique unilatéral sur les proportions:
  - seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
  - $z = 1,65$  pour un test unilatéral
  - la valeur vraie de  $\|X \subseteq Y\|$  dépasse nettement 0,75, aussi nous relevons le repère à 0,83
  - $(\|X' \subseteq Y'\| = 0,89) > 0,83$
  - le calcul de  $z$  avec  $N = 174$  donne 2,01
  
- Décision:
  - la partir affirmative de l'hypothèse  $H_2$  est acceptée avec le prédicat «semble vrai» (0,83).

▪ Critère d'échec

- Bilan des appartenances

Afin de vérifier la partie négative de l'hypothèse  $H_2$ , on utilise le tableau XX qui représente le bilan des appartenances pour l'échec dans un programme.

**Tableau XX : Bilan pour l'échec dans un programme (N=340)**

Cause	Valeur	Cause $\cap$ effet	Valeur	Inclusion	Valeur
$ X \cap P $	104,0	$ X \cap P \cap Y' $	12,7	$\ X \cap P \subseteq Y'\ $	0,12
$ X \cap P' $	121,1	$ X \cap P' \cap Y' $	17,8	$\ X \cap P' \subseteq Y'\ $	0,15
$ X' \cap P $	96,8	$ X' \cap P \cap Y' $	22,3	$\ X' \cap P \subseteq Y'\ $	0,23
$ X' \cap P' $	121,2	$ X' \cap P' \cap Y' $	33,7	$\ X' \cap P' \subseteq Y'\ $	0,28
$ X $	173,5	$ X \cap Y' $	20,0	$\ X \subseteq Y'\ $	0,12
$ X' $	166,5	$ X' \cap Y' $	45,2	$\ X' \subseteq Y'\ $	0,27
$ P $	149,2	$ P \cap Y' $	24,4	$\ P \subseteq Y'\ $	0,16
$ P' $	190,8	$ P' \cap Y' $	40,5	$\ P' \subseteq Y'\ $	0,21

Aucune valeur vraie pour les relations d'inclusion ne dépasse le repère de 0,75.

- Test statistique unilatéral sur les proportions:
  - seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
  - $z = 1,65$  pour un test unilatéral
  - aucune valeur vraie ne dépasse le repère de 0,75 ni celui de 0,67.
  
- Décision:
  - la partie négative de l'hypothèse  $H_2$  est rejetée.

Moteur d'inférence:

*Il semble vrai d'affirmer que si un sujet a été diagnostiqué positivement alors il réussira dans le programme auquel il est inscrit.*

Encore ici, le moteur d'inférence ne comprend qu'une seule règle puisque aucun pronostic sur l'échec ne peut être effectué. De même, le prédicat «semble vrai» ne s'applique que pour un diagnostic certain.

- La vérification de l'hypothèse  $H_3$

L'hypothèse  $H_3$  suppose encore une fois un lien de causalité entre le diagnostic sur la capacité d'un sujet et sa réussite dans le cours «Calcul I». Cependant, ici il s'agit d'une autre cohorte et les ensembles flous X et Y ne sont plus les mêmes que précédemment. En effet, contrairement aux données pour les bilans d'appartenances précédents qui provenaient toutes de la cohorte de 1997, ici les données proviennent de la cohorte de l'automne 1999.

Enoncé de l'hypothèse

$H_3$  :  $\|X \subseteq Y\| \geq 0,75$       *Il est plus vrai que faux que le diagnostic posé à l'aide du modèle pronostique est une condition suffisante pour prédire la réussite en tout temps du cours «Calcul I».*  
 $\|X' \subseteq Y'\| \geq 0,75$

Tests de suffisance▪ Critère de réussite

- Bilan des appartenances

Le tableau XXI représente le bilan des appartenances pour la réussite dans le cours ING1001 auquel est inscrit le sujet de la cohorte de 1999.

**Tableau XXI : Bilan pour la réussite dans le cours ING1001 (N=389)**

Cause	Valeur	Cause $\cap$ effet	Valeur	Inclusion	Valeur
$ X \cap C $	119,1	$ X \cap C \cap Y $	96,0	$\ X \cap C \subseteq Y\ $	0,81
$ X \cap C' $	25,7	$ X \cap C' \cap Y $	19,6	$\ X \cap C' \subseteq Y\ $	0,76
$ X' \cap C $	194,9	$ X' \cap C \cap Y $	92,4	$\ X' \cap C \subseteq Y\ $	0,47
$ X' \cap C' $	49,3	$ X' \cap C' \cap Y $	17,1	$\ X' \cap C' \subseteq Y\ $	0,35
$ X $	144,8	$ X \cap Y $	115,6	$\ X \subseteq Y\ $	0,80
$ X' $	244,2	$ X' \cap Y $	109,5	$\ X' \subseteq Y\ $	0,45
$ C $	314	$ C \cap Y $	170,6	$\ C \subseteq Y\ $	0,54
$ C' $	75	$ C' \cap Y $	32,6	$\ C' \subseteq Y\ $	0,43

Encore ici, la valeur vraie de  $\|X \subseteq Y\|$  dépasse 0,75 et nous allons tester la suffisance de X pour expliquer Y.

- Test statistique unilatéral sur les proportions:

- seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
- $z = 1,65$  pour un test unilatéral
- le choix du repère est de 0,75
- $(\|X \subseteq Y\| = 0,80) > 0,75$
- le calcul de z avec  $N=145$  a donné 1,29 ( $< 1,65$ )
- nous testons la suffisance de la combinaison de X avec C
- $(\|X \cap C \subseteq Y\| = 0,81) > 0,75$
- le calcul de z avec  $N=119$  a donné 1,41 ( $< 1,65$ )

- nous abaissons le repère à 0,67 et nous testons la suffisance de X
- ( $\|X \subseteq Y\| = 0,80$ ) > 0,67
- le calcul de z avec N=145 a donné 3,24 (>1,65)

- Décision:

- la partie affirmative de  $H_3$  est acceptée avec le prédicat «plus ou moins vrai» (0,67).

- Critère d'échec

- Bilan des appartenances

Le bilan des appartenances pour l'échec dans ce même cours est reproduit au tableau XXII.

**Tableau XXII : Bilan pour l'échec dans le cours ING1001 (N=389)**

Cause	Valeur	Cause $\cap$ effet	Valeur	Inclusion	Valeur
$ X \cap C $	119,1	$ X \cap C \cap Y' $	40,8	$\ X \cap C \subseteq Y'\ $	0,34
$ X \cap C' $	25,7	$ X \cap C' \cap Y' $	10,2	$\ X \cap C' \subseteq Y'\ $	0,40
$ X' \cap C $	194,9	$ X' \cap C \cap Y' $	120,3	$\ X' \cap C \subseteq Y'\ $	0,62
$ X' \cap C' $	49,3	$ X' \cap C' \cap Y' $	36,2	$\ X' \cap C' \subseteq Y'\ $	0,73
$ X $	144,8	$ X \cap Y' $	51,0	$\ X \subseteq Y'\ $	0,35
$ X' $	244,2	$ X' \cap Y' $	156,5	$\ X' \subseteq Y'\ $	0,64
$ C $	314	$ C \cap Y' $	143,4	$\ C \subseteq Y'\ $	0,46
$ C' $	75	$ C' \cap Y' $	42,4	$\ C' \subseteq Y'\ $	0,57

Aucune valeur vraie pour les relations d'inclusion n'est supérieure à 0,75.

- Test statistique unilatéral sur les proportions:

- seuil significatif de 0,05 (19 chances sur 20 de dire vrai)
- $z = 1,65$  pour un test unilatéral

- aucune valeur vraie ne dépasse le repère de 0,75, nous abaissons le repère à 0,67 et nous testons la suffisance de la combinaison de X' avec C'
  - $(\| X' \cap C' \subseteq Y' \| = 0,73) > 0,67$
  - le calcul de la variable z avec N=49 a donné 0,74 (<1,65)
- Décision:
    - la partie négative de l'hypothèse H<sub>3</sub> est rejetée.

#### Moteur d'inférence

*Généralement, il est plus ou moins vrai d'affirmer que si un sujet a été diagnostiqué positivement alors il réussira dans le cours «Calcul I».*

La vérification de ces hypothèses n'est pas suffisante pour prétendre à la généralisation.

#### **4. La vérification des limites pour la généralisation**

La théorie des ensembles flous ne nous permet pas de généraliser les résultats à une population car il n'existe pas de modèles théoriques de distribution comme c'est le cas pour les techniques statistiques. Conséquemment, pour généraliser il ne faut considérer que des ensembles ordinaires. Pour notre recherche, il s'agit des ensembles formés par les sujets qui ont obtenu un diagnostic positif certain ( $\mu_X \geq 0,95$ ) et ceux qui ont réussi avec certitude ( $\mu_Y \geq 0,95$ ). Plus spécifiquement, l'exactitude d'un pronostic de réussite se vérifie en prenant le rapport entre le nombre de sujets qui appartiennent à l'intersection de ces deux ensembles et le nombre de sujets qui ont été diagnostiqués positivement avec certitude.

- La vérification de la limite pour la généralisation longitudinale

Rappelons les hypothèses nulle et alternative:

H<sub>0</sub>: *Il n'y a pas de différences significatives entre l'exactitude du pronostic de réussite pour un cours et celle pour un programme.*

$H_1$ : *Il y a une différence significative entre l'exactitude du pronostic de réussite pour un cours et celle pour un programme.*

Le tableau XXIII représente le tableau de contingence pour calculer le  $\chi^2$ .

**Tableau XXIII : L'exactitude du pronostic de réussite**

Pronostic de réussite	Inexact	Exact	Totaux
Cours 105	2 (1,5)	40 (40,5)	42
Programme	1 (1,5)	41 (40,5)	42
<b>Totaux</b>	3	81	84

Les nombres entre parenthèses indiquent les valeurs espérées. Le calcul du  $\chi^2$  donne 0,345, une valeur nettement inférieure au seuil de 3,84 qui permet de décider, avec un seuil significatif de 0,05, le rejet de l'hypothèse nulle. Il faut donc accepter l'hypothèse nulle  $H_0$  énoncée plus tôt que: «il n'y a pas de différences significatives dans l'exactitude du pronostic de la réussite à court et à long terme». Paradoxalement, l'exactitude est plus élevée à long terme (41/42) qu'à court terme (40/42). Enfin, il faut remarquer que le nombre de sujets qui reçoivent un diagnostic positif certain n'est que de 42 sur une possibilité de 343.

- *La vérification de la limite pour la généralisation transversale*

Rappelons les hypothèses nulle et alternative:

$H_0$ : *Il n'y a pas de différences significatives entre l'exactitude du pronostic de la réussite du cours «Calcul I» pour la cohorte de 1997 et celle pour la cohorte de 1999.*

$H_1$ : *Il y a pas une différence significative entre l'exactitude du pronostic de la réussite du cours «Calcul I» pour la cohorte de 1997 et celle pour la cohorte de 1999.*

Le tableau XXIV représente le tableau de contingence pour vérifier cette hypothèse.

Tableau XXIV : La généralisation transversale

Pronostic de réussite	Inexact	Exact	Totaux
Cours 105	2 (1,35)	40 (40,65)	42
Cours ING1001	0 (0,65)	20 (19,35)	20
Totaux	2	60	62

Le calcul du  $\chi^2$  a donné 0,995, une valeur nettement inférieure à 3,84. Il faut donc accepter l'hypothèse nulle à l'effet qu'il n'y a pas de différences significatives entre l'exactitude du pronostic de la réussite d'une cohorte à l'autre.

Tous les résultats obtenus dans ce chapitre doivent être confrontés avec la question de la recherche pour décider de la validité du modèle.

## Chapitre cinquième

### Discussion sur les résultats

Rappelons-nous la question de la recherche énoncée dans le chapitre sur la problématique: «Est-ce que le modèle, issu de la théorie des ensembles flous et appliqué aux scores obtenus par une cohorte de l'École Polytechnique de Montréal à une épreuve diagnostique de mathématique, permet de prédire adéquatement la réussite des sujets dans leur programme?» Autrement dit, est-ce que la capacité en mathématique d'un étudiant (attribut X) est un bon pronostic pour sa réussite (attribut Y). Aussi, faut-il s'assurer de la fidélité dans l'interprétation des données.

#### *1. La fidélité dans l'interprétation des données*

Dans notre recherche, l'interprétation des données implique un classement dans un ensemble flou qui s'effectue selon deux processus. D'une part, les sujets sont classés dans un ensemble diagnostique X en appliquant au modèle pronostique, le score obtenu à l'épreuve de mathématique. D'autre part, les mêmes sujets sont classés dans un ensemble critère Y en appliquant à une grille, soit la cote finale au cours «Calcul I», soit des données sur la détention du diplôme et sur le dernier trimestre inscrit.

- *Le calibrage du modèle*

Le modèle pronostique est un outil pour faire une interprétation critériée des scores (ici, l'épreuve de mathématique) qui se caractérise par une fonction logistique à deux paramètres: le paramètre  $\beta$  ou césure et le paramètre  $\alpha$  qui détermine la zone d'incertitude. Également, les données utilisées pour déterminer ces paramètres proviennent de deux sources: les scores à l'épreuve de juin 1997 et les cotes pour le cours «Calcul I» dispensé lors de l'année académique 1997-1998.

Le tableau XIV (voir section 2 de l'analyse des données) illustre l'application de la méthode du modèle des groupes contraires pour estimer le paramètre  $\beta$  et présente les résultats obtenus en fonction de trois ensembles de critères de discrimination. Ainsi, on observe que l'écart entre les moyennes des groupes fort et faible s'accroît avec la discrimination. En effet, cet écart est de 10,05 points (42,51 – 32,46) pour une petite discrimination, de 12,92 points (43,79 – 30,87) pour une moyenne discrimination et de 14,70 points (45,24 – 30,54) pour une grande discrimination. Toutefois, dans chaque cas on obtient un estimé pour le paramètre  $\beta$  tout près de l'estimation globale de 37,57.

Ainsi, il faut interpréter le score de césure comme étant de 37,57 sur 60 et si on ramène celui-ci en pourcentage, nous obtenons une césure à 62,6%. Pouvons-nous faire un parallèle entre cette césure de 62,6% et la note de passage de 60% que l'on retrouve aux ordres d'enseignement secondaire et collégial? Nous croyons que oui puisque dans les deux cas il s'agit du score où nous avons le maximum d'incertitude pour prendre une décision.

La valeur du paramètre  $\alpha$  a été calculée en considérant l'épreuve en deux sous-tests: l'un regroupant les items pairs et l'autre les items impairs. Aussi, le tableau XV (voir section 2 de l'analyse de données) montre que les sujets ont obtenu en moyenne un score légèrement plus faible au sous-test des items pairs (17) qu'au sous-test des items impairs (20). Nous avons un écart-type de 5 pour ces deux sous-tests et un indice de fidélité de Livingston de 0,637. Cette valeur mitigée pour l'indice de fidélité signifie que quelques sujets ont eu un score plus élevé que la césure dans l'un des sous-tests et moins élevé dans l'autre; autrement dit, on n'a pas de constance dans le diagnostic. Cela est d'autant plus vrai si le sujet obtient un score près de la césure. De plus, la correction pour la longueur du test de l'indice de fidélité par la formule de Spearman-Brown donne 0,778, une valeur considérée faible pour une interprétation normative des scores mais lorsqu'il s'agit d'une interprétation critériée, il faut aussi regarder la distribution des scores. En effet, la moyenne des scores se situe près du score de césure et cela peut expliquer ce faible indice de fidélité.

D'ailleurs, le calcul de l'erreur-type de mesure à partir de l'indice de Livingston de 0,637 a donné une valeur de 5,525, ce qui annonce une zone d'incertitude assez grande. En effet, cette zone d'incertitude est délimitée à sa borne inférieure par la césure moins 1,95 fois l'erreur-type de mesure et à sa borne supérieure par la césure plus 1,95 fois cette erreur-type de mesure, soit  $\beta \pm 1,95 \sigma_e$ . Comme l'illustre le modèle de la figure 21, la limite supérieure de cet intervalle de confiance donne 48,3 alors que la limite inférieure est de 26,8; ici, la valeur du paramètre  $\alpha$  est de 0,278.

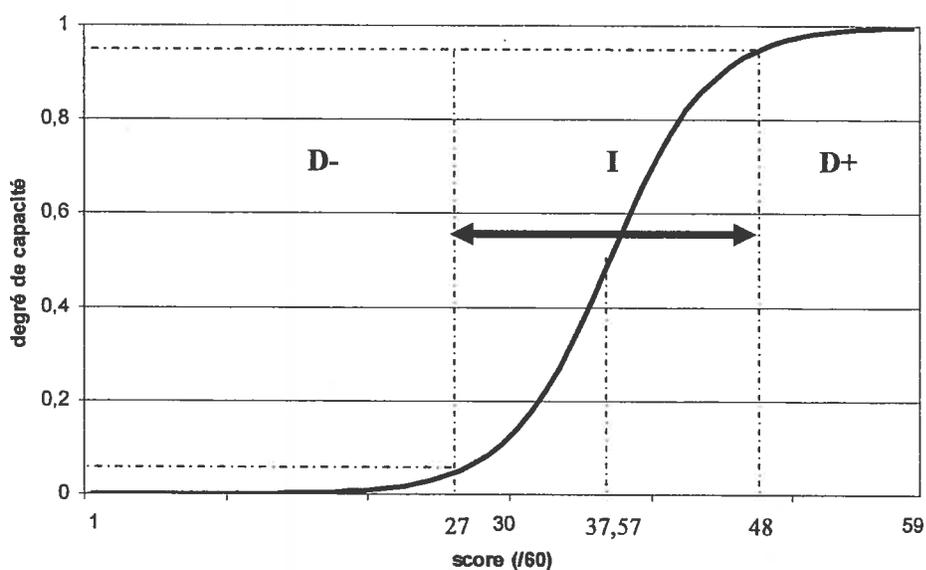


Figure 21 : Modèle pronostique

On peut comparer ce standard de performance de 48,3 ou de 81%, avec le standard de 80% que l'on rencontre couramment pour une interprétation critériée des scores; en effet, dans les deux cas il s'agit du seuil à partir duquel on est certain que le sujet maîtrise les apprentissages. Aussi, le diagnostic est certain pour un score au-dessus du standard de performance ou sous le standard de contre-performance.

- La fidélité du classement dans l'ensemble diagnostic X

On a recours au modèle pronostique pour interpréter les scores à l'épreuve diagnostique et ce, autant pour les sujets de la cohorte de 1997 ( $O_1$ ) que ceux de la cohorte de 1999 ( $O_3$ ). Avec la cohorte de 1997 ( $O_1$ ), il y avait 426 sujets qui passèrent cette épreuve mais nous en retrouvons seulement 344 parmi eux qui suivirent le cours «Calcul I» (105) dans l'année qui suivit, soit une perte de 19,2%. Nous retrouvons le même phénomène avec la cohorte de 1999 ( $O_3$ ) puisqu'il y avait 478 sujets qui passèrent l'épreuve en juin et seulement 389 d'entre eux étaient inscrits au cours «Calcul I» (ING1001), pour une perte de 18,6%.

Maintenant, étudions d'un peu plus près ce phénomène de déperdition pour la cohorte de 1997. Les deux histogrammes des figures 17 et 18 (voir section 1 de l'analyse de données), représentent les distributions des scores à l'épreuve pour respectivement les 426 et les 344 sujets. Ces deux histogrammes épousent une même forme de courbe normale. Le tableau XII, dans la même section, montre qu'il n'y a pas de différences entre ces distributions puisque les modes, les moyennes et les écart-types sont pratiquement identiques. Aussi, tous les étudiants effectuant une demande d'admission doivent passer l'épreuve diagnostique, même si quelques uns d'entre eux choisiront de poursuivre leurs études dans une autre institution. De plus, les autorités mentionnent (École Polytechnique de Montréal, 2004), que le cours «Calcul I» peut être crédité, sous certaines conditions, à des étudiants provenant de cégeps ayant une entente avec l'École. Par conséquent, nous postulons que l'échantillon de sujets ayant obtenu une cote pour le cours «Calcul I» est représentatif de l'ensemble des sujets qui ont passé l'épreuve diagnostique, et cela pour les deux cohortes.

Le modèle pronostique de la figure 21 nous indique une zone d'incertitude dans le diagnostic pour les sujets dont le score se situe entre 27 et 48. Or, l'histogramme de la figure 18 pour la cohorte de 1997 (voir section 1 de l'analyse de données) nous illustre une situation où la majorité des sujets ont un score dans cette zone d'incertitude, soient les classes de scores dont la médiane

est à 33, 38 et 43. Toutefois, on peut remarquer que la situation n'est pas la même avec l'histogramme de la figure 20 pour la cohorte de 1999; en effet, le mode est désormais à 33 et il semble y avoir moins de sujets dans ces mêmes classes de scores.

D'ailleurs, le tableau XVI (voir section 2 de l'analyse de données) montre que l'ensemble X pour la cohorte de 1997 ( $O_1$ ) est caractérisé par un taux d'appartenance de 0,52 et une entropie de 0,22 alors que ce même ensemble pour la cohorte de 1999 ( $O_3$ ) a un taux d'appartenance de 0,37 et une entropie de 0,22. Il y a donc eu un décalage entre les deux cohortes pour le classement des sujets car le taux d'appartenance est passé de 0,52 à 0,37, de même pour le mode qui est passé de 38 à 33. De plus, étant donné que les deux ensembles ont la même entropie, l'incertitude du diagnostic aura le même effet pour les deux cohortes. Nous pouvons penser que le questionnaire de l'épreuve de 1999 comprenait des items plus difficiles car seul le taux d'appartenance a diminué, l'entropie restant identique ; autrement dit, la dispersion des scores est restée la même et il y a eu un décalage du degré d'appartenance pour tout le groupe. Par conséquent, nous croyons que l'interprétation des données reste fidèle d'une cohorte à l'autre.

- *La fidélité du classement dans l'ensemble critère Y pour un cours*

Ici, les données, étant des cotes pour l'appréciation du cours «Calcul I», sont interprétées à travers la grille pour un ensemble flou à sept échelons que l'on retrouve au tableau X (voir section 5 de la méthodologie). En plus, les sujets qui échouent reprennent normalement le cours dans les trimestres qui suivent, toutefois la cote interprétée est celle provenant du résultat obtenu lors de leur première inscription. Enfin, la cote d'un sujet est traduite en un degré de réussite ou en d'autres mots, en un degré d'appartenance à l'ensemble critère Y.

La figure 19, dans la section 1 de l'analyse de données, représente l'histogramme pour la distribution des cotes pour la cohorte de 1997 ( $O_{2c}$ ). Or, on remarque que cette distribution n'épouse pas une forme normale mais possède plutôt deux modes: un mode pour la cote A et un autre pour la cote F. Toutefois, les formes de distribution différentes pour les histogrammes des figures 18 et 19 ( $O_1$  et  $O_{2c}$ )

n'influencent pas la relation causale recherchée entre la capacité  $X$  et la réussite  $Y$  puisqu'il n'est pas nécessaire de postuler la normalité des distributions avec la théorie des ensembles flous.

Maintenant, voyons comment l'interprétation de ces cotes d'appréciation du cours «Calcul I» reste stable entre les cohortes de 1997 et 1999. Le tableau XVI (voir section 2 de l'analyse de données) montre un taux d'appartenance et une entropie respectivement de 0,52 et 0,11 pour la cohorte de 1997 ( $O_{2c}$ ) alors que ces mêmes indicateurs donnent respectivement 0,57 et 0,18 pour la cohorte de 1999 ( $O_4$ ). Ceci démontre donc la fidélité dans l'interprétation des cotes pour le cours «Calcul I» d'une cohorte à l'autre.

- *La fidélité du classement dans l'ensemble critère  $Y$  pour un programme*

Les données relatives à l'observation en  $O_{2p}$  pour la réussite dans un programme (voir figure 14 dans la section 2 de la méthodologie) sont interprétées à la lumière de la grille du tableau XI (voir section 5 de la méthodologie) pour donner des degrés d'appartenance à un ensemble flou  $Y$  à sept échelons. Avec cette grille, le diplôme donne une appartenance de 1,0, sinon c'est le dernier trimestre inscrit qui détermine le degré d'appartenance. Le tableau XVI (voir section 2 de l'analyse de données) met donc en évidence que cet ensemble  $Y$  est précis puisqu'il se caractérise par un taux d'appartenance de 0,81 et une entropie de 0,01; cela s'explique par le fait que les sujets décrocheurs abandonnent dans les premiers trimestres et non après quatre ans d'études. Ainsi, avec une entropie près de 0, seul le diplôme constitue un élément déterminant pour caractériser l'appartenance à cet ensemble critère  $Y$ .

Voyons maintenant comment l'interprétation de cette donnée est fidèle. À l'audition de la Commission de l'éducation de l'Assemblée nationale en juin 2005, le directeur général de l'École Polytechnique affirmait que le contrat de performance prévoyait un taux de diplomation de 73,9% pour 2008. Toutefois, comme le montre le tableau XIII (voir section 1 de l'analyse de données), le nombre de diplômés pour la cohorte de 1997 était de 274. Or, en faisant le rapport entre ce nombre et celui de 426 que nous avons en juin 1997, nous

obtenons un taux de diplomation de 64,3%. Toutefois, le même tableau XIII indique un taux de diplomation global de 80,6% si on fait le rapport entre le nombre de diplômés (274) et le nombre de sujets suivis (340). Puisque nous ne savons pas quel pourcentage des 426 sujets étaient inscrits dans un programme à l'automne 1997, il devient difficile de trouver le taux réel mais la cible de 73,9% semble plausible. Aussi, l'interprétation de la donnée «diplôme» nous semble fidèle.

## 2. *La validité du modèle pronostique*

Le modèle pronostique est un outil pour faire une interprétation critériée des scores à l'épreuve diagnostique. Donc, établir la validité du modèle pronostique signifie aussi certifier la validité pronostique de cette épreuve. Les bilans d'appartenance du chapitre précédent nous ont permis, par les moteurs d'inférence, d'établir une relation causale entre un diagnostic de capacité et la réussite. Cependant, nous n'avons pu obtenir de relation entre un diagnostic d'incapacité et l'échec. Est-ce qu'en ajustant le modèle pronostique, nous pouvons obtenir des résultats plus concluants? C'est ce que nous allons établir maintenant.

- *La validité interne*

Rappelons que l'hypothèse  $H_1$  servait à vérifier la validité interne de l'épreuve. Or, le moteur d'inférence obtenu ne comprenait que la règle affirmative pour prédire la réussite avec le prédicat «plus vrai que faux» (0,75). Un score de césure trop élevé peut expliquer ce résultat.

Aussi, un ajustement du modèle pronostique a été effectué en abaissant le paramètre  $\beta$  jusqu'à ce que les valeurs vraies des deux implications principales soient égales. Avec une valeur de 34 pour  $\beta$ , nous avons obtenu une valeur vraie de 0,75 pour les implications principales  $\|X \subseteq Y\|$  et  $\|X' \subseteq Y'\|$ ; cette dernière expression est désormais significative pour le repère «plus ou moins vrai» (0,67;  $z = 1,66$ ). Toutefois, la valeur vraie de  $\|X \cap D \subseteq Y\|$  a donné 0,83 et est

significative au repère «plus vrai que faux» (0,75;  $z = 1,80$ ). Donc, après l'ajustement du modèle, nous obtenons le moteur d'inférence que voici:

- 1: *Il est plus vrai que faux d'affirmer que si un sujet de la cohorte de 1997 détient un diplôme d'études collégiales en sciences de la nature et qu'il a été diagnostiqué positivement alors il réussira le cours «Calcul I».*
2. *Il est plus ou moins vrai d'affirmer que si un sujet de la cohorte de 1997 a été diagnostiqué négativement alors il échouera le cours «Calcul I».*

Le nouvel ensemble diagnostic X ainsi formé après l'ajustement du modèle pronostique se caractérise par un taux d'appartenance de 0,64 et une entropie de 0,21. Alors que l'entropie n'a presque pas changé, il est compréhensible que le taux d'appartenance ait progressé de 0,52 à 0,64 puisque nous avons abaissé le score de césure de 37,57 à 34.

- *La validité du pronostic de réussite dans un programme*

Établir un pronostic de réussite ou d'échec pour le cours «Calcul I» n'est pas suffisant pour certifier la validité pronostique de l'épreuve diagnostique. Il faut en plus connaître le pronostic à long terme.

Rappelons que c'est la vérification de l'hypothèse  $H_2$  qui nous permet de valider le pronostic de réussite à long terme dans un programme. Or, encore une fois le moteur d'inférence obtenu lors de l'analyse de données ne comprenait que la règle affirmative sur le pronostic de la réussite et restait muet sur le pronostic de l'échec. Cependant, la valeur vraie de la relation causale est plus élevée que précédemment (0,89) puisqu'elle est significative avec le prédicat «semble vrai» (0,83). Il semble paradoxal que l'on puisse prédire la réussite avec une plus grande certitude dans le programme que dans le cours «Calcul I». Mais ce ne l'est pas puisque les étudiants ayant échoué ce cours le reprennent plus tard.

Avec le modèle ajusté à un paramètre  $\beta$  de 34, la valeur vraie pour le pronostic de réussite a descendu à 0,87 ce qui entraîne aussi un abaissement de la signification au prédicat «plus vrai que faux» (0,75;  $z = 3,56$ ). En plus, aucune

valeur vraie ne dépasse le repère le plus faible (0,67) pour prédire l'échec. Donc, le moteur d'inférence demeure avec une seule règle affirmative:

1. *Il est plus vrai que faux d'affirmer que si un sujet a été diagnostiqué positivement alors il réussira dans le programme auquel il est inscrit.*

L'absence d'une règle négative pour établir un pronostic d'échec vient nous signaler que l'information est incomplète. D'autres attributs non observés peuvent sans doute expliquer les échecs à long terme dans un programme.

Le tableau XXIII (voir section 4 dans l'analyse de données) représente le tableau de contingence illustrant l'exactitude du pronostic de réussite dans un cours et dans un programme. Malgré le fait que l'on retrouve peu de cas inexacts, il ressort que l'hypothèse nulle fut acceptée avec une signification de 19 chances sur 20 de dire vrai. Nous concluons que si nous prédisons la réussite d'un sujet dans le cours «Calcul I», nous prédisons en même temps sa réussite dans le programme et également sa réussite professionnelle en généralisant selon le schéma de la figure 16 (voir section 6 de la méthodologie).

- La validité externe

En appliquant le modèle pronostique avec la cohorte de 1999, nous pouvons observer la répétition de ses effets. Aussi, la vérification de l'hypothèse  $H_3$  relative à cette cohorte va nous permettre de valider à l'externe le modèle pronostique. Rappelons-nous que le moteur d'inférence ne comprenait qu'une seule règle affirmative pour prédire la réussite et restait muet sur le pronostic de l'échec.

Avec l'ajustement du modèle, nous obtenons une valeur vraie de 0,74 pour l'expression  $\|X \subseteq Y\|$ , cette valeur vraie étant significative pour un prédicat «plus ou moins vrai» (0,67;  $z = 1,70$ ). Nous avons aussi une valeur vraie de 0,80 pour l'expression  $\|X' \cap C' \subseteq Y\|$  mais celle-ci n'est pas significative pour un prédicat «plus ou moins vrai» (0,67;  $z = 1,38$ ). Ainsi, le moteur d'inférence demeure:

- 1: *Il est plus ou moins vrai d'affirmer que généralement si un sujet a été diagnostiqué positivement alors il réussira le cours «Calcul I».*

L'absence d'une règle pour établir un pronostic de l'échec est un phénomène que l'on retrouve également avec un modèle de régression linéaire. En effet, un tel modèle fait passer une droite pour minimiser la somme des erreurs au carré et l'un des paramètres du modèle est l'ordonnée à l'origine. Or, une valeur non négligeable pour ce paramètre signifie que pour une valeur faible de la variable indépendante «x» nous retrouvons une valeur moyenne pour la variable dépendante «y». Autrement dit, si le diagnostic est l'incapacité (x faible) on n'a pas un pronostic de l'échec (y faible) ; il faudrait plutôt chercher à expliquer l'échec en combinant d'autres attributs un peu comme on le ferait avec un modèle de régression où on ajouterait des variables pour expliquer la variance au critère.

Le nouvel ensemble diagnostic X pour la cohorte de 1999 se caractérise désormais par un taux d'appartenance de 0,50 et une entropie de 0,25. En effet, le taux d'appartenance a progressé de 0,37 à 0,50 car on a abaissé le score de césure.

Quant au tableau XXIV (voir section 4 de l'analyse de données), il représente le tableau de contingence illustrant l'exactitude du pronostic de réussite dans le cours «Calcul I» pour les deux cohortes étudiées. Encore une fois, l'hypothèse nulle a été acceptée et nous concluons que généralement il est possible d'établir un pronostic de réussite pour le cours «Calcul I» en interprétant le score à l'épreuve diagnostique à travers le modèle pronostique.

Résumons-nous, le modèle pronostique permet de prédire la réussite dans un programme mais non l'échec. Donc, on ne peut pas l'appliquer directement dans le cadre du modèle pour un processus d'enseignement différencié que nous avons à la figure 1 (voir la section 2 de la problématique) . Par conséquent, il

faudrait l'intégrer plutôt dans une analyse d'appartenance avec d'autres attributs intervenants qui seraient susceptibles d'influencer l'échec.

Le modèle pronostique ajusté comprend désormais un standard de performance de 44,8 et un standard de contre-performance de 23,2. En d'autres mots, cela signifie qu'un sujet ayant obtenu un score supérieur ou égal à 45 sur 60 recevra un diagnostic de capacité alors que celui dont le score est inférieur ou égal à 23 sur 60 se verra décerner un diagnostic d'incapacité. Et un sujet avec un score entre ces deux limites de 23 et 45 aura un diagnostic incertain.

## Conclusion

La société du XXI<sup>e</sup> siècle a-t-elle vraiment le choix de rendre le plus efficace possible son système d'éducation puisque les attentes sociales et économiques requièrent de plus en plus une réussite scolaire pour le plus grand nombre ? En ce sens, le modèle du processus d'enseignement différencié constitue une avenue plausible pour augmenter la réussite scolaire puisqu'il permet d'orienter l'élève ou l'étudiant vers un enseignement adapté à son niveau. Or, cette tâche d'orientation requiert l'utilisation d'un instrument de mesure dont la validité pronostique envers le critère de réussite ait été éprouvée. La présente recherche vise donc à élaborer un modèle pronostique pour une mesure référée à un critère de réussite en recourant à la théorie des ensembles flous. Les résultats obtenus par l'application de cette théorie aux données fournies par l'École Polytechnique de Montréal vont nous permettre de: souligner les forces du modèle obtenu, d'identifier les défis pour améliorer ce modèle, de reconnaître les limites d'utilisation et de proposer d'autres situations où l'on peut l'appliquer.

Nous pouvons distinguer cinq principales forces pour ce modèle. Premièrement, nous retrouvons l'avantage où il n'est plus nécessaire d'assumer que les données proviennent d'une mesure effectuée avec une échelle d'intervalle et par corollaire, on n'a pas besoin non plus d'assumer que la distribution des scores épouse la forme d'une courbe normale. Deuxièmement, le modèle permet de prédire la réussite et l'échec dans un cours dispensé quelques mois après l'administration de l'épreuve diagnostique ; le pronostic sur la réussite dans un programme peut aussi se faire à plus long terme mais avec une plus faible certitude. Troisièmement, la zone d'incertitude du modèle vient compenser une fidélité imparfaite pour l'instrument de mesure et on peut utiliser les données issues de son administration sans pour cela passer par un processus complexe de standardisation. Quatrièmement, l'interprétation des résultats est facilitée par le fait que les valeurs vraies pour les relations d'implication peuvent être traduites directement dans un raisonnement verbal en recourant aux prédicats qualitatifs.

Enfin, le score à un instrument de mesure est traduit en une donnée qui peut être traitée conjointement avec des données qualitatives pour produire des règles d'inférence moins floues et ce, afin d'établir un pronostic de réussite ou d'échec.

Aussi, nous pouvons identifier trois défis à relever pour améliorer le modèle. Premièrement, il est difficile d'obtenir un pronostic de l'échec mais l'ajout d'attributs significatifs dans le moteur d'inférence pourrait sans doute contribuer à une amélioration des résultats ; par exemple, au niveau postsecondaire on pourrait représenter par un ensemble flou supplémentaire le travail rémunéré des étudiants qui constitue, selon certaines études, un facteur de l'échec scolaire. Et deuxièmement, il faut pouvoir effectuer une analyse d'items afin d'améliorer la qualité de l'instrument de mesure ; or, en représentant les réponses de chaque sujet à un item comme un ensemble il est possible de trouver la valeur vraie du degré d'inclusion de cet ensemble dans l'ensemble flou constitué par les scores pour ces mêmes sujets. Enfin, un troisième défi consiste à déterminer les limites acceptables de la zone d'incertitude pour une fidélité imparfaite de l'instrument.

Il faut en plus reconnaître deux limites à ce modèle. Premièrement, nous sommes habitués à raisonner avec précision et certitude, or le modèle nous amène plutôt à raisonner en logique floue. Deuxièmement, les paramètres du modèle ne peuvent être généralisés à une très grande population de sujets mais doivent plutôt être ajustés au fil des ans.

Un modèle pronostique permettant d'établir un pronostic sur la réussite et l'échec d'un sujet peut être appliqué dans beaucoup d'autres situations en éducation. Pour l'admission dans des programmes universitaires contingentés, on utilise des seuils pour la cote R de rendement au collégial. Plusieurs cégeps font passer une épreuve diagnostique en français à leurs nouveaux étudiants afin de les orienter ou non vers des cours de mise à niveau. Dans les dernières années du secondaire, le nouveau régime pédagogique prévoit d'orienter les élèves dans des voies générale et semi-spécialisée. Pour le premier cycle du secondaire, on propose de plus en plus des cheminements adaptés aux élèves ; par

exemple, on pourrait mettre l'emphase sur la compétence en lecture pour les élèves susceptibles de décrocher. Au primaire aussi, on a besoin de poser un diagnostic sur chaque élève au début d'un cycle afin d'offrir si nécessaire les services d'adaptation scolaire adéquats ou de différencier l'enseignement.

Dans le débat public sur la réforme en éducation, nous constatons que plusieurs intervenants remettent en cause l'évaluation des apprentissages. Toutefois, en regardant de plus près les énoncés pour les échelles de niveaux de compétence, on y retrouve des prédicats qualitatifs associés à des critères d'évaluation ; la théorie des ensembles flous pourrait sans doute contribuer à rendre les communications sur la performance de l'élève plus cohérentes.

Vu l'importance de prédire autant l'échec que la réussite en éducation, il devient important de poursuivre les recherches pour valider ce modèle pronostique. Dans la présente recherche, nous n'avons pas tenu compte du contexte d'administration de l'épreuve diagnostique qui peut influencer le pronostic de l'échec. En effet, aucune décision de certification, de sélection ou d'orientation n'était prise suite au score obtenu par un sujet et cela peut atténuer la motivation à répondre. Par conséquent, il est plausible que le diagnostic négatif, posé à partir de l'interprétation des scores à l'épreuve, sous-estime la capacité réelle du sujet.

Aussi, il faut valider ce modèle pronostique en le comparant à une autre approche telle par exemple l'approche bayésienne. Alors, on aurait recours à une interprétation critériée classique des scores et on établirait deux probabilités conditionnelles : la probabilité de réussite dans le programme étant donné une décision de maîtrise à l'épreuve et la probabilité d'échec dans le programme étant donné une décision de non-maîtrise à l'épreuve. Si les résultats étaient comparables à ceux obtenus alors, on pourrait conclure que ce n'est pas le modèle lui-même qui ne permet pas d'établir le pronostic pour l'échec d'un sujet.

## **Bibliographie**

- Allen, M. J. & Yen, W. M. (1979). *Introduction to Measurement Theory*.  
Monterey : Brooks Cole Publishing.
- Ammar, S., Bifulco, R., Duncombe & W., Wright, R. (2000). Identifying  
Low-Performance Public Schools. *Studies in Educational  
Evaluation, 26*, 259-287.
- Andrich, M. (1978). Relationships Between the Thurstone and Rasch  
Approaches to Item Scaling. *Applied Psychological  
Measurement, 2*, 449-460.
- Auger, R. & Séguin, S.P. (1992). Le testing adaptatif avec interprétation  
critérielle de praticabilité du TAM pour l'évaluation sommative  
des apprentissages au Québec. *Mesure et évaluation en  
éducation, 15*, 103-145.
- Baillargeon, G. (1989). *Probabilités, statistiques et techniques de régression*.  
Trois-Rivières : Les Éditions SMG.
- Bassey, M. (1998). *Fuzzy generalisation : an approach to building an  
educational theory*. Communication présentée à la British  
Educational Research Association Annual Conference de  
Queen's University of Belfast, Irlande du Nord. Document  
consulté le 17 décembre 2001 sur le site internet :  
[www.leeds.ac.uk/educol/documents/0000000801.htm](http://www.leeds.ac.uk/educol/documents/0000000801.htm).
- Bégin, J. (1993). *Introduction à la psychométrie*. Document inédit, Université  
du Québec à Montréal.
- Bertrand, R. (1988). Pourquoi de nouvelles théories de la mesure. *Mesure et  
évaluation en éducation, 11*, 5-26.
- Berk, R.A. (Éd.). (1980). *Criterion-Referenced Measurement : The State of the  
Art*. Baltimore : The John-Hopkins University Press.

- Blais, J.G. & Laurier, M. (1997). La détermination de l'unidimensionalité de l'ensemble des scores à un test. *Mesure et évaluation en éducation*, 20, 65-90.
- Blais, J.G. & Fournier, M. (1997). Réflexions sur la modélisation statistique des scores aux test. *Mesure et évaluation en éducation*, 19, 69-94.
- Blais, J.G. (2003). Synthèse sur les fondements de la mesure et sur la contribution de la modélisation moderne. Dans Blais, J.G. et Raïche, G. (Éds), *Regards sur la modélisation de la mesure en éducation et en sciences sociales*. (pp. 9-32). Québec: Les Presses de l'Université Laval.
- Bloom, B. S. (1982). *Taxonomie des objectifs pédagogiques: 1 domaine cognitif*. Traduit de l'américain par Lavallée, M. Québec : Presses de l'Université du Québec.
- Brennan, R.L. (1980). Applications of Generalizability Theory. Dans Berk, R. (Éd.), *Criterion-Referenced Measurement : The State of the Art*. (pp.186-232). Baltimore : John Hopkins University.
- Brown, F.G.(1980). *Guidelines for Test Use: A commentary on the Standards for Educational and Psychological Tests*. Washington DC : National Council on Measurement in Education.
- Cardinet, J. (1986). *Évaluation scolaire et mesure*. Bruxelles : Université de Boeck.
- Carver, R.P. (1974). Two Dimensions of Tests : Psychometric and Edumetric. *American Psychologist*, July 1974, 512-518.
- Charlot, B. (1995). *Les sciences de l'éducation, un enjeu, un défi*. Paris : Ed. ESF.
- Chou, Y.-L. (1975). *Statistical Analysis, second edition*. NewYork : Éd. HRW.
- Commission de l'éducation, Assemblée nationale (2005). *Audition des dirigeants des établissements d'enseignement universitaire en application de la Loi sur les établissements d'enseignement universitaire*. Québec : Auteur.

- Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements*. New York : Éd. John Wiley & Sons.
- Crowther, C. S. & Batchelder, W. H. (1995). A measurement theoretic analysis of the fuzzy logic model of perception. *Psychological Review*, 102, 396-408.
- Dassa, C. & Laurier, M.D. (2003). Le diagnostic pédagogique comme moyen d'informer. Dans Laurier, M.D. (Éd.). *Évaluation et Communication* (pp.103-130). Montréal : Les éditions Québecor.
- Debays, B. (1998). *De l'épistémologie au politique : l'unité de la pensée de Karl Popper*. Cahiers d'épistémologie publiés par l'Université du Québec à Montréal, cahier no. 9807.
- De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation*. Paris : Presses universitaires de France.
- De Landsheere, V. (1988). *Faire réussir faire échouer, la compétence minimale et son évaluation*. Paris : Presses universitaires de France.
- Desmarais, M.C. & Pu, X. (2004). *Computer Adaptive Testing with Bayesian Networks: A comparison with IRT*. Netherland : Kluwer Academic Publishers.
- Dictionnaire Larousse (1991). *Le Petit Larousse en couleurs*. Paris : Auteur.
- Dubois, P.H. (1970). *A History of Psychological Testing*. Boston : Allyn and Bacon.
- Echternacht, G. (1980). *New directions for testing and measurement: Measurement Aspects of Title 1 Evaluations*. New York : Éd. Jossey-Bass Inc.
- École Polytechnique de Montréal (2004). *La structure des programmes*. Document consulté le 8 mars 2004 sur le réseau internet à l'adresse : [www.polymtl.ca/etudes/bc/structure.php](http://www.polymtl.ca/etudes/bc/structure.php).

- École Polytechnique de Montréal (2005). *Règlements généraux et particuliers des études supérieures*. Document consulté le 27 mai 2005 sur le réseau internet à l'adresse :  
[www.polymtl.ca/sg/docs\\_officiels/2510re03.htm](http://www.polymtl.ca/sg/docs_officiels/2510re03.htm)
- Embretson, S.E. (1999). Issues in the Measurement of Cognitive Abilities. Dans Embretson, S.E. & Hershberger, S.L. (Éds), *The New Rules of Measurement*. (pp.1-15). Princeton : Lawrence Erlbaum Associates, Publishers.
- Flood, R. L. & Carson, E. R. (1990). *Dealing with Complexity*. New York : Plenum Press.
- Fourali, C. (1997). Using Fuzzy Logic in Educational Measurement: The Case of Portfolio Assessment. *Evaluation and Research in Education*, 11, 129-147.
- Freissinet, C., Cunge, J., Erlich, M. & Vauclin, M. (1998). *Modélisation mathématique et théorie des sous-ensembles flous. Application à l'estimation des imprécisions dans la modélisation du devenir des produits phytosanitaires dans les eaux souterraines*. CNFGG-Rapport quadriennal 1995-1998.
- Gitomer, D.H. & Rock, D.(1993). Addressing Process Variables in Test Analysis. Dans Frederiksen, N., Mislevy, R.J. & Bejar, I. I., (Éds), *Test Theory for a new Generation of Tests*, (pp.243-268). Princeton : Éd. Lawrence Erlbaum Associates, Hillsdale.
- Glass, G.V. & Hopkins, K.D. (1984). *Statistical Methods in Education and Psychology*. Princeton: Éd. Prentice-Hall.
- Grasland, C. (1997). *À la recherche d'un cadre théorique et méthodologique pour l'étude des maillages territoriaux*. Communication présentée aux Entretiens Jacques Cartier, Lyon, France, 8-10 déc. 1997.
- Hansen, B.K. (2000). *Analog Forecasting of Ceiling and Visibility using Fuzzy Sets*. 2<sup>nd</sup> Conference on Artificial Intelligence, American Meteorological Society, pp.1-7.
- Hashway, R.M. (1998). *Assesment and Evaluation of Developmental Learning*. Connecticut : Praeger Publishers.

- Hawkes, L.W., Derry, S. & Kandel, A. (1992). Fuzzy Expert Systems for an Intelligent Computer-Based Tutor. Dans Kandel, A. (Éd.), *Fuzzy Expert Systems*, (pp.237-258). Boca Raton : CRC Press.
- Heiser, W. J. (1997). Cluster differences scaling with a within-clusters loss component and a Fuzzy Successive Approximation Strategy to avoid local minima. *Psychometrika*, 62, 63-83.
- Huang, M. J. (1999). A Fuzzy Student Modeling with two intelligent agents. *Journal on Educational Computing Research*, 21, 99-113.
- Huitt, W. (1996). *Measurement and Evaluation : Criterion-versus Norm-Referenced Testing*. Document consulté le 20 janvier 2005 sur le réseau internet à l'adresse :  
<http://chiron.valdosta.edu/whuitt/col/measeval/cnmref.html>
- Jameson, J. (1996). Numerical Uncertainty Management in User and Student Modeling: An Overview of Systems and Issues. *User Modeling and User Adapted Interaction*, 5.
- Johnson, B. & Onwuegbuzie, A. (2004). Mixed Methods Research : A Research Paradigm Whose Time has gone. *Educational Researcher*, 33, 14-26.
- Kandel, A. (Éd.). (1992). *Fuzzy Expert Systems*. Florida : CRC Press inc.
- Katz, S. & Lesgold, A. (1992). *Use of Fuzzy Modeling Techniques in a Coached Practice Environment for Electronics Troubleshooting*. EDRS Availability : Microfiche, Paper.
- Klir, G. J. & Yuan, B. (1995). *Fuzzy Sets and Fuzzy Logic : Theory and Applications*. New Jersey: Éd. Prentice Hall.
- Kosko, B. (1993). *Fuzzy thinking, The New Science of Fuzzy Logic*. NewYork:Éd. Hypérion.
- Laperrière, A. (1997). Les critères de scientificité des méthodes qualitatives. Dans Poupart, Deslauriers, Groulx, Laperrière, Mayer, Pires (Éds), *La recherche qualitative : enjeux épistémologiques et méthodologiques*, (pp.365-399. Boucherville : Éd. Gaëtan Morin.
- Laurencelle, L. (1997). La capacité discriminante d'un instrument de mesure. *Mesure et évaluation en éducation*, 20, 25-40.

- Legendre, R. (1988). *Dictionnaire actuel de l'éducation*. Paris : Éd. Larousse.
- L.R.Q. *Lois et Règlements du Québec*. Éditeur officiel du Québec.
- Lohman, D.F. & Ippel, M.J. (1993). Cognitive Diagnosis : From Statistically Based Assessment Toward Theory-Based Assessment. Dans Frederiksen, N., Mislevy, R.J. & Bejar, I. I., (Éds), *Test Theory for a new Generation of Tests*. (pp.41-71). Princeton : Éd. Lawrence Erlbaum Associates.
- Lopes, A.L. & Lanzer, E.A. & Barcia, R. (1997). *Fuzzy Cross-Evaluation of the Performance of Academic Departments within a University*. Federal University of Santa Carina, Brazil. Document consulté le 6 septembre 2002 sur le réseau internet à l'adresse : [www.cirpa-acpri.ca/prevConferences/toronto97/97\\_proceedings/lopes.html](http://www.cirpa-acpri.ca/prevConferences/toronto97/97_proceedings/lopes.html)
- Lord, F. M. & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. NewYork : Éd. Addison-Wesley Publishing Company.
- Lossin, S.-H. (2005). *Decision Making with Imprecise and Fuzzy Probabilities – a Comparison*. 4th International Symposium on Imprecise Probabilities and Their Applications, Pittsburg, PA.
- Marcoulides, G.A. (1999). Generalizability Theory: Picking Up Where the Rasch IRT Model Leaves Off? Dans Embretson, S.E. & Horshberger, S.L. (Éds), *The New Rules of Measurement*. (pp.129-152). Princeton : Éd. Lawrence Erlbaum.
- Martin, P. (1986). *Élaboration et implantation d'une approche systémique d'enseignement en électrotechnique au niveau collégial I*. Mémoire de maîtrise présenté à l'Université du Québec à Montréal.
- McDonald, R. P. (1999). *Test Theory*. Princeton : Lawrence Erlbaum Associates, Publishers.
- MEQ (2003). *Indicateurs de l'éducation*. Gouvernement du Québec.
- MEQ (2004-1). *La relance au collégial en formation technique-2003*. Gouvernement du Québec.

- MEQ (2004-2). *Le cheminement des élèves, du secondaire à l'entrée à l'université*. Gouvernement du Québec.
- MEQ (2004-3). *La réussite des garçons, des constats à mettre en perspective*. Gouvernement du Québec.
- Messick, S. (1993). *Foundations of validity : meaning and consequences in psychological assessment*. Princeton: Educational Testing Service.
- Miles, M.B. et Huberman, A.M. (1991). *Qualitative Data Analysis*. Palo Alto : Sage.
- Mislevy, R. J. (1993). Foundations of a New Test Theory. Dans Frederiksen, N., Mislevy, R.J. & Bejar, I. I. (Éds), *Test Theory for a new Generation of Tests*, (pp.19-39). Princeton : Lawrence Erlbaum Associates.
- Nedelsky, L. (1954). Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement*, 14, 3-19.
- Plaisance, E. et Vergnaud, G. (1993). *Les sciences de l'éducation*. Paris : Éd. La Découverte.
- Popham, W. J. (1981). *Modern Educational Measurement*. Princeton : Prentice-Hall.
- Pryor, R. & Hesketh (1989). Making things clearer by making them fuzzy : Counseling Illustrations of a Fuzzy Graphic Rating Scale. *Career Development Quarterly*, 38, 135-147.
- Ragin, C. C. (1987). *The Comparative Method : Moving beyond Qualitative and Quantitative Strategies*. Los Angeles : University of California Press.
- Ragin, C. C. (2000). *Fuzzy-Set Social Science*. Chicago : The University of Chicago Press.
- Ralston, A. & Reilly, E. D. (1983). *Encyclopedia of Computer Sciences and Engineering*. New York : VNR.
- Rasmani, K.A. (2002). *A Data-Driven Fuzzy Rule-Based Approach for Student Academic Performance Evaluation*. Thèse de doctorat, Université d'Édimbourg, UK.

- RHDC (2004). *Emploi-Avenir*. Éd. par le ministère des Ressources Humaines et du Développement des Compétences, Gouvernement du Canada.
- St-Yves, A. (1982). *Psychologie de l'apprentissage-enseignement*. Québec : Presses de l'Université du Québec.
- Samejima, F. (1995). A Cognitive Diagnosis Method Using Latent Trait Models: Competency Space Approach and its relationship with DiBello and Stout's Unified Cognitive-Psychometric Diagnosis Model. Dans Nichols, P.D., Chipman, S.F. & Brenman, R.L. (Éds), *Cognitively Diagnostic Assessment*. (pp.391-410). Princeton: Lawrence Erlbaum.
- Sangali, A. (1998). *The Importance of Being Fuzzy*. Princeton : University Press.
- Schneider, M. et Kandel, A. (1992). General Purpose Fuzzy Expert Systems intelligents. Dans Kandel, A. (Éd.), *Fuzzy Expert Systems*. (pp.23-42). Boca Raton : CRC Press.
- Shaycoft, M. F. (1979). *Handbook of criterion-referenced testing : development, evaluation, and use*. NewYork : Garland STPM Press.
- Siegel, S. (1956). *Non-parametric Statistics*. NewYork : McGraw-Hill.
- Smithson, M. (1989). *Ignorance and Uncertainty*. NewYork : Springer-Verlag.
- Smithson, M. (2005). Fuzzy Set Inclusion. *Sociological Methods & Research*, 33, 431-461.
- Snow, R.E. & Lohman, D.F. (1993). Cognitive Psychology, New Test Design, and New Test Theory : An introduction. Dans Frederiksen, N., Mislavy, R.J. & Bejar, I. I. (Éds), *Test Theory for a new Generation of Tests Test Theory for a new Generation of Tests* (pp.1-17). Princeton : Lawrence Erlbaum Associates.
- Statistique Canada (2001). *Population âgée de 15 ans et plus selon le plus haut niveau de scolarité atteint, provinces et territoires*. Gouvernement du Canada.
- Statistique Canada (2005). *Dernier communiqué de l'Enquête sur la population active (4 février 2005)*. Gouvernement du Canada.

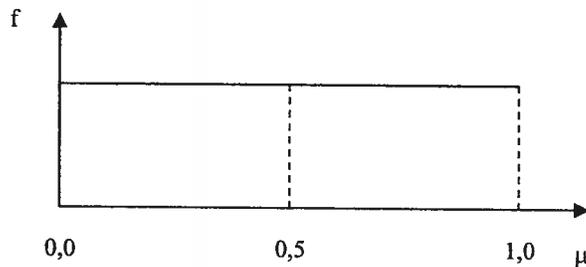
- Subkoviak, M.J. (1980). Decision-Consistency Approaches. Dans Berk, R.A. (Éd.), *Criterion-Referenced Measurement : The State of the Art*. (pp.129-185). Baltimore : The John Hopkins University Press.
- Tatsuoka, K.K. (1995). Architecture of Knowledge Structures and Cognitive Diagnosis: A Statistical Pattern Recognition and Classification Approach. Dans Nichols, P.D., Chipman, S.F. & Brenman, R.L. (Éds), *Cognitively Diagnostic Assessment*, (pp.327-360). Princeton : Lawrence Erlbaum.
- Tourneur, Y. (1989-1). Le traitement des facettes cachés dans l'application des dispositifs expérimentaux aux situations scolaires. *Mesure et évaluation en éducation*, 12, 29-52.
- Tourneur, Y. (1989-2). Les domaines d'application de la théorie de la généralisabilité. *Mesure et évaluation en éducation*, 12, 53-68.
- Van der Maren, J.M. (1996). *Méthodes de recherche pour l'éducation*. Montréal: Les Presses de l'Université de Montréal.
- Violato, C. & McDougall, D. & Marini, A. (1992). *Educational Measurement and Evaluation*. Des Moines : Kendall-Hunt.
- Von Altrock, C. (2000). *Fuzzy Logic and NeuroFuzzy in Business and Finance*. NewYork : Prentice-Hall.
- Voisard, A.M. (2004). *Le français à l'école se porte mal. Des tests à en perdre son français*. Le journal «La Presse» du 23 octobre 2004, Montréal.
- Wright, B.D. (1999). Fundamental Measurement for Psychology. Dans Embretson, S.E. & Hershberger, S.L. (Éds), *The new rules of measurement*, (pp.65-104). Princeton: Lawrence Erlbaum Associates.
- Yi, D., Lewis, C. & Pashley, P. J. (1993). Computerized Mastery Testing Using Fuzzy Set Decision Theory. *Applied Measurement in Education*, 6, 181-193.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8, 338-353.
- Zwick, R. (2002). *Fair game? The use of Standardized Admissions Tests in Higher Education*. NewYork: Routledge Falmer.

## Annexe I

### L'entropie d'une distribution uniforme

#### *1. La distribution uniforme de l'appartenance à un ensemble flou*

Une distribution uniforme signifie que pour chaque degré d'appartenance à un ensemble A quelconque il y a le même nombre d'éléments. La figure ci-dessous illustre la forme de cette distribution uniforme.



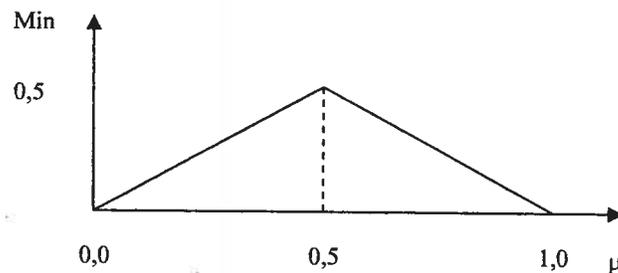
La relation pour calculer l'entropie est:

$$E(A) = \frac{|A \cap A'|}{|A \cup A'|} = \frac{\sum \text{Min}(\mu_A, 1-\mu_A)}{\sum \text{Max}(\mu_A, 1-\mu_A)}$$

Pour une distribution continue, la sommation devient une intégrale. Alors un point de discontinuité existe à  $\mu=0,5$ . En effet, la fonction Min prend la valeur de  $\mu_A$  lorsque le degré d'appartenance se situe entre 0 et 0,5 et de  $1-\mu_A$  lorsqu'il se situe entre 0,5 et 1,0. Il en va de même pour la fonction Max qui prend la valeur de  $1-\mu_A$  entre 0 et 0,5 et de  $\mu_A$  entre 0,5 et 1,0.

#### *2. L'intégrale du numérateur*

Le numérateur prend la forme de la figure ci-dessous.

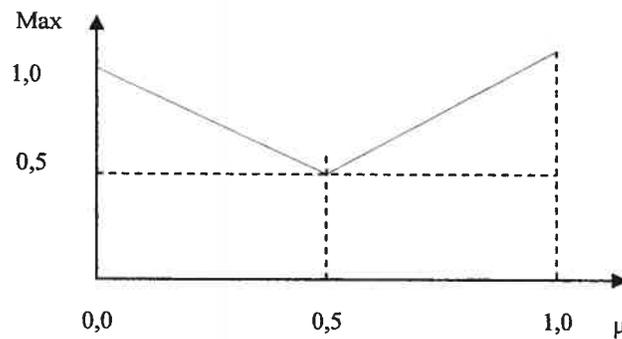


Nous avons:

$$\begin{aligned} \int \text{Num} &= \int_0^{0,5} \mu_A d\mu + \int_{0,5}^1 1-\mu_A d\mu \\ &= \frac{(0,5 \times 0,5)}{2} + \frac{(0,5 \times 0,5)}{2} \\ &= 0,25 \end{aligned}$$

### 3. L'intégrale du dénominateur

Le dénominateur prend plutôt la forme de la figure ci-dessous.



Nous avons:

$$\begin{aligned} \int \text{dénom} &= \int_0^{0,5} 1-\mu d\mu + \int_{0,5}^1 \mu d\mu \\ &= \frac{(0,5 \times 1)}{2} + \frac{(0,5 \times 0,5)}{2} + \frac{(0,5 \times 0,5)}{2} \\ &= 0,75 \end{aligned}$$

### 4. L'entropie de la distribution uniforme

Nous reprenons la formule originale en remplaçant les valeurs pour le numérateur et le dénominateur.

$$\begin{aligned} E(A) &= \frac{\int \text{num}}{\int \text{dénom}} \\ &= 0,25 / 0,75 \\ &= 0,33 \quad \text{c.q.f.d} \end{aligned}$$

## Annexe II

### Devis pour les données de la recherche à l'École Polytechnique

#### *1. Le but*

Le but de la recherche est de prédire la réussite des sujets dans leur programme à partir des résultats qu'ils ont obtenus au test diagnostique en mathématique administré lors de leur admission.

#### *2. Méthode*

Élaborer un modèle à partir de la cohorte de 1997 qui sera suivie jusqu'au trimestre Été 2003 et appliquer ce modèle à la cohorte de 1998 qui sera suivie jusqu'au trimestre Été 2004.

#### *3. Requête de données*

- *Support et structure*

Les données devraient être sous la forme d'un fichier Excel (.xls). Il faudrait une feuille distincte pour chaque cohorte. Chaque sujet, identifié par son matricule, spécifie une et une seule ligne du tableur. Chaque attribut spécifie une et une seule colonne.

- Cohorte de 1997

Pour chaque sujet admis au trimestre Automne 1997 et qui a passé le test diagnostique de math en juin 1997, il faut les données pour les attributs suivants:

Matricule

Date de naissance

Sexe

Diplôme d'accueil (ex : DEC gén. Sciences <200.B0>)

Cote au cours ING1005, Calcul I (ex : A+)

Nature du programme suivi (ex : BIELE pour Bacc. en génie électrique)

Si possible : Inscription et réussite du cours de mise à niveau en math (ex : non-inscrit, ou inscrit et réussite)

Dernier trimestre inscrit dans le programme (ex : Hiver 2002)

Diplôme obtenu (B. Sc.A. ou B. ing. : vrai ou faux)

Réponses aux 60 items du test diagnostique en math

- Cohorte de 1998

Pour chaque sujet admis au trimestre Automne 1998 et qui a passé le test diagnostique de math en juin 1998, il faut exactement les mêmes données que pour la cohorte 1997.

### Annexe III

#### Apparition des nouveaux sigles (cours) et correspondance avec les anciens sigles

Ancien sigle	Années									Titre du nouveau cours
	1997-1998	1998-1999	1999-2000	2000-2001	2001-2002	2002-2003	2003-2004	2004-2005	2005-2006	
327		MTH2301								Méthode statistique en ingénierie
105			ING1001		Fin					Calcul I
115			ING1002		Fin					Calcul II
135 ou (125+126)			ING1003							Équations différentielles
421				MTH2110						Méthodes mathématiques de la physique I
423				MTH2111						Méthodes mathématiques de la physique II
426 ou (420+424)				MTH2120						Analyse appliquée
215				MTH2200						Analyse numérique
340				MTH2305						Probabilités pour ingénieurs
					ING1005					Calcul I, 2 cr
ING1001, 3 cr					ING1006					Algèbre linéaire, 2cr
ING1002, 3 cr					ING1007					Calcul II, 2 cr
MTH2200						MTH2210				Calcul scientifique pour ingénieurs
MTH2301								MTH2302		Probabilités et statistiques pour l'ingénieur civil