Revue des sciences de l'éducation

La généralisabilité des seuils de performance établis par des appréciateurs : étude de trois procédés

Gérard Scallon

Volume 7, Number 1, Winter 1981

URI: https://id.erudit.org/iderudit/900313ar
DOI: https://doi.org/10.7202/900313ar

See table of contents

Publisher(s)

Revue des sciences de l'éducation

ISSN

0318-479X (print)
1705-0065 (digital)

Explore this journal

Cite this article

Scallon, G. (1981). La généralisabilité des seuils de performance établis par des appréciateurs : étude de trois procédés. *Revue des sciences de l'éducation*, *7*(1), 3–21. https://doi.org/10.7202/900313ar

Article abstract

This study compared three procedures in which the test evaluators determine the threshold of performance on a ten item exam, namely, the procedures of Angoff, Ebel and Nedelsky. Analysis and variance results showed that the thresholds varied from one evaluator to another and from one procedure to another. However, the concordance among evaluators using the Nedelsky procedure was the highest. Moreover, according to the generalizability study, the differentiation between the questions proved to be reliable for all evaluators and all procedures. Although this research has substantiated the subjectivity characteristic of procedures in which evaluators determine the threshold of performance, the results indicate areas of further research both in item analysis and in the concept of minimal competence.

Revue des sciences de l'éducation, vol. VII, no 1, hiver 1981

La généralisabilité des seuils de performance établis par des appréciateurs : étude de trois procédés

Gérard Scallon*

Résumé — Cette recherche avait pour but de comparer trois procédés faisant appel à des appréciateurs pour déterminer le seuil de performance à un examen de dix questions : les procédés d'Angoff, d'Ebel et de Nedelsky. D'après les résultats de l'analyse de la variance, les seuils obtenus ne sont pas comparables d'un procédé à l'autre ni d'un appréciateur à l'autre. Toutefois, la concordance entre les appréciateurs est apparue la plus élevée dans le cas du procédé de Nedelsky. De plus, d'après l'étude de généralisabilité, la différenciation entre les questions s'est avérée fiable et ce, quels que soient les appréciateurs et les procédés. Bien que cette recherche ait mis en évidence la subjectivité qui caractérise la détermination d'un seuil de performance par des appréciateurs, les résultats obtenus suggèrent des avenues à explorer, tant du côté de l'analyse des questions d'un examen que du côté de la notion de compétence minimale.

Abstract — This study compared three procedures in which the test evaluators determine the threshold of performance on a ten item exam, namely, the procedures of Angoff, Ebel and Nedelsky. Analysis and variance results showed that the thresholds varied from one evaluator to another and from one procedure to another. However, the concordance among evaluators using the Nedelsky procedure was the highest. Moreover, according to the generalizability study, the differentiation between the questions proved to be reliable for all evaluators and all procedures. Although this research has substantiated the subjectivity characteristic of procedures in which evaluators determine the threshold of performance, the results indicate areas of further research both in item analysis and in the concept of minimal competence.

Resumen — Esta investigación tenía por objeto comparar tres procedimientos mediante la utilización de apreciadores que determinen el nivel de desempeño en un exámen de diez preguntas : los procedimientos de Angoff, de Ebel y de Nedelsky. Según los resultados del análisis de la varianza, los niveles obtenidos no son comparables ni de un procedimiento a otro, ni de un apreciador a otro. De todas formas la concordancia entre los apreciadores fué mas elevada en el caso del procedimiento de Nedelsky. Además, según el estudio de generalidad, la diferenciación entre las preguntas se mostraron comparables sean cuales fueren los apreciadores y los procedimientos. Aún cuando esta investigación puso en evidencia la subjetividad que caracteriza la determinación de un nivel de desempeño por los apreciadores, los resultados obtenidos sugieren nuevas vias a explorar, tanto del lado del análisis de las preguntas de un exámen, como del lado de la noción de la competencia mínima.

Zusammenfassung — Das Ziel dieser Untersuchung ist ein Vergleich zwischen drei Verfahren, welche Bewerter benützen, um die Leistungsschwelle bei einer aus zehn (10) Fragen bestehenden Prüfung zu bestimmen : die Verfahren von Angoff, Ebel und Nedelsky. Den Ergebnissen der Varianzanalyse nach ist es unmöglich, die gefundenen Schwellen zu vergleichen, weder unter den einzelnen Verfahren, noch

* Scallon, Gérard : professeur, Université Laval.

unter den einzelnen Bewertern. Dennoch schien im Verfahren von Nedelsky die Ubereinstimmung unter den Bewertern am höchsten zu sein. Gemäss der Studie über die Möglichkeit der Verallgemeinerung erscheint die Differenzierung unter den Fragen zuverlässig, und zwar unabhängig von den Bewertern und Verfahren. Obgleich diese Untersuchung die Subjektivität aufzeigt, die einer Bestimmung von Leistungsschwellen durch Bewerter zueigen ist, eröffnen die Resultate neue Wege, sowohl bei der Analyse der Prüfungsfragen als auch beim Begriff der Mindestkompetenz.

Introduction

La notion de mesure centrée sur un critère (criterion-referenced measurement) existe déjà depuis plusieurs années dans le domaine de l'évaluation pédagogique. Le développement de programmes d'enseignement individualisé et l'avènement de la pédagogie de maîtrise ont fait de ce type de mesure une cible d'intérêt tant pour l'évaluation formative que pour l'évaluation sommative des étudiants.

Le test centré sur un critère est reconnu aujourd'hui comme étant l'instrument de choix pour intervenir de façon continue dans la progression de chaque étudiant dans ses apprentissages. Il s'est également imposé à d'autres contextes d'évaluation, là où il s'agit par exemple d'établir la compétence d'un individu au sortir d'un programme d'études. La mesure centrée sur la compétence (competency-based measurement) est devenue une caractéristique de nombreux programmes de formation professionnelle aux États-Unis comme en témoigne l'ouvrage récent de Bunda et Sanders (1979).

Quel que soit le contexte de l'évaluation, la détermination du seuil de performance[1] est l'un des problèmes majeurs que doivent résoudre ceux qui utilisent des instruments de mesure conçus dans une perspective d'interprétation « critériée » des résultats. De nombreux procédés ont été suggérés à titre de solutions possibles et les écrits sur le sujet contiennent d'excellentes synthèses, entre autres celle de Millman (1973), celle de Meskauskas (1976) et celle de Hambleton et Eignor (1979).

Un certain nombre de procédés exigent la collaboration directe de personnes devant agir comme appréciateurs du contenu et de la difficulté des questions d'un test. D'autres procédés se basent sur des résultats obtenus auprès de groupes d'étudiants servant de points de repère. Enfin, il existe toute une gamme de procédés qui sont caractérisés par une approche rationnelle ou théorique centrée sur la théorie de la décision et qui sont accompagnés d'une évaluation des conséquences possibles d'un seuil de performance à choisir ou d'un nombre minimum de questions à inclure dans un examen.

Le dénominateur commun à tous ces procédés est la part d'arbitraire ou de subjectivité qu'ils impliquent à divers degrés. Les procédés qui apparaissent le plus directement visés par cette affirmation sont fort probablement ceux qui font appel à des appréciateurs. C'est du moins ce qui ressort d'un texte récent de Glass (1978, p. 249) dans lequel plusieurs procédés suggérés dans les écrits de recherche pour déterminer le seuil de performance ont été sévèrement critiqués.

Le but de la présente recherche est d'examiner et de comparer entre eux trois des procédés suggérés pour déterminer le seuil de performance en faisant appel au jugement d'appréciateurs : celui de Nedelsky (1954), celui d'Ebel (1972) et celui d'Angoff (1971).

Description des procédés étudiés

Les procédés de Nedelsky, d'Ebel et d'Angoff ont, pour point de départ commun, la notion de compétence minimale. En effet, les appréciateurs devant utiliser l'un ou l'autre de ces procédés doivent se faire une idée de ce que doit être un étudiant de compétence minimale. Il s'agit d'un étudiant à peine ou tout juste compétent pour réussir un cours ou un programme d'études.

Le procédé de Nedelsky s'applique à un examen composé de questions à choix de réponses. Pour chaque question l'appréciateur doit d'abord identifier le ou les leurres qu'un 'étudiant de compétence minimale' pourrait reconnaître comme tels et ensuite écrire l'inverse du nombre de choix qui restent. Par exemple, avec une question à cinq choix dont un seul est la bonne réponse, un appréciateur écrirait $\frac{1}{3}$ après avoir identifié deux leurres que, d'après lui, un étudiant de compétence minimale pourrait rejeter comme mauvaises réponses. La somme des inverses pour l'ensemble des questions correspond au seuil de performance établi par l'appréciateur. Avec plusieurs appréciateurs on prendra alors la moyenne des seuils de performance ainsi établis à laquelle s'ajoute un certain nombre de fois le degré de dispersion (écart-type) entre les appréciateurs. Pour comparer le procédé de Nedelsky à d'autres procédés, seule la moyenne des seuils de performance établis par les appréciateurs a été retenue.

Dans le cas du procédé d'Ebel, l'appréciateur doit évaluer les questions d'un examen selon deux dimensions : le degré de difficulté (difficile, moyen, facile) et le degré de pertinence (essentiel, important, acceptable et discutable). On obtient ainsi une grille à deux dimensions composée de douze cases dans lesquelles sont placées les questions. Chaque appréciateur (ou l'ensemble des appréciateurs) doit ensuite écrire dans chaque case la proportion de questions de cette case qu'un étudiant de compétence minimale pourrait réussir. Cette proportion est multipliée par le nombre de questions placées dans cette case par tous les appréciateurs (la même question pouvant être dénombrée plus d'une fois). Un calcul simple permet d'obtenir un produit moyen pour l'ensemble des appréciateurs et des questions qui est le seuil de performance recherché.

Le procédé d'Angoff, le plus simple des trois procédés étudiés dans cette recherche, consiste essentiellement, pour un appréciateur, à estimer directement la probabilité de succès d'un étudiant de compétence minimale et ce, pour chacune des questions du test. La somme des probabilités estimées par un appréciateur, pour l'ensemble des questions, correspond au seuil de performance établi par cet appréciateur. La moyenne des seuils ainsi établis par les appréciateurs est le seuil de performance recherché.

Le procédé de Nedelsky s'applique uniquement à des questions du type « choix de réponses ». Les procédés d'Ebel et d'Angoff conviennent également à ce type de question mais peuvent être utilisés avec des questions où il n'y a pas de choix de réponses.

Recherches antérieures et questions soulevées

Très peu de recherches ont été entreprises au sujet des procédés suggérés pour déterminer le seuil de performance à un test et notamment dans le cas des procédés faisant appel au jugement d'appréciateurs.

Meskauskas et Webster (1975) rapportent avoir essayé le procédé de Nedelsky avec un examen de recertification pour les médecins membres de l'American Board of Internal Medicine. Cet examen s'inscrivait dans le cadre d'une politique de formation continue mise de l'avant par la Commission pour les médecins en exercice ; l'examen a été administré sur une base de volontariat. Aux fins de la recertification, une note minimale de passage était requise et le procédé de Nedelsky a été expérimenté avec les six membres du comité chargé de l'élaboration de l'examen. Les auteurs rapportent que les seuils de performance déterminés par les appréciateurs étaient très différents les uns des autres, ce qui indiquait un faible consensus parmi ces appréciateurs. Le procédé de Nedelsky a dû être abandonné pour faire place à un procédé normatif (Meskauskas et Webster, 1975, pp. 580-581).

Andrew et Hecht (1976) ont réalisé une étude comparative entre le procédé de Nedelsky et celui d'Ebel. Deux groupes de quatre appréciateurs chacun ont été formés avec des personnes ayant participé à la rédaction de 180 questions à choix de réponses devant faire partie d'un examen de certification en sciences de la santé. Chaque groupe a utilisé le procédé de Nedelsky et celui d'Ebel mais l'ordre de présentation des procédés a été inversé entre les groupes. Pour contrôler l'effet d'entraînement ou de mémoire entre les procédés, le procédé d'Ebel a été utilisé avec les questions portant un numéro pair alors que celui de Nedelsky, avec les questions portant un numéro impair. Les résultats rapportés par Andrew et Hecht sont caractérisés par une différence appréciable entre les seuils de performance établis par les deux procédés : Nedelsky (52%), Ebel (68.4%). Une analyse de la variance des seuils obtenus a révélé une différence significative entre les seuils établis selon le procédé alors qu'aucune différence significative n'a été obtenue entre les groupes d'appréciateurs.

En ce qui a trait à la détermination du seuil de performance, certaines conclusions se dégagent immédiatement de ces recherches. D'une part, il apparaît hasardeux de se fier à des appréciateurs considérés individuellement lorsqu'un seul procédé (comme celui de Nedelsky, par exemple) est utilisé. D'autre part il serait possible de se fier à des appréciateurs constitués en groupes. Dans ce deuxième cas, cependant, on ne pourrait utiliser n'importe quel procédé faisant appel à des appréciateurs si on devait établir un seuil de performance à un examen de compétence ou de certification dans une situation réelle.

Les études rapportées jusqu'ici ne permettent pas de tirer des conclusions lorsque des appréciateurs (pris comme individus), des procédés et des questions (qui composent un examen) sont combinés dans une même expérience. Pourtant, lorsqu'il s'agit d'établir un seuil de performance ou une note minimale de passage avec des procédés comme ceux auxquels il a été fait allusion, la trame de fond se joue au niveau d'une question, avec un appréciateur et selon le procédé. Le seuil de performance à établir n'est autre que le fruit d'une compilation statistique lorsque, avec un seul procédé utilisé, il y a plusieurs répliques de l'unité « question, appréciateur ».

C'est à la suite de ces considérations qu'une étude de généralisabilité (du type G) est apparue pertinente. Dans une telle étude, les appréciateurs, les procédés et les

questions doivent idéalement être considérés comme des dimensions croisées. En présumant que la notion de compétence minimale et la subjectivité qui entre en jeu dans l'estimation de la probabilité de réussir une question sont des facteurs pouvant amener, en tout ou en partie, des différences entre des appréciateurs, entre des procédés et entre des questions (et davantage si celles-ci sont de degrés différents de difficulté), il s'agit de se demander si de telles différences peuvent être généralisées à divers contextes. Par exemple, si les questions d'un examen étaient perçues par des appréciateurs comme différentes en difficulté, pour un étudiant de compétence minimale, peut-on affirmer que ces différences sont indépendantes du procédé utilisé et/ou des appréciateurs ? La théorie de la généralisabilité développée par Cronbach et collaborateurs (1972) et l'algorithme de symétrie développé par Cardinet, Tourneur et Allal (1976), dans le cadre de cette théorie, ont été utilisés pour répondre à une telle question.

Déroulement de l'expérience

Onze étudiants gradués inscrits à un cours de deuxième cycle du Département de mesure et évaluation à la Faculté des sciences de l'éducation de l'Université Laval, intitulé « Principes d'évaluation formative », ont participé à l'expérience à titre d'appréciateurs. Un sous-test de dix questions ayant chacune cinq choix de réponses a été choisi à même un examen de compréhension de texte utilisé il y a quelques années par le ministère de l'Éducation du Québec pour les élèves de sixième année du niveau primaire.[2]

Les dix questions du sous-test utilisé sont en rapport avec le participe passé. Ce choix a été effectué dans la perspective d'utiliser un test dont le contenu, le participe passé, serait familier aux appréciateurs. Bien que reliées à un même contenu, les questions du sous-test couvrent des aspects variés : identification de la règle d'accord, identification du mot avec lequel s'accorde un participe passé inséré dans une phrase simple (le participe passé pouvant être employé seul, avec « être » ou avec « avoir ») et application de certaines règles d'accord.

Pour les onze appréciateurs, les procédés ont été utilisés dans l'ordre suivant : Angoff, Ebel et Nedelsky. Cet ordre a été choisi de façon à réduire le plus possible les effets de séquence entre les procédés. Les appréciateurs n'étaient familiers avec aucun des procédés. Chaque appréciateur disposait de trois copies du sous-test pour y inscrire les probabilités de réussite selon le procédé. Les copies ont été recueillies après chaque procédé et, dans le cas du procédé d'Ebel, les proportions attendues de réussites ont été déterminées en groupe. Pour chacun des procédés, l'appréciateur devait concevoir ce qu'un étudiant de compétence minimale devrait au moins ou à peine maîtriser, en ce qui a trait au participe passé, pour réussir le test. Le seuil à choisir serait celui qui servirait de note minimale pour passer du primaire au secondaire si la connaissance du participe passé devait servir de critère unique à une telle décision. Il s'agit bien entendu d'une situation fictive.

La procédure expérimentale qui vient d'être décrite impose aux données qui ont été recueillies une structure à trois dimensions croisées : appréciateurs x procédés x questions. La procédure utilisée souffre de certaines limites qu'il est important de signaler.

On pourrait nous reprocher le manque d'expérience ou d'entraînement des appréciateurs utilisés. Alors que dans certains domaines comme l'observation de comportements et l'analyse de l'enseignement il est possible d'entraîner des observateurs, il ne semble pas que l'on puisse exercer avec autant d'efficacité des appréciateurs à déterminer un seuil de performance. La même difficulté semble d'ailleurs caractériser certaines recherches portant sur des procédés d'évaluation de productions complexes (examens à réponses élaborées). Dans deux études traitant de l'évaluation de compositions littéraires on a eu recours à des candidats à l'enseignement dont plusieurs n'avaient presque pas d'expérience (Hales et Tokar, 1975 ; Hugues, Keeling et Tuck, 1980). Dans l'étude de Coffman et Kurfman (1968), maintes fois citée dans les écrits en évaluation de compositions à cause de son schème expérimental très raffiné, il est dit peu de choses au sujet de la qualité des quatre lecteurs utilisés. À la fin de leur étude les auteurs signalent qu'il serait souhaitable que les lecteurs de compositions comparent leurs évaluations en guise d'entraînement (Coffman et Kurfman, 1968, p. 106).

Le petit nombre de questions et l'ordre unique dans lequel les procédés ont été utilisés ont pu amener un effet de séquence ou de mémoire d'un procédé à l'autre. Toutefois, il apparaît difficile de contrôler un tel effet en permutant simplement les procédés entre eux. Il est raisonnable de penser, par exemple, qu'un appréciateur qui connaît le procédé de Nedelsky ait tendance à transposer au procédé d'Angoff la technique d'analyse des leurres qui s'y rattache. Tout en admettant que l'ordre unique des procédés utilisés ne constitue pas une situation idéale on peut à tout le moins convenir qu'il apparaît invraisemblable ou peu probable que le procédé d'Angoff influence les données à obtenir avec le procédé de Nedelsky. Ces dernières sont établies par dérivation au moyen d'une appréciation des leurres. C'est à ce titre que l'ordre unique des procédés utilisés a été considéré comme présentant le moins de risques de contamination d'un procédé à l'autre et il s'agit là bien sûr d'une présomption qui est implicite dans le schème expérimental de la présente étude.

Enfin, comme autre limite de notre étude, le caractère fictif de la situation présentée aux appréciateurs ajoute à l'ambiguïté qui caractérise la notion de « compétence minimale » ; il faut convenir qu'au Québec, l'usage d'un seuil de performance aux fins de certification ou de « promotion » ne semble pas encore tellement répandu. Il eût donc été difficle d'inscrire notre étude dans un contexte réel et de recruter des appréciateurs expérimentés.

Ces limites seraient sans aucun doute sérieuses si des prises de décision concrètes devaient être recommandées, à la suite de cette étude, quant à l'usage possible dans une situation réelle de certification ou de promotion, de l'un des seuils établis par les trois procédés. Dans le contexte de cette étude pilote, et compte tenu de ces limites, nous avons tenté néanmoins de dégager certaines conclusions pouvant guider d'autres recherches en ce domaine.

Méthodologie

La probabilité estimée de réussir chaque question, pour un étudiant de compétence minimale, a été inscrite dans un tableau à trois dimensions contenant 330 valeurs

numériques comprises entre 0 et 1,00 (11 appréciateurs x 3 procédés x 10 questions). Dans la terminologie de l'analyse de la variance il s'agit d'un schème à trois facteurs croisés avec une seule observation par cellule. Bien que ce schème soit traité dans de nombreux ouvrages en statistique (Kirk, 1968, pp. 228 et suiv. ; Winer, 1971, pp. 394 et suiv. ; Keppel, 1973, pp. 426 et suiv.) il ne représente pas une situation idéale à cause du problème de l'estimation du terme d'erreur. Toutefois le schème avec une seule observation par cellule est typique de la méthodologie utilisée dans la majorité des études en rapport avec les composantes de la variance et de la théorie de la généralisabilité (voir Medley et Mitzel, 1963 ; Endler, 1966 et Cronbach, Gleser, Nanda et Rajaratnam, 1972).

La méthodologie utilisée dans notre étude comporte deux volets : une analyse de la variance des résultats obtenus et une étude de généralisabilité des seuils de réussite. En ce qui a trait à l'analyse de la variance, les procédés, les appréciateurs et les questions ont été considérés comme fixes, ce qui limite la signification des différences obtenues aux procédés, aux appréciateurs et aux questions impliqués comme tels dans l'étude. Pour l'étude de généralisabilité chacune des trois dimensions a été considérée à tour de rôle comme étant fixe ou aléatoire et ce, dans l'esprit même de l'algorithme de symétrie proposé par Cardinet, Tourneur et Allal (1976).

Présentation des résultats et analyse de la variance

Le tableau 1 présente les seuils qui ont été obtenus avec chacun des appréciateurs et les seuils moyens obtenus par procédé pour l'ensemble des appréciateurs. Notons que chacune des moyennes rapportées au bas du tableau 1 pourrait être la note minimale de passage à l'examen de dix questions sur le participe passé. Dans le cas du procédé de Nedelsky, la valeur 3,53 devrait être augmentée d'un certain nombre de fois « k » la valeur de l'écart-type des seuils établis par les appréciateurs (la valeur de « k » étant déterminée par consensus entre les appréciateurs). Il faut se rappeler que cette valeur n'a pas été considérée dans la présente étude et ce dans le but de rendre comparables les seuils provenant des estimations faites par les appréciateurs.

D'après les moyennes rapportées au bas du tableau 1, on observe une certaine différence entre les procédés quant au seuil qui pourrait être établi pour réussir l'examen de 10 questions. Le procédé d'Ebel a produit le seuil le plus exigeant des trois (4,38 sur 10) alors que celui de Nedelsky a produit le seuil le plus indulgent (3,53 sur 10).

On peut observer une certaine divergence entre les appréciateurs quant aux seuils établis avec un même procédé. Dans une certaine mesure, l'écart-type des seuils établis par les appréciateurs considérés individuellement reflète le degré de concordance entre ces appréciateurs. La valeur de 0,40, associée au procédé de Nedelsky, est la plus faible des trois valeurs d'écart-type et suggère que les appréciateurs seraient plus cohérents entre eux avec ce procédé qu'avec les deux autres procédés. Le procédé d'Angoff, lorsque comparé aux deux autres procédés, est caractérisé par la dispersion la plus élevée entre les appréciateurs (écart-type : 1,48) et on voit que d'après l'appréciateur H, le seuil exigé pour réussir l'examen aurait été de 2,1 alors que d'après l'appréciateur K, le seuil aurait été de 6,9.

Tableau 1

Sommes des probabilités de succès d'un étudiant de compétence minimale estimées par chacun des appréciateurs et selon chaque procédé, pour les dix questions de l'épreuve.

Appréciateur	procédé Angoff	Ebel	Nedelsky
A	3,46	5.50	3,73
B	2,50	2,70	3,98
C	3,30	3,90	3,32
D	4,40	3,40	3,48
E	5,80	5,40	3,65
F	5,50	4,90	2,98
G	3,15	4,20	4,32
H	2,10	4,40	3,74
I	4,70	4,80	3,14
J	4,70	4,50	3,26
K	6,90	4,50	3,18
Moyenne :	4,23	4,38	3,53
Écart-type :	1,48	0,83	0,40

Les données de base ont été soumises à une analyse de la variance pour éprouver statistiquement les différences qui viennent d'être signalées. Avec un schème qui ne contient qu'une seule observation par cellule, il est apparu utile de vérifier si l'interaction triple (procédés x appréciateurs x questions) pouvait être utilisée comme terme d'erreur. Le test de non-additivité de Tukey (voir Winer, 1971, pp. 394-397) a été effectué sur les données de base pour donner un rapport F non significatif à α : 0,25 (F : 0,952 pour 1 et 179 degrés de liberté). Puisque le rapport F obtenu ne dépasse pas la valeur critique correspondant au seuil de signification de 0,25 l'hypothèse d'un modèle additif a été conservée, ce qui implique que l'interaction triple a pu être considérée comme un estimé de l'erreur expérimentale (voir Winer, 1971, p. 478 et Kirk, 1968, p. 228).

Les résultats de l'analyse de la variance apparaissent au tableau 2. Ces résultats font voir que les moyennes des probabilités estimées de succès pour un étudiant de compétence minimale sont significativement différentes d'un procédé à l'autre, d'un appréciateur à l'autre et d'une question à l'autre ($p \leqslant 0{,}01$). Il faut cependant être prudent dans l'interprétation de ces différences puisque les interactions doubles sont toutes significatives à un niveau de probabilité inférieur à 0,01. L'importance de ces interactions signifie que les différences observées entre les procédés, entre les appréciateurs et entre les questions ne doivent pas être interprétées de façon absolue. Par exemple, en ce qui a trait aux procédés (P) et aux appréciateurs (A), l'importance de l'interaction PA signifie

que la différence observée entre les procédés est conditionnée par la présence de tel appréciateur plutôt que de tel autre alors que globalement la différence entre les procédés est significative. Cette interprétation s'applique vraisemblablement aux trois dimensions prises deux à deux dans l'analyse. Toutefois, la relation entre les procédés et les appréciateurs, qui apparaît la plus importante à considérer ici, a été examinée de plus près par l'étude des effets principaux simples associés à ces deux dimensions et par un graphique de l'interaction entre ces mêmes dimensions.

Tableau 2

Analyse de la variance des probabilités de succès assignées aux dix questions par les appréciateurs, d'après les trois procédés.

Source de variation	Somme des carrés	D.L.	C.M.	F	P
Procédés (P)	0,45862	2	0,229311	13,234	(0,01)
Appréciateurs (A)	1,04821	10	0,104821	6,049	(0,01)
Questions (Q)	2,43293	9	0,270325	15,601	(0,01)
P x A	1,97870	20	0,098935	5,710	(0,01)
P x Q	0,65754	18	0,036530	2,108	(0,01)
A x Q	2,40495	90	0,026722	1,542	(0,01)
P x A x Q, e	3,11900	180	0,017328		

(P, A et Q ont été considérés comme fixes)

L'analyse de la variance des effets principaux simples apparaît au tableau 3. Les résultats de cette analyse suggèrent que la différence entre les procédés n'est significative (à $p \leq 0{,}01$) que pour six des onze appréciateurs. D'autre part, la différence entre les appréciateurs n'est significative que dans le cas des procédés d'Angoff et d'Ebel mais n'est pas significative dans le cas du procédé de Nedelsky.

Bien que les résultats de l'analyse de la variance suggèrent à première vue des différences significatives entre les procédés et entre les appréciateurs, l'étude des effets principaux simples indique que d'une part on ne peut affirmer que la différence entre les procédés est significative, quel que soit l'appréciateur, et que d'autre part, on ne peut affirmer que la différence entre les appréciateurs est significative, quel que soit le procédé. Cette même conclusion sera reprise avec les coefficients de généralisabilité rapportés dans la dernière partie de cette étude.

Le graphique de l'interaction « procédés x appréciateurs » permet de visualiser les résultats rapportés jusqu'ici en regard de ces deux dimensions. Le profil des appréciateurs a été tracé pour chacun des procédés. Les trois profils obtenus apparaissent à la figure 1. Afin de dégager le sens de l'interaction, les appréciateurs ont été disposés en ordre de sévérité décroissante : ainsi, l'appréciateur E est celui dont la note minimale moyenne de passage pour les trois procédés est la plus élevée (cette note minimale moyenne étant

déduite des données du tableau 1) ; à l'opposé, l'appréciateur B s'est avéré le moins exigeant de tous, si on considère la note minimale moyenne qu'il aurait produite en se servant des trois procédés.

Tableau 3

Analyse des effets principaux simples pour l'étude de l'interaction procédés x appréciateurs

Source de variation	Somme des carrés	D.L.	C.M.	F	P
Entre les procédés :					
appr. A	0,24558	2	0,12279	7,086	(0,01)
appr. B	0,12896	2	0,06448	3,721	N.S. [1]
appr. C	0,02323	2	0,01161	0,670	N.S.
appr. D	0,06176	2	0,03088	1,782	N.S.
appr. E	0,26150	2	0,13075	7,546	(0,01)
appr. F	0,34656	2	0,17328	10,000	(0,01)
appr. G	0,08286	2	0,04143	2,391	N.S.
appr. H	0,28051	2	0,14025	8,094	(0,01)
appr. I	0,17331	2	0,08665	5,001	(0,01)
appr. J	0,12171	2	0,06085	3,512	N.S.
appr. K	0,71136	2	0,35568	20,526	(0,01)
Entre les appréciateurs :					
proc. Angoff	2,18314	10	0,21831	12,599	(0,01)
proc. Ebel	0,68164	10	0,06816	3,934	(0,01)
proc. Nedelsky	0,16215	10	0,01621	0,936	N.S.
(estimé de la variance d'erreur : 0,017328 avec 180 degrés de liberté)					

1. N.S. : non significatif à $p \leqslant 0{,}01$.

La direction ou l'orientation des profils de la figure 1 est quelque peu différente d'un procédé à l'autre. L'orientation horizontale du profil associé au procédé de Nedelsky suggère que la concordance entre appréciateurs est plus grande avec ce procédé qu'avec les deux autres procédés. On voit que la hauteur de ce profil varie approximativement entre 3 et 4. On se rappellera d'ailleurs que l'étude des effets principaux simples n'a indiqué aucune différence significative entre les appréciateurs avec le procédé de Nedelsky. Le profil associé au procédé d'Angoff présente une certaine pente pour l'ensemble des appréciateurs disposés en degré de sévérité décroissante, ce qui indique que les seuils établis par ce procédé reflètent mieux le degré de sévérité des appréciateurs que les seuils établis par le procédé de Nedelsky. Le profil associé au procédé d'Ebel apparaît se situer entre celui du procédé d'Angoff et celui du procédé de Nedelsky.

Quelques hypothèses se présentent d'elles-mêmes pour expliquer les résultats obtenus. Le fait que des juges, des examinateurs ou des appréciateurs diffèrent en degré de

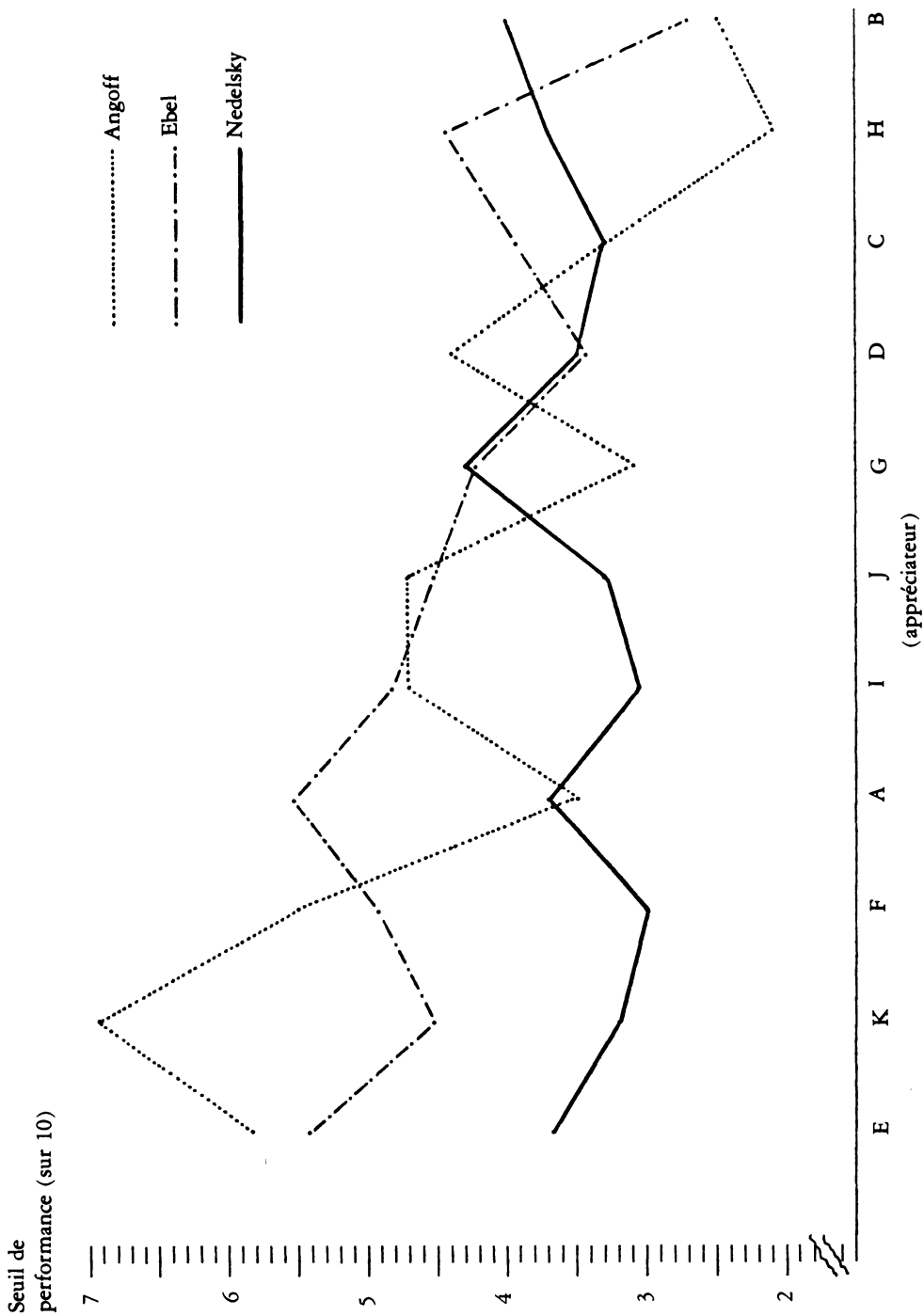

Figure 1 : Profils des seuils de performance établis par les appréciateurs selon le procédé. (Note : les appréciateurs ont été disposés en ordre de sévérité décroissante pour l'ensemble des procédés)

sévérité dans des situations non structurées est un phénomène bien connu dans le domaine de l'évaluation. Le procédé d'Angoff, par son caractère global, prêterait davantage le flanc à cette facette de la subjectivité alors que le procédé de Nedelsky, plus analytique au niveau des leurres avec des questions à choix de réponses, jouerait un rôle de régulateur de cette subjectivité. Pour autant qu'une telle explication est valable et que la concordance est une qualité recherchée, l'amélioration à apporter à des procédés de détermination d'un seuil de performance devrait être du côté de techniques pouvant « objectiver » ou « dimensionnaliser » davantage l'estimation de la probabilité de réussir une question pour un étudiant de compétence minimale.

Le changement de profil observé d'un procédé à l'autre et qui suggère un certain gain en concordance depuis le procédé d'Angoff, en passant par le procédé d'Ebel, jusqu'au procédé de Nedelsky, peut être expliqué autrement. Il faut se rappeler que cet ordonnancement des procédés correspond à celui dans lequel les procédés ont été utilisés dans l'expérience. Il est donc possible que le changement de profil d'un procédé à l'autre soit dû à un effet d'entraînement. Si tel était le cas, des stratégies pourraient être développées pour exercer davantage les personnes devant agir comme appréciateurs pour déterminer un seuil de performance.

Quelle que soit la stratégie invoquée pour améliorer la concordance entre appréciateurs il n'est pas garanti que le seuil de performance qui sera obtenu avec un procédé sera adéquat. Même avec une concordance élevée, le seuil de performance exigé pour réussir un examen peut être trop ou pas assez exigeant. Cette question semble reliée au degré de familiarité que des appréciateurs peuvent avoir avec une situation donnée.

La généralisabilité des seuils de performance

Les résultats analysés jusqu'ici et les conclusions qui s'en dégagent ont été centrés sur les appréciateurs et les procédés. Afin d'étayer davantage ces conclusions en tenant compte des trois dimensions considérées dans cette recherche, une étude de généralisabilité a été réalisée à partir des résultats de l'analyse de la variance telle que décrite au tableau 1.

Les composantes de la variance ont été estimées à partir des carrés moyens rapportés au tableau 1 pour chacun des trois modèles suivants : le modèle à effets fixes (celui utilisé pour l'analyse de la variance proprement dite), un modèle mixte où seule la dimension « procédés » est fixe et le modèle à effets aléatoires. La méthode de calcul utilisée, qui ne peut être exposée ici, est traitée dans plusieurs ouvrages en rapport avec la théorie de la généralisabilité (Endler, 1966 ; Cronbach et al., 1972 ; Cardinet, Tourneur et Allal, 1976). En analyse classique de la variance il n'est pas coutume de varier ainsi les modèles mais dans l'esprit de la théorie de la généralisabilité, l'estimation des composantes de la variance selon divers postulats quant à la nature des dimensions étudiées a pour objet la détection des principales sources de variation qui caractérisent un dispositif expérimental (voir par exemple Medley et Mitzel, 1963, pp. 312 et suiv.). Les composantes estimées de la variance ainsi que leurs pourcentages pour les trois modèles retenus apparaissent au tableau 4.

Tableau 4

Composantes estimées de la variance et leurs pourcentages selon trois modèles d'analyse de la variance.

Source	P, A et Q fixes		P fixe A et Q aléat.		P, A et Q aléatoires	
	composantes	%	composantes	%	composantes	%
Proc. (P)	1,93[a]	4,5	1,01	2,4	1,01	2,7
Appr. (A)	2,92	6,8	2,60	6,3	0,00[b]	—
Quest. (Q)	7,67	17,9	7,38	17,9	6,80	17,8
P x A	8,16	19,0	8,16	19,7	8,16	21,4
P x Q	1,74	4,0	1,75	4,2	1,75	4,6
A x Q	3,13	7,3	3,13	7,6	3,13	8,2
P x A x Q,e	17,33	40,4	17,33	41,9	17,33	45,4
Total :	42,88	100	41,36	100	38,18	100

a) les valeurs numériques des composantes de la variance doivent être multipliés par 10^{-3}
b) la valeur zéro a été substituée à une valeur négative de cette composante de la variance.

L'examen des pourcentages apparaissant au tableau 4 indique à première vue qu'une proportion relativement élevée de la variance totale est associée à l'interaction triple P x A x Q confondue à l'erreur. Cette proportion varie approximativement de 40% à 45% d'un modèle à l'autre. L'importance relativement élevée de la variance d'erreur indique que le dispositif expérimental choisi pour cette étude est caractérisé par plusieurs sources de variation non contrôlées. Les pourcentages du tableau 4 révèlent de plus deux autres sources de variation non négligeables dans les trois modèles : les questions avec 17,8% ou 17,9% de l'interaction procédés x appréciateurs avec 20% environ.

On peut remarquer que les diverses sources de variation, à l'exception des procédés, demeurent relativement stables, en proportion, d'un modèle à l'autre. Pour mieux saisir l'influence des diverses sources de variation identifiées sur la fiabilité des seuils de performance, l'algorithme de Cardinet, Tourneur et Allal (1976) a été appliqué aux composantes de la variance du tableau 4 qui sont en rapport avec le modèle à effets aléatoires. Les divers coefficients de généralisabilité estimés, lorsque chaque dimension de l'analyse est considérée à tour de rôle comme facette de différenciation ou facette de généralisation, apparaissent au tableau 5.

Les coefficients de généralisabilité en rapport avec les questions comme facette de différenciation offrent un certain intérêt pour la présente étude. Dans la mesure où les questions sont réellement différentes en difficulté, pour un élève de compétence minimale, les coefficients de généralisabilité permettent d'évaluer en quelque sorte le

Tableau 5

Valeurs estimées des coefficients de généralisabilité selon divers dispositifs quant au choix des facettes de différenciation et de généralisation.

Facettes de...			Le problème étant de...		$\hat{\rho}^2$
	généralisation				
différenciation	aléatoire	fixée	différencier :	peu importe :	
Q	A	P	la difficulté des questions, estimée avec les trois procédés utilisés	les appréciateurs	0,901
Q	P	A	la difficulté des questions, estimée par les onze appréciateurs	le procédé	0,865
Q	P,A		la difficulté des questions	les appréciateurs et le procédé	0,830
A	P	Q	les estimations faites par les appréciateurs pour les dix quest.	le procédé	0,087
A	Q	P	les estimations faites par les appréciateurs avec les trois procédés	les questions	0,753
A	P,Q		les estimations faites par les appréciateurs	le procédé et les questions	0,000[1]
P	Q	A	les procédés utilisés par les onze appréciateurs	les questions	0,841
P	A	Q	les procédés utilisés pour estimer la diff. des dix questions	les appréciateurs	0,569
P	A,Q		les procédés utilisés	les questions et les appréciateurs	0,485

1. le seul composant de la variance, au numérateur, est nul.

degré auquel cette différenciation est indépendante des appréciateurs ou des procédés. Un coefficient élevé, c'est-à-dire une faible dépendance, indiquerait que le dispositif expérimental utilisé dans la présente étude mesure adéquatement la difficulté perçue des questions quels que soient les appréciateurs ou les procédés (selon la facette de généralisation choisie).

La différenciation entre les appréciateurs et la différenciation entre les procédés, selon les facettes de généralisation qui les accompagnent, sont la contrepartie de l'étude. Si, par exemple, le dispositif expérimental devait établir une différenciation fiable entre les procédés et ce, quels que soient les questions et les appréciateurs, il faudrait conclure que les procédés utilisés dans cette étude ne sont pas équivalents et ne peuvent donc pas être utilisés indistinctement l'un de l'autre.

Les diverses questions en rapport avec la généralisabilité des résultats de l'étude ont été formulées dans le tableau 5 dans le but de faciliter l'interprétation des divers coefficients de généralisabilité qui y sont rapportés.

Pour ce qui est de la différenciation entre les questions, les coefficients de généralisabilité apparaissant au tableau 5 suggèrent que le dispositif expérimental utilisé dans cette recherche mesurerait adéquatement les différences en difficulté estimée des questions avec les trois procédés à l'étude (facette de généralisation fixée) et ce, quels que soient les appréciateurs (coefficient de 0,901). De plus, si on accepte la valeur de 0,80 comme standard à atteindre pour un coefficient de généralisabilité (Cardinet, Tourneur et Allal, 1972, p. 129), ces différences en difficulté des questions, telles que perçues par les onze appréciateurs, peuvent être également généralisées à tout procédé (0,865) et peuvent être généralisées à la fois à tout procédé et à tout appréciateur (0,830).

Lorsqu'il s'agit de différencier les unes des autres les estimations faites par les appréciateurs, le coefficient de généralisabilité le plus élevé des trois qui apparaissent en regard de cette facette de différenciation est 0,753. D'après le standard de 0,80 pour un coefficient acceptable de généralisabilité, le dispositif expérimental de la présente étude ne permettrait pas de mesurer adéquatement les estimations faites par les appréciateurs avec les trois procédés (facette fixée), quelles que soient les questions. Pour atteindre ce standard, il faudrait utiliser un minimum de 13 questions dans quel cas le coefficient de généralisabilité passerait de 0,753 à 0,798. Toujours en ce qui a trait à la différenciation entre les appréciateurs, mais quel que soit le procédé, le coefficient obtenu est relativement très faible (0,087). Ce résultat est compatible avec une constatation faite auparavant dans l'étude des effets principaux simples : la différence entre les appréciateurs bien que significative globalement, n'est pas significative pour tous les procédés pris un à un. La différenciation obtenue entre les estimations faites par les appréciateurs ne serait donc pas indépendante des questions et des procédés utilisés.

Lorsqu'il s'agit de différencier entre les procédés, le coefficient de généralisabilité de 0,841, le plus élevé des trois coefficients, signifie que le dispositif utilisé différencie adéquatement les estimations faites selon ces procédés par les onze appréciateurs, quelles que soient les questions. Il faut souligner que cette différenciation peut apparaître

indésirable dans un certain sens. Les constatations faites à ce sujet nous amènent en effet à craindre que les différences déjà soupçonnées entre les procédés, qui se sont avérées significatives globalement, soient généralisables aux questions posées à l'égard du participe passé. Cependant, la différenciation entre les procédés n'apparaît pas indépendante des appréciateurs lorsque la facette 'questions' est fixée (0,569) ni des appréciateurs et des questions (0,485). Lors de l'étude des effets simples, la différence entre les procédés ne s'est avérée significative que pour six des onze appréciateurs, ce qui laissait déjà entendre que la différenciation entre les procédés n'était pas indépendante des appréciateurs.

À partir des coefficients de généralisabilité obtenus dans cette étude on peut avancer certaines interprétations. Il apparaît que les estimations faites par des personnes, de la difficulté de questions traitant du participe passé, peuvent être mesurées adéquatement, quels que soient les appréciateurs ou les procédés et ce, en dépit de certaines limites qui ont été soulignées quant au choix de ces appréciateurs. Ceci ne signifie en rien que les seuils établis sont adéquats car l'interprétation qui vient d'être avancée n'est valable que pour les différences établies en difficulté perçue des questions et non pour la difficulté même de ces questions.

Les différences observées entre les appréciateurs dépendraient et des questions et des procédés utilisés. Certains appréciateurs peuvent avoir des degrés de sévérité différents selon le procédé alors que la sévérité manifestée par d'autres appréciateurs serait indépendante du procédé. La sévérité des appréciateurs peut également varier à des degrés divers selon la question examinée. Il est probable qu'une question manifestement facile ou manifestement très difficile soit l'objet d'une plus grande concordance entre les appréciateurs que ne le serait une question de difficulté moyenne. Une telle hypothèse pourrait être vérifiée dans une autre étude.

Enfin, la différenciation entre les procédés apparaît fiable lorsqu'il s'agit de généraliser cette différenciation aux questions tout en considérant les onze appréciateurs uniquement. Les coefficients obtenus lorsque le procédé est considéré comme facette de différenciation indiquent qu'il est difficile de différencier entre les seuils de performance établis selon le procédé, quels que soient les appréciateurs ou quels que soient les appréciateurs et les questions.

En résumé, la différenciation entre des questions ayant trait au participe passé est généralisable à tout appréciateur et à tout procédé faisant appel à des appréciateurs. En contrepartie, les procédés utilisés avec les onze appréciateurs amènent des seuils différents les uns des autres et ce, quelles que puissent être les questions ayant trait au participe passé.

Conclusion

Trois procédés faisant appel à des appréciateurs pour déterminer le seuil de performance à un examen ont été étudiés dans cette recherche : le procédé d'Angoff, celui d'Ebel et celui de Nedelsky.

Les résultats obtenus confirment ceux déjà rapportés dans des recherches précédentes à savoir que les seuils établis ne sont pas comparables d'un procédé à l'autre ni d'un appréciateur à l'autre. De plus, les questions se rapportant au participe passé se sont avérées significativement différentes en difficulté d'après les estimations faites par les appréciateurs. Cependant, les effets dus aux procédés, aux appréciateurs et aux questions ne sont pas indépendants les uns des autres comme le laisse entrevoir la présence d'interactions significatives entre les trois dimensions traitées dans l'étude, prises deux à deux.

Les profils des seuils de performance établis par les appréciateurs, selon le procédé, indiquent une concordance entre ces appréciateurs plus grande dans le cas du procédé de Nedelsky que dans le cas de chacun des deux autres procédés.

Comme deuxième volet de cette recherche, une étude de généralisabilité a été réalisée dans le but d'estimer la fiabilité des différenciations observées. Les coefficients de généralisabilité les plus élevés ont été obtenus pour différencier entre les questions, quels que soient les appréciateurs et les procédés. La différenciation entre les appréciateurs s'est avérée peu fiable et la différenciation entre les procédés serait généralisable à toute question se rapportant au participe passé à la condition toutefois de se limiter aux onze appréciateurs utilisés dans l'étude.

Les différences observées entre les seuils de performance établis selon le procédé et selon l'appréciateur indiquent qu'on ne peut faire appel indifféremment à tout procédé ou à tout appréciateur pour déterminer la note minimale de passage à un examen se prêtant à une interprétation « critériée » des résultats. Le nombre d'étudiants qui seraient échoués varierait donc d'un procédé à l'autre et d'un appréciateur à un autre. Il est encore plus facile de supposer que des élèves se situant à proximité d'un certain niveau de compétence minimale seraient livrés aux caprices de la situation pour être ballotés entre la compétence et l'incompétence. Le problème de la subjectivité que pose l'utilisation d'un procédé faisant appel à des appréciateurs pour établir un seuil de performance demeure donc entier.

Les résultats obtenus dans cette recherche ne font pas que conduire à un bilan négatif. C'est du moins ce qui se dégage de la concordance relativement élevée entre les appréciateurs dans le cas du procédé de Nedelsky et de la fiabilité avec laquelle il a été possible de différencier la difficulté perçue des questions.

Le gain en concordance, observé depuis le procédé d'Angoff jusqu'au procédé de Nedelsky, peut être expliqué en partie par l'effort plus ou moins grand d'analyse qui est exigé de la part des appréciateurs par chacun des procédés. Le procédé d'Angoff est global. Celui d'Ebel propose un schème à deux dimensions pour classifier les questions d'un examen. Le procédé de Nedelsky, le plus analytique des trois, exige une certaine interprétation du degré d'attraction des leurres puisqu'il s'applique à des questions à choix de réponses. Cette constatation devrait nous amener à encourager le développement de techniques particulières pour guider davantage le travail d'analyse des appréciateurs. Dans le cas de questions à choix de réponses il nous faudrait accéder à une meilleure

compréhension des raisons profondes pouvant expliquer le choix d'un leurre par un étudiant de compétence minimale de façon à mieux distinguer les erreurs grossières des simples erreurs. Dans le cas de questions à réponse élaborée ou à réponse brève, pour lesquelles le procédé de Nedelsky ne s'applique pas dans sa version originale, il faudrait en arriver à anticiper les réponses pouvant caractériser un étudiant de compétence minimale et, à partir de ces réponses, estimer la probabilité de succès à une question pour ce type d'étudiant. Ce sont là quelques suggestions parmi bien d'autres qui pourraient orienter la recherche en ce domaine.

Pour ce qui est de la fiabilité des seuils de performance, le dispositif expérimental utilisé dans cette recherche s'est avéré adéquat pour différencier la difficulté perçue des questions et ce, quels que soient les appréciateurs et les procédés. Dans une certaine mesure, ce résultat est une indication que des procédés faisant appel à des appréciateurs peuvent être sensibles aux différences de difficulté perçue des questions et ce, en dépit du choix qui a été fait de ces appréciateurs et du petit nombre de questions utilisées. Cette façon de voir la détermination du seuil de performance par des appréciateurs suggère à première vue une stratégie expérimentale qui mérite d'être explorée davantage. Par exemple, en ayant à notre disposition des examens de niveaux variés de difficulté, établis expérimentalement auprès d'étudiants, il serait alors possible d'étudier des améliorations qu'on pourrait apporter à un procédé en particulier dans le but de rendre le seuil de performance indépendant des appréciateurs ou de tout autre effet de contexte.

Même si une concordance élevée était obtenue entre des appréciateurs, le problème de la détermination du seuil de performance n'est pas entièrement résolu pour autant. Avec un degré élevé de concordance le seuil établi peut être inadéquat c'est-à-dire qu'il peut être trop ou pas assez exigeant. Le seuil de performance idéal est celui qui différencie adéquatement les individus qui doivent échouer de ceux qui doivent réussir. Pour dire si le seuil établi à l'aide d'un procédé connu est adéquat, il nous faudrait un point de repère permettant d'évaluer jusqu'à quel point nous étions justifiés de faire échouer certains étudiants et d'en faire réussir d'autres. Avec ce point de repère, la notion de compétence minimale serait moins abstraite. Pour prendre un exemple avec le contenu de l'examen qui a été utilisé dans cette recherche, peut-on avancer l'hypothèse que la probabilité de réussir certaines questions pour un étudiant de compétence minimale serait estimée avec plus de justesse si on révélait à des appréciateurs la nature des tâches et des problèmes pour lesquels la maîtrise du participe passé est prérequise ?

Les procédés suggérés jusqu'à maintenant pour faire déterminer le seuil de performance à un examen, par des appréciateurs, feront l'objet de critiques sérieuses pour plusieurs années encore. Devant une telle situation nous serions tentés d'avoir recours à des méthodes arbitraires, faute de mieux. Le problème n'est pas résolu pour autant et de telles méthodes se prêtent difficilement à l'expérimentation.

NOTES

1. l'expression « seuil de performance » est une traduction des expressions « performance standard » ou « minimum passing score » utilisées dans les ouvrages américains.
2. d'après les informations obtenues auprès du ministère de l'Éducation il s'agit d'un test expérimental qui s'inscrivait, il y a quelques années, dans le cadre des tests de rendement en français pour la fin du niveau primaire.

RÉFÉRENCES

Andrew, Barbara J. et James T. Hecht, A Preliminary Investigation of two Procedures for Setting Examination Standards, *Educational and Psychological Measurement,* vol. 36, no 1, 1976, p. 45-50.

Angoff, William H., Scales, Norms, and Equivalent Scores, Chap. 15, p. 508-600 dans : Robert L. Thorndike (éd.), *Educational Measurement,* 2ème edition, Washington D.C. : American Council on Education, 1971.

Bunda, Anne et James R. Sanders, éditeurs, *Practices and Problems in Competency-Based Measurement,* Washington D.C. : National Council on Measurement in Education, 1979.

Cardinet, Jean, Yvan Tourneur et Linda Allal, The Symmetry of Generalizability Theory : Applications to Educational Measurement, *Journal of Educational Measurement,* vol. 13, no 2, 1976, p. 119-135.

Coffman, William E. et Dana Kurfman, A Comparison of Two Methods of Reading Essay Examinations, *American Educational Research Journal,* vol. 5, no 1, 1968, p. 99-107.

Cronbach, Lee J., G.C. Gleser, H. Nanda et N. Rajaratnam, *The Dependability of behavioral Measurement : Theory of Generalizability for scores and profiles,* New York : Wiley, 1972.

Ebel, Robert L., Determination of the Passing Score, p. 492-496 dans : *Essentials of Educational Measurement,* Englewood Cliffs, New Jersey : Prentice-Hall Inc, 1972.

Endler, Norman S., Estimating Variance Components from Mean Squares for Random and Mixed Effects Analysis of Variance Models, *Perceptual and Motor Skills,* 1966, vol. 22, p. 559-570.

Glass, Gene V., Standards and Criteria, *Journal of Educational Measurement,* vol. 15, no 4, 1978, p. 237-261.

Hales, L.W. et E. Tokar, The Effect of the Quality of Preceding Responses on the Grades assigned to Subsequent Responses to an Essay Question, *Journal of Educational Measurement,* vol. 12, no 2, 1975, p. 115-117.

Hambleton, Ronald K. et Daniel R. Eignor, Issues and Methods for Standard-Setting, sixième unité dans : *Criterion-Referenced* Test Development and Validation Methods, AERA Training Program Materials, avril 1979.

Hughes, David C., Brian Keeling et Bryan F. Tuck, The Influence of Context Position and Scoring Method on Essay Scoring, *Journal of Educational Measurement,* vol. 17, no 2, 1980, p. 131-135.

Keppel, Geoffrey, *Design and Analysis : A Researcher's handbook,* Englewood Cliffs, N.J. : Prentice-Hall, 1973.

Medley, Donald M. et Harold L. Mitzel, Measuring Classroom Behavior by Systematic Observation, Chapitre 6, p. 247-329 dans : N.L. Gage (éditeur), *Handbook of Research on Teaching : a Project of The American Educational Research Association,* vol. 46, no 1, 1976, p. 133-158.

Meskauskas, Hohn A. et G. W. Webster, The American Board of Internal Medicine Recertification Examination Process and Results, *Annals of Internal Medicine,* 82, 1975, p. 577-581.

Millman, Jason, Passing Scores and Test Lengths for Domain-Referenced Measures, *Review of Educational Research,* vol. 43, no 2, 1973, p. 205-216.

Nedelsky, Léo, Absolute Grading Standards for Objective Tests, *Educational and Psychological Measurement,* vol. 14, no 1, 1954, p. 3-19.

Winer, B.J., *Statistical Principles in Experimental Design,* 2ème edition, New York : McGraw-Hill, 1971.