Mesure et évaluation en éducation

Les directives accompagnant une épreuve scolaire : sur quels sujets doivent-elles porter ?

Huguette Croteau et Serge P. Séguin

Volume 18, numéro 2, 1995

URI : https://id.erudit.org/iderudit/1092279ar
DOI : https://doi.org/10.7202/1092279ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Croteau, H. & Séguin, S. P. (1995). Les directives accompagnant une épreuve scolaire : sur quels sujets doivent-elles porter ? *Mesure et évaluation en éducation*, *18*(2), 81–103. https://doi.org/10.7202/1092279ar

Résumé de l'article

Le rôle important des directives de test, pour ce qui est des aspects logique et empirique de la fidélité, nous a incités à connaître, de façon la plus exhaustive possible, les directives proposées dans la documentation scientifique. Un inventaire, puis une synthèse, ont été effectués et la pertinence des sujets relevés a été vérifiée dans le contexte québécois de la mesure. Il en résulte, validée, *une liste-critère des sujets des manuels de test*. Pour les concepteurs de tests, cette liste fournit un outil de planification du contrôle des erreurs de mesure ; pour les utilisateurs de tests, elle constitue un outil d'évaluation de l'exhaustivité et de la pertinence du contenu des manuels de test.

Mesure et évaluation en éducation
Vol. 18 n° 2, pages 81 à 103.

Les directives accompagnant une épreuve scolaire: sur quels sujets doivent-elles porter?

Huguette Croteau
Université du Québec à Montréal et Évatest[1]

Serge P. Séguin
Université du Québec à Montréal

Le rôle important des directives de test, pour ce qui est des aspects logique et empirique de la fidélité, nous a incités à connaître, de façon la plus exhaustive possible, les directives proposées dans la documentation scientifique. Un inventaire, puis une synthèse, ont été effectués et la pertinence des sujets relevés a été vérifiée dans le contexte québécois de la mesure. Il en résulte, validée, une liste-critère des sujets des manuels de test. Pour les concepteurs de tests, cette liste fournit un outil de planification du contrôle des erreurs de mesure ; pour les utilisateurs de tests, elle constitue un outil d'évaluation de l'exhaustivité et de la pertinence du contenu des manuels de test.

(directives de test, consignes, instructions, manuels de test, guide de test, cahier du candidat, cahier de l'examinateur, standardisation)

Test instructions are important to logical and empirical reliability. This led us to perform an inventory of the test instructions suggested in the literature and to synthesize the topics found. The result of our comprehensive survey is a validated list of criteria, which is comprised of selected topics whose relevance is verified in the Québec measurement context. For test developers, this list is a tool for the planification of the control of errors of measurement. Test users, on the other hand, will find it useful in evaluating the exhaustiveness and the pertinence of test manuals contents.

(directions, instructions, test's manuals, test's guides, standardization)

Préambule: mise au point terminologique

Dans cet article, le terme **test** est retenu pour désigner tout **examen** ou **épreuve** ou toute instrumentation contenue dans les **méthodes d'évaluation** telles que définies comme suit par le Comité consultatif mixte (1993):

Les **méthodes d'évaluation** désignent les différentes stratégies et les différentes techniques que peuvent utiliser les enseignants pour obtenir des informations à des fins d'évaluation. Ces stratégies et ces techniques incluent, sans toutefois s'y limiter, les observations, les questions et les tests reliés aux manuels scolaires et aux programmes d'étude, les tests papier-crayon, l'interrogation orale, les modèles de production servant de référence, les entrevues, l'évaluation par les pairs et l'auto-évaluation, les tests standardisés critériés et normatifs, les évaluations de la performance, les échantillons de production écrite, les expositions, les dossiers de présentation, les évaluations de projet et de produit. Diverses appellations ont été proposées pour décrire des sous-ensembles de ces stratégies et de ces techniques. Parmi les plus courantes, on retrouve «évaluation authentique», «évaluation de la performance», et «évaluation alternative». Toutefois, pour les fins du présent document, le terme «méthode d'évaluation» comprend toutes les stratégies et les techniques qui peuvent servir à recueillir des informations sur les progrès des élèves par rapport aux connaissances, aux habiletés, aux attitudes et aux comportements visés.

Nous retenons aussi les définitions de concepteurs et d'utilisateurs, formulées comme suit par le Comité consultatif mixte (1993):

Par **concepteurs**, on entend les personnes qui construisent ces méthodes de même que les personnes qui décident des politiques régissant les divers programmes d'évaluation. Quant aux **utilisateurs**, ils comprennent les personnes qui choisissent et utilisent les méthodes d'évaluation, celles qui retiennent les services de concepteurs ou celles qui prennent des décisions fondées sur les résultats et les conclusions d'une évaluation. Ces rôles peuvent se chevaucher lorsqu'un enseignant ou un instructeur élabore et administre un instrument pour ensuite corriger et interpréter les réponses de l'élève, ou lorsqu'un ministère de l'Éducation ou un système scolaire local commandite l'élaboration ou l'implantation d'un programme d'évaluation et de services de correction en vue de décisions fondées sur cette évaluation.

Uniformiser est ici synonyme de **contrôler** ou de **standardiser** les conditions de passation, de correction, de notation et de communication des résultats, que les qualités métrologiques de l'instrument de mesure soient connues ou non.

Nous utilisons **sujets** et non *directives* ou *sujets de directives* parce que les sujets des manuels de test contiennent des directives mais aussi d'autres données.

Nous utilisons aussi **manuels de test** plutôt que *guides de test* parce que, le produit étant le test, les manuels décrivent comment utiliser ce produit. Ils sont au nombre de deux: le **manuel d'utilisation,** qui s'adresse aux responsables de l'évaluation, aux administrateurs et aux correcteurs du test, et le **manuel du candidat.**

Pour les responsables de l'évaluation, les administrateurs et pour les correcteurs du test, la **documentation** inclut le manuel d'utilisation, la clé de correction et la fiche de verdict. Pour le candidat, la documentation inclut le plus souvent le manuel du candidat, le questionnaire et la feuille de réponse. Ces documents ou d'autres, constituant le **matériel du test,** devraient être identifiés dans les manuels de test.

Problématique

L'approche curriculaire, utilisée en formation professionnelle (FP) au secondaire, implique l'intégration de l'évaluation à la planification des programmes. Dans cette approche, les mesures utilisées pour encadrer l'élaboration de tests valides et produisant des résultats fidèles, incluent plusieurs guides dont le *Guide de présentation des épreuves en formation professionnelle* (Gouvernement du Québec, 1990) qui traite des directives de test accompagnant ces épreuves.

Comme l'évaluation sommative des apprentissages en formation professionnelle au secondaire, aux fins de la sanction des études, relève à 30% du ministère de l'Éducation du Québec et à 70% des commissions scolaires, ces organismes conçoivent des tests pouvant être administrés à plusieurs groupes d'élèves et par différents administrateurs. Dans ces cas, pour éviter des risques d'écarts indus dans les résultats en uniformisant les situations de «testing» d'un endroit ou d'un administrateur à un autre, on compte sur les directives de test et sur la compétence des administrateurs en mesure et évaluation. C'est dans ce contexte que les directives de test prennent toute leur importance. Le concepteur de test a la responsabilité d'arrêter les conditions d'administration du test qu'il élabore et de les décrire dans les manuels d'utilisation du test.

L'élaboration des directives de tests écrits est simplifiée grâce au guide précité (Gouvernement du Québec, 1990) dont les exemples sont facilement adaptables. Par contre, dans le cas des tests utilisant une grille d'observation, l'élaboration des directives et l'adaptation des exemples

présentent des difficultés. Le guide suggère de se référer à des exemples de directives de tests ministériels de divers programmes. L'étude de ces exemples nous a fait douter de l'exhaustivité des sujets traités et de la pertinence de certaines directives.

Lorsqu'on se tourne vers la documentation scientifique pour solutionner ces problèmes, nous constatons que plusieurs associations et spécialistes en mesure et en évaluation (Morissette, 1993; Airasian et Terrasi, 1994; Gouvernement du Québec, 1990; AERA, APA, NCME, 1985; Traxler, 1951; Gouvernement du Québec, 1988-1990; Mehrens et Lehmann, 1987; Cangelosi, 1990; Gronlund, 1993; Tousignant et Morissette, 1990) précisent qu'il est important de présenter des directives claires, complètes et concises mais ils ne présentent que quelques éléments à traiter dans les manuels de test.

Nous n'avons pas trouvé, dans la documentation scientifique, de solution simple aux problèmes de l'exhaustivité des sujets des manuels de test et de leur pertinence. En effet, l'organisation de la présentation des directives et la diversité des sujets varient grandement d'une source à une autre, ce qui laisse supposer qu'il n'y a pas de consensus dans ce domaine. Nous avons rencontré aussi des difficultés reliées à la terminologie jusqu'à ce que nous trouvions les définitions du Comité consultatif mixte (1993).

Les objectifs de notre recherche découlent de la difficulté de trouver réponses à nos questions. Ces dernières nous ont incités à vouloir savoir quelles sont les directives proposées dans la documentation scientifique. Les objectifs étaient donc d'élaborer une liste-critère aussi exhaustive que possible des sujets des manuels de tests, selon la documentation scientifique, et de vérifier la pertinence des sujets dans le contexte québécois de la mesure des apprentissages scolaires.

Le produit de cette recherche consiste en une *liste-critère validée des sujets des manuels de test* conçue «à l'intention des concepteurs de méthodes d'évaluation et d'instruments de mesure ainsi que de leurs utilisateurs» (Comité consultatif mixte, 1993: 3).

Partie d'une problématique en formation professionnelle, cette recherche nous a menés à un champ beaucoup plus vaste, qui touche l'ensemble des tests de tous les ordres d'enseignement, du primaire à l'université, et qui s'adresse à tous les intervenants, tant concepteurs qu'utilisateurs. En effet, les sujets de la liste-critère s'appliquent, comme dans le cas du document préparé par le Comité consultatif mixte (1993: 4)

> aux évaluations effectuées par les enseignants aux niveaux élémentaire, secondaire [...] également au niveau post-secondaire [...] sur les évaluations standardisées qui sont élaborées à l'extérieur des établissements d'enseignement, par des maisons d'édition spécialisées, des ministères de l'Éducation et des juridictions locales.

Contexte conceptuel

Il importe ici de rappeler comment, théoriquement, les sujets des manuels de tests peuvent avoir des incidences sur les qualités métrologiques des tests.

Quand les résultats sont non fidèles, les décisions basées sur ces résultats sont non valides, comme le précisent Selltiz et al. [(1977; rapporté par Legendre (1993)]: «[...] un instrument non fidèle ne peut être valide [...].»

L'évaluation de la fidélité d'un instrument de mesure implique, selon Stanley (1971), trois types d'opérations, à savoir logique, empirique et statistique. Au-delà des aspects statistiques de la fidélité, Stanley invite à considérer les aspects logique et empirique, fortement interdépendants, qui concernent le but visé et les opérations pour estimer ou apprécier la fidélité. C'est que dès qu'on mesure un objet ou les caractéristiques d'un individu, il y a de l'erreur de mesure ou de l'infidélité, selon Feldt et Brennan (1989): «Except for isolated, exceptional cases, all measurements must be presumed to contain error.»

Pour obtenir une fidélité acceptable des résultats, il faudrait préciser les sources de variance et décrire, dans les manuels de test, les moyens de contrôle prévus pour tenter de minimiser les erreurs de mesure. Ces moyens sont nécessaires parce que, à moins qu'un effort délibéré ne soit fait pour assurer un niveau acceptable de fidélité, les tests en éducation tendent, selon Martuza (1977), à être peu ou pas suffisamment fidèles:

> In general, variable measurement errors tend to be quite sizable in educational test scores (or, equivalently, educational tests tend to be unreliable) unless a deliberate effort is made to insure an acceptable level of reliability (p. 9).

Il y a deux sortes d'erreurs de mesure: l'erreur constante (systématique ou permanente) et l'erreur aléatoire (temporaire, variable, ou due au hasard). L'erreur constante se distingue par le fait qu'elle se répète systématiquement d'une situation de «testing» à une autre tandis que

l'erreur aléatoire, comme son nom l'indique, est due au hasard. Pour distinguer l'erreur aléatoire de l'erreur constante, prenons l'exemple suivant en mesure physique: si l'on se pèse à plusieurs reprises sur une balance de salle de bain, par exemple, l'erreur aléatoire pourrait être due à un ajustement à zéro (de la balance avant de se peser) variable à chaque occasion, tandis qu'une erreur constate, résulterait du fait que la balance serait mal calibrée (par exemple: valeur erronée du zéro). L'intérêt principal de l'identification de l'erreur de mesure et des facteurs de biais résultant en erreurs systématiques est que ces dernières sont parfois récupérables tandis que les erreurs aléatoires ne le sont pas.

Martuza (1977) spécifie que les erreurs aléatoires peuvent parfois devenir des erreurs systématiques. Par exemple, un éclairage inadéquat ou un local mal chauffé sont des sources d'erreurs aléatoires mais peuvent devenir des sources d'erreurs systématiques si ces situations ne sont pas corrigées d'une administration à une autre.

L'erreur due au hasard est introduite dans le processus de la mesure, selon Martuza (1977), par des facteurs associés aux candidats (tels l'état de santé, le niveau de motivation, l'humeur), aux conditions générales de «testing» (tels le bruit, l'éclairage, la température), aux procédures de correction et à l'instrument (tels des directives ambiguës adressées aux candidats, des items ambigus, un nombre insuffisant d'items dans le test). Selon Clemans (1971), trois catégories de participants peuvent influencer l'environnement dans lequel la mesure est obtenue: le concepteur du test, l'administrateur (incluant le correcteur) et le candidat.

Il est plus facile de contrôler les sources d'erreurs provenant des concepteurs et des administrateurs que celles provenant des candidats. Thorndike (1951) relève des sources de variance provenant des individus et les répartit en caractéristiques générales, spécifiques, durables et temporaires. Il y a aussi les conditions d'administration et les facteurs divers. L'auteur nous met en garde contre l'interchangeabilité de ce classement selon la situation de «testing». Morissette (1993) fournit aussi une *classification de quelques sources de variation dans les résultats d'un examen,* qui ne diffère que partiellement de la liste de Thorndike. Selon Magnusson (1966), c'est l'analyse de la situation de «testing» qui permet de dégager les facteurs pouvant créer de l'erreur de mesure ou de l'erreur systématique.

Les listes de Thorndike et de Morissette ne sont pas directement utilisables lors de la conception d'un test parce qu'elles se concentrent sur

la distinction des sources de variance et non sur les sujets des directives pouvant permettre de minimiser les erreurs en provenance de ces sources. Par contre, Clemans s'est penché sur cet aspect et, en 1971, il a fourni une liste de vérification des sujets (*topics*) qui devraient être traités dans les manuels de test. Cette liste, bien qu'elle ne contienne que 27 sujets et sous-sujets, a servi de point de départ pour notre recherche.

Le diagramme de la page suivante permet de comprendre l'importance des sujets des manuels de test en ce qui concerne les qualités métrologiques des instruments de mesure.

Instruments d'observation, erreur de mesure et fidélité

Généralement, dans la documentation scientifique, les directives d'épreuves écrites sont traitées séparément des directives de tests utilisant une grille d'observation: ceci laisse supposer que les sujets des manuels de test diffèrent selon le type de test. Toutefois, l'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education [AERA, APA, NCME], (1985) affirment que tout test (celui utilisant une grille d'observation est ici inclus dans la définition de test) est affecté par les conditions d'administration. Il faut cependant reconnaître que les conditions d'administration et de correction d'un test utilisant une grille d'observation sont plus complexes, et que le risque d'erreur de mesure est d'autant plus grand, que dans une situation de test papier crayon à réponses choisies ou brèves. Popham (1990) rappelle d'ailleurs qu'en situation d'observation d'une performance psychomotrice ou d'une production ou d'une réponse élaborée à une question à développement oral ou écrit, les résultats des élèves sont fortement dépendants de l'observateur:

> The observers themselves are the pivotal people in any observation enterprise [...]. Proper training of observers, complete with specific directions, practice sessions, and interobserver disagreement resolution, is imperative (p. 290).

La stabilité de tout observateur ou correcteur (fidélité intra-juge) et la concordance des observations ou notations entre les observateurs (fidélité inter-juges) ainsi que le contrôle des erreurs de mesure et l'uniformisation des situations de «testing» (fidélité logique et empirique) sont donc autant d'intérêt que la constance des résultats (fidélité statistique) aux examens écrits.

Importance des sujets des manuels de test en ce qui concerne les qualités métrologiques des instruments de mesure

VALIDITÉ

«[...] **un instrument non fidèle ne peut être valide** [...]»
Selltiz et al. [(1977), rapporté par Legendre (1993)]

FIDÉLITÉ ET ERREUR DE MESURE

Dès qu'on mesure un objet ou les caractéristiques d'un individu, **il y a de l'erreur de mesure ou de l'infidélité**, selon Feldt et Brennan (1989).
Il y a deux sortes d'erreurs de mesure: **l'erreur constante** (systématique ou permanente) et **l'erreur aléatoire** (temporaire, variable, ou due au hasard).

SOURCES DE VARIATION DANS LES RÉSULTATS

Il est nécessaire de **distinguer les sources de variation** pour pouvoir contrôler ce qui est contrôlable.
Les sources de variation peuvent provenir des **concepteurs** et des **utilisateurs** de tests.

MOYENS DE CONTRÔLE DES ERREURS DE MESURE: LES SUJETS DES MANUELS DE TEST

Les **sujets des manuels de test** s'adressent aux **utilisateurs** de tests et ont pour but de contrôler les erreurs de mesure et d'uniformiser les situations de «testing».
Quant à la **liste-critère validée des sujets des manuels de test** ainsi que la **liste des recommandations**, elles s'adressent tant aux **concepteurs** qu'aux **utilisateurs** de tests.
La **liste-critère validée** des sujets des manuels de test est un **outil de planification** du contrôle des erreurs de mesure et **un outil d'évaluation** de l'exhaustivité et de la pertinence du contenu des manuels.

À l'intérieur de ce cadre conceptuel, nous considérons que les sujets des manuels de test concernent la fidélité logique et empirique des résultats et qu'ils permettent de communiquer aux utilisateurs les moyens de contrôle prévus par le concepteur pour tenter de minimiser les erreurs de mesure et uniformiser les situations de «testing». Ce contrôle, il va de soi, n'est que conditionnel au fait que les manuels de test soient suivis par les utilisateurs.

Méthode

L'étude a été conduite en deux temps pour répondre, respectivement, aux critères d'exhaustivité des sujets et de pertinence de chacun des sujets, constituants essentiels de leur validité.

Pour assurer l'exhaustivité des sujets des manuels de test à couvrir, nous avons retenu et analysé 13 sources. Au moment du choix des sources, le document intitulé *Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada* (Comité consultatif mixte, 1993) n'avaient pas encore été édité. Les sources choisies sont considérées comme les principales références en langue anglaise ou française utilisées pour la construction de tests ou d'examens. Quatre sources sont québécoises et de langue française alors que les neuf autres sont nord-américaines et de langue anglaise: Morissette, 1993; Airasian et Terrasi, 1994; Gouvernement du Québec, 1990; AERA, APA, NCME, 1985; Ebel et Frisbie, 1991; Traxler, 1951; Gouvernement du Québec, 1988-1990; Mehrens et Lehmann, 1987; Clemans, 1971; Popham, 1990; Cangelosi, 1990; Gronlund, 1993; Tousignant et Morissette, 1990.

Les composantes du discours ont été dégagées et associées à des thèmes d'intérêt que nous appelons sujets des manuels de test. La stratégie générale d'organisation des données a nécessité des étapes de codification, de production de sommaires et de recherche de thèmes intégrateurs. Cette classification a été expérimentée auprès d'étudiants universitaires intéressés par notre recherche. Une dernière vérification des références a permis de répondre à des questions comme: est-ce possible qu'il y ait seulement neuf sources sur 13 qui spécifient, par exemple, *le but du test ?* Lorsqu'il y avait un doute, un retour aux sources était effectué pour une dernière vérification.

Nous ne prétendons pas avoir intégré tous les sujets suggérés dans les 13 sources. Certains sujets nous ont peut-être échappé puisqu'il fallait les

déduire à partir de leur présentation sous forme narrative; on trouvait également plusieurs sujets dans une même phrase. L'éparpillement des sujets dans certains textes est aussi un aspect qui nous a causé divers tracas: nous comprenons que cet éparpillement est voulu par certains auteurs afin d'éviter une liste trop technique à suivre aveuglément. Cette crainte cependant rend la tâche d'élaboration des manuels de test très complexe et c'est pourquoi nous désirions synthétiser cette matière en une liste-critère. En extrayant, des sources retenues, les sujets des manuels de test, nous avons relevé des sujets qu'il ne nous semblait pas opportun de conserver dans la liste-critère. Par contre, nous les avons conservés sous forme de liste appelée *Recommandations aux concepteurs de tests,* parce qu'il s'agit bien de recommandations.

Parce que la liste-critère a été établie d'une façon empirique, nous présumons que tous ses sujets sont matière à vérification à travers des recherches futures.

Nous avons élaboré aussi un index des sujets parce que: 1) certains sujets sont abordés dans les deux manuels (le manuel d'utilisation et le manuel du candidat); 2) il est souvent nécessaire de pouvoir les trouver facilement et rapidement; 3) certains sujets sont abordés dans plus d'une chapitre d'un même manuel; 4) un sujet peut parfois être traité à plus d'un endroit dans un même chapitre; par exemple, le sujet qui traite des *qualifications des administrateurs* nous renvoie aux sujets 2.1 et 2.6: le sujet 2.1 traite des qualifications générales et le sujet 2.6, des qualifications spécifiques. L'index a facilité le classement des sujets proposés dans les diverses sources. Il devrait aider les concepteurs de manuels à repérer le ou les endroits où traiter de tel ou tel sujet. Il devrait être utile aussi aux utilisateurs de tests qui doivent se retrouver dans tout ce matériel fort varié.

La liste-critère des sujets des manuels de test selon la documentation scientifique a été soumise à une étude empirique visant à apprécier la pertinence de chacun des sujets. Pour ce faire, la liste-critère a été transformée en grille d'évaluation en ajoutant une échelle d'appréciation à chacun des sujets de la liste. L'échelle utilisée était simple: le sujet était pertinent ou non. Cependant, si le sujet était considéré pertinent, on demandait aux juges d'être plus précis et de choisir entre un peu, moyennement ou très pertinent (l'extrait de la grille d'évaluation, sujets 5.1 à 5.4, est présenté en annexe). Les juges étaient aussi invités à émettre des commentaires ou des suggestions directement sur la grille d'évaluation. La cueillette d'information s'est effectuée par courrier.

La grille d'évaluation de la pertinence des sujets a servi à la collecte des données et a permis de connaître l'opinion de gens concernés par l'évaluation des apprentissages. Nous avons sollicité 30 personnes (taux de réponses de 97%) engagées en évaluation des apprentissages aux ordres d'enseignement primaire et secondaire. Les 29 répondants se répartissent en trois groupes: 11 docimologues universitaires, 10 responsables de l'évaluation dont 4 au ministère de l'Éducation du Québec et 6 dans des commissions scolaires, et 8 administrateurs de tests.

L'ensemble du primaire et du secondaire est ainsi couvert: formation générale des jeunes, formation générale des adultes, formation professionnelle et technique, formation à distance, bien que nous n'ayons pas tenté d'obtenir un échantillon scientifiquement représentatif de l'ordre d'enseignement précollégial.

Résultats

Les résultats de la recherche documentaire ont donné la *Liste-critère des sujets des manuels de test selon la documentation scientifique.* Les sujets sont regroupés en sept chapitres pour le manuel d'utilisation qui s'adresse aux administrateurs (incluant les correcteurs) et en deux chapitres pour le manuel du candidat. Les titres des chapitres du manuel d'utilisation sont: 1) Informations générales; 2) Préparation des administrateurs; 3) Préparation des candidats; 4) Préparation du matériel et de l'équipement; 5) Préparation du local; 6) Administration du test; 7) Correction et interprétation des résultats. Quant au manuel du candidat, les titres des chapitres sont: 1) Informations générales ; 2) Informations sur la notation.

De plus, les résultats de la recherche ont donné une liste de *Recommandations aux concepteurs de tests:* elle se compose d'une quinzaine de recommandations, selon la documentation scientifique.

Un extrait de la liste-critère est présenté au tableau 1, accompagné d'une légende. On y trouve les références précises à chacune des sources qui mentionnent un sujet ou le sous-sujet donné. La colonne intitulée N indique le nombre de sources qui proposent d'inclure dans le manuel d'utilisation le sujet ou le sous-sujet mentionné. Voici un exemple de lecture du tableau 1: trois sources (Airasian et Terrasi, 1994; Gronlund, 1993; Traxler, 1951) proposent de traiter le sujet 6.3 dans le manuel d'utilisation du test.

Légende des titres des colonnes de la liste-critère

AT94	Airasian et Terrasi[1]
C90	Cangelosi
C71	Clemans
MEQ88	Ministère de l'Éducation du Québec[2]
EF91	Ebel et Frisbie
G93	Gronlund
ML87	Mehrens et Lehmann
MEQ90	Ministère de l'Éducation du Québec[3]
M93	Morissette
P90	Popham
AERA...85	AERA, APA, NCME
TA51	Traxler
TM90	Tousignant et Morissette

Exemples d'information contenue dans la liste:

p267	page 267
1A1	numéro du sujet donné dans Clemans (1971)
5.2P	standards 5.2 Primaire (voir Abréviations ci-contre)
1.1	numéro de sujet donné dans MEQ90

Abréviations utilisées dans la colonne AERA...85

P: Primaire[4]: sujets qui doivent être présents dans les manuels ou les guides (à justifier en cas d'absence).

S: Secondaire: sujets qu'il serait souhaitable d'avoir mais qui parfois sont difficilement réalisables dans plusieurs situations. Ces sujets aident à évaluer et à comparer les tests entre eux (ils n'ont pas à être justifiés en cas d'absence).

C: conditionnel: sujets primaires ou secondaires, selon la situation.

NB: Commentaire: ceci n'est pas un standard mais un commentaire accompagnant le standard.

---

1 Voir les références complètes dans la bibliographie.
2 MEQ88 fait référence à Gouvernement du Québec (1988-1990) dans cet article.
3 MEQ90 fait référence à Gouvernement du Québec (1990) dans cet article.
4 En référence aux catégories définies dans AERA, APA, NCME (1985) (traduction libre).

Tableau 1

Liste-critère des sujets des manuels de test selon la documentation scientifique

Sujets proposés pour le «Manuel d'utilisation»	N	Sources: AT94	C90	C71	MEQ88	EF91	G93	ML87	MEQ90	M93	P90	AERA...85	TA51	TM90
6. Administration du test														
6.1 Procédure d'admission des candidats	2		IVC									15.3NB		
6.2 Création d'un climat sain, non stressant. Aider les élèves à relaxer, les rassurer en disant que la raison d'être du test est de prendre de bonnes décisions de classement pour la poursuite de leur étude. Être calme.	6	p6314			p58			p31		p270	p331		p356	
6.3 Lecture des renseignements et des directives inclus dans le guide du candidat, d'une voix claire, en utilisant exactement les mots de l'auteur (trop souvent les candidats négligent de lire attentivement et ne peuvent donc suivre malgré une motivation et un intérêt élevé de réussir).	3	p6314					p102						p363	
6.4 Présentation de la tâche à accomplir avec suffisamment de détails pour que les candidats sachent ce qu'ils ont à faire (spécifier la répartition des points).	5						p101		3.1.3	p267		3.22P		p91

(Suite à la page suivante.)

Tableau 1 (suite)

Liste-critère des sujets des manuels de test selon la documentation scientifique

Sujets proposés pour le «Manuel d'utilisation»	N	Sources: AT94	C90	C71	MEQ88	EF91	G93	ML87	MEQ90	M93	P90	AERA...85	TA51	TM90
6.5 Façon de répondre aux questions des candidats:														
- prévoir des réponses aux questions possibles des candidats	9	p6311			V C1	p206		p31	3.1.3	p270	p332	3.22NB		p91
- répondre sans dédain, sarcasme ou impatience, peu importe le genre de question	1	p6313												
- s'assurer de ne pas donner d'indices de réponses aux items	3		p63			p218		p31						
- être sensible aux émotions des candidats et adopter des techniques qui permettent aux candidats de donner le maximum de performance	1	p6314												
- répondre aux questions pendant la période prévue seulement	3	p6314				p204							p364	
- répondre aux questions de manière à ce que tous les candidats en profitent	3		p63							p269	p332			
- prévoir la gestion du matériel défectueux	2	p6312		V C2										

Le nombre total de sujets et de sous-sujets inventoriés dans les 13 sources est de 427. Le tableau 2 fait voir le nombre de sujets et de sous-sujets (sss) par source et par manuel. En regroupant et en juxtaposant les sujets inventoriés, nous obtenons un total de 125 sss (90 sujets et 35 sous-sujets), soit plus du double de la source qui en fournissait le plus.

Tableau 2

Nombre de sujets et de sous-sujets (sss) recensés par chacune des 13 sources

Sources	Manuel d'utilisation	Manuel du candidat	Total
M93	46	8	54
AT94	47	4	51
MEQ90	27	18	45
AERA...85	37	5	42
EF91	29	11	40
TA51	32	7	39
MEQ88	31	5	36
ML87	26	2	28
C71	24	3	27
P90	18	2	20
C90	14	4	18
G93	12	4	16
TM90	9	2	11
Total sss recensés	352 sss	75 sss	427 sss
Total sss différents	103 sss	22 sss	125 sss
Total des sujets	70 sujets	20 sujets	90 sujets

L'analyse quantitative des données a permis de constater que, dans la documentation scientifique, aucun sujet ne fait l'unanimité, plus de la moitié des sujets et des sous-sujets (54%) sont mentionnés dans au moins trois sources, tandis que 46% sont mentionnés dans une ou deux sources seulement. Ces résultats confirment notre intuition première à l'effet que les sujets proposés dans la documentation scientifique ne sont pas nécessairement les mêmes d'une source à une autre, mais qu'ils se complètent plutôt.

Les résultats de l'étude de la pertinence ont donné la *Liste-critère validée des sujets des manuels de test* sur laquelle nous reviendrons. La grille d'évaluation de la pertinence des sujets a été remplie par 29 juges. Nous avons retenu la règle de décision suivante: le sujet ou le sous-sujet est accepté si les réponses cotées 2 et 3 (moyennement et très pertinent) totalisent plus de 50% des réponses au sujet en question. Dans le cas contraire, le sujet est rejeté. Un extrait du sommaire des résultats (ceux des sujets du chapitre 3 concernant la *Préparation des candidats*), incluant ces chiffres et les décisions d'acceptation ou de rejet des sujets ou de sous-sujets, est présenté au tableau 3. Ce tableau contient aussi la moyenne des réponses, le nombre de répondants et la fréquence des réponses par valeur scalaire. Remarquons le sujet 3.6: la moyenne des réponses est de 1,46 et seulement 12 des 28 juges (43%) considèrent ce sujet comme moyennement ou très pertinent: le sujet est donc rejeté.

Les caractéristiques du sujet 3.6 mériteraient d'être reconsidérées. En effet, ce sujet est particulièrement controversé par les docimologues universitaires qui ont participé à notre recherche. Aussi, dans la documentation scientifique, ce sujet est proposé par trois sources en plus d'être traité par le Comité consultatif mixte (1993) à la ligne directrice #5: «Démontrer l'effet sur les résultats d'évaluation de facteurs tels que [...], les stratégies de «passation de tests.»

Le tableau 4 présente le nombre de sujets classés selon leur valeur scalaire moyenne, cette valeur variant entre 0 et 3. On constate que 94% des sujets et sous-sujets ont des réponses moyennes de plus de 1,50 c'est-à-dire qu'ils sont considérés par les juges, comme moyennement ou très pertinents.

Tableau 3

Sommaire des résultats de l'évaluation de la liste-critère et décisions

Sujets proposés pour le «Manuel d'utilisation», selon la documentation scientifique	Réponse moyenne	Nombre de répondants	Fréquence par valeur scalaire 0*	1	2	3	Pourcentage des réponses 2 et 3	A: accepté R: rejeté
3. Préparation des candidats								
3.1 Information à communiquer aux candidats afin de les motiver:								
- moment de l'administration	2,55	29	1	3	4	21	86%	A
- explication du but du test	2,82	29	1	0	2	26	97%	A
- contenu général sur lequel portera le test	2,69	29	1	0	6	22	97%	A
- importance de chaque partie du contenu	2,34	29	1	4	8	16	83%	A
- type de performance exigée	2,51	29	1	1	9	18	93%	A
- conditions de réussite	2,51	29	1	1	9	18	93%	A
- façon de corriger et de noter	2,17	29	2	5	8	14	76%	A
- importance du test (sélection, sanction, promotion, etc.)	2,65	29	2	0	4	23	93%	A
3.2 Explications concernant l'utilisation des résultats afin de réduire l'anxiété et de s'assurer de la coopération des candidats.	2,27	29	2	3	9	15	83%	A
3.3 Encouragement à se concentrer sur la tâche à accomplir (plutôt que sur les émotions négatives: nervosité excessive, prévision de l'échec et de ses conséquences), encourager une attitude positive et ne jamais minimiser l'importance du test auprès des candidats.	2,00	29	4	4	9	12	72%	A

* 0: non pertinent, 1: peu pertinent, 2: moyennement pertinent, 3: très pertinent

Tableau 3 (suite)

Sommaire des résultats de l'évaluation de la liste-critère et décisions

Sujets proposés pour le «Manuel d'utilisation», selon la documentation scientifique	Réponse moyenne	Nombre de répondants	Fréquence par valeur scalaire 0*	1	2	3	Pourcentage des réponses 2 et 3	A: accepté R: rejeté
3.4 Questions des candidats: répondre de préférence avant l'administration du test.	2,31	29	3	1	9	16	86%	A
3.5 Préparation mentale (apprendre le domaine couvert par le test) et physique (être en forme, avoir suffisamment dormi).	1,72	29	6	5	9	9	62%	A
3.6 Développement de l'habileté des candidats à faire des tests.	1,46	28	6	10	5	7	43%	R
3.7 Précision du matériel que les candidats devraient apporter lors de l'administration du test (crayon et efface en double, etc.).	2,75	29	1	1	2	25	93%	A
3.8 Communication des directives préliminaires la veille du test.	1,72	29	6	5	9	9	62%	A

* 0: non pertinent, 1: peu pertinent, 2: moyennement pertinent, 3: très pertinent

Tableau 4

Barème de fréquences des réponses moyennes

Réponses moyennes (Valeur scalaire)	Nombre de sujets et de sous-sujets	Fréquence cumulative	Fréquence cumulative en %
2,50 - 2,99	41	41	33%
2,00 - 2,49	44	85	68%
1,50 - 1,99	33	118	94%
1,00 - 1,49	5	123	98%
0,50 - 0,99	2	125	100%
0,00 - 0,49	0	125	100%

Dans le dernier tableau (tableau 5), on remarque que seulement 6 des 125 sujets et sous-sujets n'ont pas le 50% de consensus requis pour être acceptés; 29 sous-sujets ont un degré de consensus variant entre 50 et 69%; 90 sous-sujets ont un degré de consensus inter-juges de plus de 70% dont 37 sous-sujets atteignent un degré de consensus inter-juges supérieur à 90%.

Tableau 5

Barème de fréquences des sujets moyennement et très pertinents, par degré de consensus

Degré de consensus inter-juges	Nombre de sujets et de sous-sujets	Fréquence cumulative	Fréquence cumulative en %
90%-99%	37	37	30%
80%-89%	21	58	46%
70%-79%	32	90	72%
60%-69%	19	109	87%
50%-59%	10	119	95%
0%-49%	6	125	100%

Après avoir enlevé les six sujets et sous-sujets rejetés et avoir pris en considération l'intégration de quelques sujets par suite des commentaires des juges, nous obtenons une *liste-critère validée* de 84 sujets: 64 pour le manuel d'utilisation et 20 pour le manuel du candidat.

Les commentaires et suggestions de quelques juges concernaient surtout l'ajout d'expressions comme *s'il y a lieu, au besoin, si c'est pertinent, si disponible, si nécessaire*, à sept sujets. Nous laissons ces ajouts à la discrétion des concepteurs parce que l'ajout de ces expressions pourrait s'appliquer à presque tous les sujets et que nous n'avions pas consensus relativement à ces ajouts. Les sujets ne s'appliquent pas tous à un test donné. Par exemple, dans le chapitre relatif à la préparation des administrateurs, le sujet 2.7 *Façon d'utiliser les grilles d'observation*, s'applique uniquement aux grilles d'observation. Autre exemple, dans le manuel du candidat, parmi les renseignements généraux, le sujet 1.6 *Respect des règles de santé et de sécurité,* s'applique surtout au test en formation professionnelle. Bien qu'il y ait quelques sujets spécifiques comme ces exemples, il nous semble préférable d'avoir une seule liste incluant quelques sujets spécifiques à certains tests, plutôt que d'avoir plusieurs listes, une par type de test. Les concepteurs devraient donc sélectionner les sujets qui s'appliquent à chacun des différents tests qu'ils élaborent.

Conclusion

L'élaboration de manuels de test de qualité est un domaine complexe, d'abord parce qu'il implique de nombreux sujets à traiter et ensuite, parce que la documentation scientifique n'était pas tout à fait transparente et unanime sur les sujets à traiter dans les manuels de test. C'est cette situation qui nous a motivés à faire un inventaire et une synthèse des sujets proposés dans la documentation scientifique. L'outil qui en résulte a été validé dans le contexte québécois de la mesure des apprentissages scolaires. Nous espérons que les résultats de cette étude contribueront à l'avancement de l'équité en mesure et évaluation en fournissant aux concepteurs de tests un outil de planification du contrôle des erreurs de mesure et aux utilisateurs de tests, un outil d'évaluation de l'exhaustivité et de la pertinence du contenu des manuels de tests.

Cet outil est présenté dans un livret[2] édité en octobre 1995, comprenant la liste-critère validée et une liste complémentaire de recommandations selon la documentation scientifique, accompagnées d'un index. Ce guide pratique et facile d'utilisation ne compte qu'une quarantaine de pages et présente seulement les résultats qui importent lors de la conception ou de l'évaluation des manuels de test. Il s'intitule *Manuels de test et d'examen: l'importance des directives.*

Il reste beaucoup à faire dans le domaine des manuels de test. Il conviendra éventuellement de distinguer les sujets considérés obligatoires de ceux qui sont considérés facultatifs. Par obligatoire, on entend les sujets à mettre absolument dans tout manuel de test et à justifier en cas d'absence; par facultatif, on entend ceux qui ne s'appliquent pas nécessairement à tous les tests et dont on n'a pas à justifier l'absence. AERA, APA, NCME (1985) font cette distinction ainsi que le gouvernement du Québec (1990).

D'autres recherches pourraient tenter de répondre à diverses questions dont: Quelle est la façon la plus efficace et concise de formuler un sujet de manuels de test? Serait-il préférable de remettre le manuel d'utilisation à l'administrateur du test suffisamment de temps avant l'administration pour qu'il puisse mettre le contenu en oeuvre? Est-ce que les sujets des manuels de test sont lus et appliqués?

Au cours des dernières décennies, l'accent semble avoir été mis, d'une part, sur la validité de contenu et, d'autre part, sur l'aspect statistique de la fidélité, à l'exclusion des aspects logique et empirique auxquels font référence les manuels de tests. L'étude que nous venons de rapporter et la liste qui en résulte permettent d'ouvrir un vaste champ d'exploration auquel il conviendrait d'accorder un intérêt particulier pour qu'il puisse récupérer la place qui lui revient.

NOTES

1. Évatest est une entreprise qui offre des services d'évaluation et de développement de tests, d'épreuves ou d'examens.

2. Publié par Évatest, avenue 3635 Ridgewood app. 303, Montréal (Québec) H3V 1B4. Commande postale : faire parvenir un chèque ou un mandat-poste de 10,90 $ par livret (incluant les frais de manutention et d'expédition). Nous acceptons seulement l'envoi contre remboursement (C.O.D.) pour les commandes téléphoniques (514) 739-8959.

RÉFÉRENCES

AERA, APA, NCME, voir American ...

Airasian, P. W. & Terrasi, S. (1994). Test administration. In International Encyclopedia of Education (pp. 6311-6315).

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington (DC) : American Psychological Association.

Cangelosi, J. S. (1990). Designing tests for evaluating student achievement. White Plains (N.Y.): Longman.

Clemans, W. V. (1971). Test administration. In R. L. Thorndike (Éd.), Educational Measurement (pp. 188-201). Washington (DC): American Council on Education.

Comité consultatif mixte. (1993). Principes d'équité relatifs aux pratiques d'évaluation des apprentissages scolaires au Canada. Edmonton (Alberta): Center for Research in Applied Measurement and Evaluation, University of Alberta.

Ebel, R. L. & Frisbie, D. A. (1991). Essentials of educational measurement. Englewood Cliffs (N.J.): Prentice-Hall.

Feldt, L. S. & R. L. Brennan. (1989). Reliability. In Linn (Éd.), Educational measurement (p. 113), Washington (DC) : American Council on Education.

Gouvernement du Québec (1990). Guide de présentation des épreuves en formation professionnelle. Québec: ministère de l'Éducation, Direction du développement de l'évaluation. Formation professionnelle au secondaire.

Gouvernement du Québec (1988-1990). Éléments de docimologie (fascicules 1,2,3,4). Québec: ministère de l'Éducation. Les publications du Québec.

Gronlund, N. E. (1993). How to make achievement tests and assessments (5e éd.). Needham Heights (Mass.): Allyn and Bacon.

Legendre, R. (1993). Dictionnaire actuel de l'éducation (2e éd.). Montréal: Guérin.

Magnusson, D. (1966). Test theory. Don Mills (Ont.): Addison-Wesley.

Martuza, V. R. (1977). Applying norm-referenced and criterion-referenced measurement in education. Boston: Allyn and Bacon.

Mehrens, W. A. & Lehman, I. J. (1987). Using standardized tests in education (4e éd.). N.Y.: Longman.

Morissette, D. (1993). Les examens de rendement scolaire (3e éd.). Québec: Presses de l'Université Laval.

Popham, W.J. (1990). Modern educational measurement: a practitionner's perspective. Englewood Cliffs (N.J.): Prentice-Hall.

Stanley, J. C. (1971). Reliability. In R.L.Thorndike (Éd.), Educational measurement (pp. 356-378). Washington (DC) : American Council on Education.

Thorndike, R. L. (1951). Reliability. In E. F. Lindquist (Éd.), Educational Measurement (pp. 560-620). Washington (DC): American Council on Education.

Tousignant, R. & Morissette, D. (1990). Les principes de la mesure et de l'évaluation des apprentissages (2e éd.). Boucherville (Qué.): Gaétan Morin.

Traxler, A. E. (1951). Administering and scoring the objective test. In E. F. Lindquist (Éd.) Educational Measurement (pp. 329-365). Washington (DC): American Council on Education.

ANNEXE
Grille d'évaluation de la liste-critère

Sujets proposés pour le «Guide de l'administrateur du test» selon la documentation	Mettre un ✓ dans la case appropriée* 0	1	2	3
5. Préparation du local				
5.1 Bruit: éviter les sonneries de cloche ou de téléphone. S'il y a plus d'un administrateur éviter de parler ensemble, ne pas parler inutilement avant le test et ne jamais parler pendant le test.	☐	☐	☐	☐
5.2 Interruption: une indication sur la porte d'entrée (exemple: «Examen en cours, prière de ne pas déranger») peut permettre d'éviter d'être interrompu.	☐	☐	☐	☐
5.3 Aménagement des lieux nécessaire à l'obtention de résultats valides:				
- local adéquat	☐	☐	☐	☐
- chaises et des tables de travail raisonnablement confortables et espaces de travail suffisants	☐	☐	☐	☐
- plan d'emplacement des sièges dans la salle, au besoin	☐	☐	☐	☐
- espace suffisant entre les bureaux pour éviter le copiage	☐	☐	☐	☐
- permettre une circulation dans la salle pendant le test	☐	☐	☐	☐
- prévoir une horloge murale précise et visible par tous ou un autre moyen d'indiquer le temps. De préférence avoir un deuxième moyen de contrôler l'heure.	☐	☐	☐	☐
5.4 Conditions physiques des lieux:				
- vérifier que l'éclairage soit adéquat	☐	☐	☐	☐
- s'assurer que la température est modérée	☐	☐	☐	☐
- s'assurer d'une circulation d'air convenable.	☐	☐	☐	☐

* 0 = non pertinent, 1 = peu pertinent, 2 = moyennement pertinent, 3 = très pertinent