Méthodes en sciences humaines

L

M

D

Introduction aux théories des tests

en psychologie et en sciences de l'éducation

- Méthode détaillée
- Exemples concrets
- Tableaux de synthèse

Dany Laveault
Jacques Grégoire

4e édition

OFFERT

+ EN LIGNE

Questionnaires de compréhension
Ressources complémentaires

deboeck SUPÉRIEUR

Méthodes en sciences humaines

L

M

D

Introduction aux théories des tests

en psychologie et en sciences de l'éducation

- Méthode détaillée
- Exemples concrets
- Tableaux de synthèse

Dany Laveault
Jacques Grégoire

4e édition

deboeck SUPÉRIEUR

RESSOURCES NUMÉRIQUES

EN FIN DE CHAPITRE

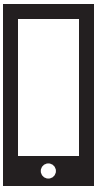
- Questionnaires de compréhension

EN LIGNE

- Notions d'inférence statistique
- Webographie
- Liens utiles
- Glossaire des symboles

Repérez les codes QR et accédez directement à votre ressource :

Scannez le code avec votre téléphone ou votre tablette



OU

Tapez l'URL dans votre navigateur

Pour toute information sur notre fonds et les nouveautés dans votre domaine de spécialisation, consultez notre site web : **www.deboecksuperieur.com**

4e édition
Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve

Dépôt légal :
Bibliothèque nationale, Paris : juin 2023
Bibliothèque royale de Belgique, Bruxelles : 2023/13647/045

ISSN 1373-0231
ISBN 978-2-8073-5107-3

SOMMAIRE

Ressources complémentaires

www.lienmini.fr/51073-compl

AVANT-PROPOS

Chaque nouvelle édition d'*Introduction aux théories des tests en psychologie et en sciences de l'éducation* cherche à refléter les progrès accomplis dans le domaine de la mesure et de l'évaluation tout en réaffirmant les fondements théoriques en psychométrie et en édumétrie qui ont particulièrement bien traversé les années. On trouvera des traces de ce parcours en consultant la liste des références du livre qui constitue une riche bibliographie des principaux ouvrages accumulés au cours des ans et qui s'étendent sur plus de 50 ans. C'est que, dans la discipline qui est la nôtre, le savoir s'est construit de manière inclusive, chaque nouveau modèle bâtissant sur les plus anciens, sans nécessairement les invalider.

Dans ce contexte où l'ancien et le nouveau cohabitent, la principale difficulté pour les constructeurs et concepteurs d'instruments de mesure et d'évaluation est de savoir s'y retrouver. Pour les auteurs de ce livre, c'est de pouvoir guider le lecteur dans un domaine où le savoir est en constante expansion en faisant les meilleurs choix pour le contenir dans un format qui demeure accessible. Après tout, c'est le but de tout ouvrage d'introduction. Il aura été atteint si, après avoir lu ce livre ou en avoir consulté l'une ou l'autre section, le lecteur sera mieux orienté et pourra trouver réponse à ses questions avec l'aide des nombreuses références et des conseils qui lui sont fournis.

En ce sens, la quatrième édition demeure fidèle à notre conception du domaine de la mesure et de l'évaluation telle que formulée dans l'avant-propos de la troisième édition :

> *Tout n'est pas quantifiable ou ne mérite pas de l'être, mais ce qui pourrait l'être mérite que l'on s'y intéresse correctement. La mesure n'est pas une nécessité, mais à la « dictature des chiffres » nous opposons l'autorité que la mesure peut apporter à l'étude scientifique en éducation et en psychologie. L'autorité de la mesure ne vient pas des chiffres eux-mêmes, ni des données quantifiables. Elle s'appuie sur une utilisation appropriée des modèles théoriques des tests ainsi que sur la justesse des inférences que nous en retirons. Cet ouvrage montre que c'est possible et propose les méthodologies appropriées pour y parvenir.*

Pour rester fidèle à notre conception de départ, cette quatrième édition a dû faire des choix et apporter des changements à l'édition précédente. Le passage à une version numérique a ouvert de nouvelles possibilités pour le lecteur dont nous avons cherché à tirer bénéfice. Voici une courte énumération des changements apportés :

- Physiquement, le nombre de chapitres est inchangé ainsi que le nombre de pages. Nous y sommes parvenus en libérant l'espace occupé par l'annexe 1 de la troisième édition portant sur les notions d'inférence statistique. Le contenu de cette annexe est maintenant disponible gratuitement en ligne et permet aux lecteurs qui le souhaitent de revoir leurs notions de base en statistiques inférentielles.
- Comme le livre est utilisé non seulement comme ouvrage de référence, mais aussi comme manuel de cours au niveau gradué et post-gradué, nous avons introduit un questionnaire interactif à la fin de chaque chapitre pour permettre au lecteur-étudiant de vérifier sa propre compréhension des principales notions. Ces questionnaires sont accessibles via un code QR (ou l'URL correspondante).
- Les principales « nouveautés » se trouvent au chapitre 1 sur la construction d'un instrument de mesure et au chapitre 7 sur les modèles de réponse à l'item et les nouvelles technologies. Dans le cas du chapitre 1, nous avons fait état des progrès réalisés dans la construction des tests, qu'il s'agisse de la conception des tests eux-mêmes ou de la spécification des items. Il en est résulté de nouvelles sections sur les formats d'items, les capacités accrues de rédaction des items et, au final, sur une véritable méthodologie de rédaction des items. Dans le chapitre 7, à l'autre extrémité du développement d'un instrument de mesure, se sont ajoutés les nouveaux modèles de réponse à l'item et leur utilité dans la création de banques d'items, qu'il s'agisse du testing sur ordinateur dans sa forme la plus simple jusqu'au testing adaptatif, sa forme la plus aboutie. Avec la multiplication des plateformes numériques pour l'administration des tests, il est également devenu nécessaire d'aborder la question de la comparabilité des résultats obtenus de façon numérique ou papier-crayon.
- De façon plus globale, les nouvelles sections des autres chapitres font écho aux évolutions récentes en matière de validité des résultats, notamment avec la version la plus récente des standards américains (chapitre 4). Nous avons aussi ajouté un nouvel outil pour estimer la fidélité de cohérence interne (chapitre 3).

En abordant l'influence croissante de l'évaluation dans nos sociétés, Charles Hadji (2012) soulève une question fondamentale : « Faut-il avoir peur de l'évaluation ? ». La question mérite d'être posée, car l'évaluation peut avoir des conséquences insoupçonnées sur la vie des gens. Naturellement, nous sommes d'avis que nous ne devrions pas craindre l'évaluation, surtout lorsque nous prenons le soin de le faire dans le respect et l'intérêt des personnes. Ceci étant dit, avec l'évaluation viennent une responsabilité et une série d'obligations de compétences dont la lecture de ce livre permet de s'acquitter en grande partie. Non, il ne faut pas avoir peur de l'évaluation, mais craindre plutôt ceux ou celles qui s'improvisent « évaluateurs » sans se soucier de fonder leur prise de décision sur une science établie. Avec ce livre, nous espérons contribuer à poser un jalon dans le but de voir reconnaître l'importance d'une compétence essentielle à développer pour ceux et celles qui aspirent à devenir des professionnels de l'évaluation, tant en psychologie qu'en sciences de l'éducation.

Dany Laveault
Jacques Grégoire

CHAPITRE 1

LA CONSTRUCTION D'UN INSTRUMENT DE MESURE

1. Le processus de développement d'un test

La construction d'un test en psychologie ou en éducation est un processus complexe et, parfois, de longue haleine. C'est pourquoi ce processus doit être documenté pour en assurer le suivi dans le temps, mais aussi pour rendre compte du rationnel qui en a guidé les étapes de développement et de la mise en œuvre. Ce sont là des conditions essentielles pour que le test produise des résultats valides et soit utilisé dans des conditions détaillées et transparentes. Downing (2006) distingue douze étapes pour le développement d'un test :

1. Plan d'ensemble et fonctions du test
2. Définition du contenu
3. Spécification du test
4. Rédaction des items
5. Conception et assemblage du test
6. Édition et production du test
7. Administration du test
8. Notation et calcul d'un résultat
9. Établissement d'un seuil de réussite
10. Communication des résultats
11. Banque d'items
12. Rédaction d'un rapport technique

Tout concepteur de test ou utilisateur des résultats doit être en mesure de fournir la documentation appropriée et des explications pour justifier la confiance à accorder aux résultats obtenus au moyen de ce test et aux affirmations que l'on peut faire à partir de ceux-ci. C'est le but de ce premier chapitre qui abordera déjà les étapes préliminaires 1 à 5. D'autres étapes seront abordées plus en détail dans les chapitres 2 à 7.

1.1. Première étape : Le choix des fonctions et des modalités d'évaluation

La première étape du modèle de Downing (2006) consiste à établir un plan d'ensemble et à déterminer les fonctions du test. La première question que doit se poser la personne désireuse de construire ou d'utiliser un test consiste à se demander : « À quoi va-t-il servir ? ». Par exemple, un test de mathématique peut avoir pour fonction de sélectionner des sujets, de diagnostiquer des difficultés d'apprentissage ou encore de déterminer si un élève maîtrise les compétences attendues en fin d'année scolaire. De même, un questionnaire d'anxiété peut être utilisé pour recruter des personnes possédant certaines caractéristiques de personnalité ou pour évaluer l'effet d'un médicament anxiolytique. Le plus souvent, un même test ne remplit qu'une seule fonction à la fois.

Les usages prévus des résultats à un test affectent profondément les caractéristiques recherchées et vice versa. Ce sont les caractéristiques du test qui feront en sorte que les résultats obtenus permettront de soutenir la validité des affirmations que nous souhaitons faire et réciproquement, c'est le type d'évaluation que l'on compte faire des résultats qui conduiront au choix des caractéristiques préférables du test. À cette étape-ci, le concepteur de test devra définir non seulement la fonction du test, mais la modalité d'évaluation des résultats. Voici un aperçu des options :

Au niveau des fonctions du test :

1. **Évaluation bilan, sommative ou certificative**. Elles consistent toutes à porter un jugement sur un ensemble de performances pour, au minimum, en établir un bilan. En éducation, l'évaluation sommative présente ce bilan suite à une période d'apprentissage de durée variable selon l'importance du bilan. L'évaluation bilan et l'évaluation sommative sont dites « certificatives » lorsque les résultats sont comparés à un seuil de performance (par exemple, une note de passage) pour parvenir à une décision de réussite ou d'échec, de promotion ou tout autre type de décision, comme l'admission à un ordre professionnel. La fonction n'est pas de sélectionner les meilleurs candidats, mais ceux qui atteignent et dépassent le seuil de performance pour se qualifier.
2. **Évaluation formative**. C'est un mode d'évaluation continue qui consiste à porter attention aux résultats des apprentissages en cours de formation en vue d'en favoriser la progression. À la différence des trois précédentes, c'est une évaluation qui porte sur des aspects bien délimités et réduits des apprentissages en cours. Idéalement, elle implique l'examinateur et le sujet examiné dans une démarche de corégulation où la rétroaction et l'autoévaluation jouent un rôle important (Allal, 2020 ; Laveault & Allal, 2016). Cette démarche peut être fortement instrumentée et associée à une évaluation critériée (Scallon, 1988, 2000), ou encore très informelle, dans laquelle les tests sont remplacés par des méthodes d'observation ou des formes de dialogue ou d'interaction entre formateurs et formés.
3. **Évaluation diagnostique.** C'est un mode d'évaluation ponctuelle qui vise à déterminer les caractéristiques personnelles d'un individu qui peuvent avoir un impact sur sa capacité générale d'adaptation. En éducation, elle porte sur les caractéristiques, relativement stables d'un apprenant (intelligence, motivation scolaire, troubles de l'attention, hyperactivité, etc.), qui peuvent avoir une incidence sur l'apprentissage. En psychologie, elle porte sur des caractéristiques

d'intérêt telles que les traits de personnalité, le style de leadership (pour un emploi ou une promotion), le degré d'anxiété, la créativité, l'impulsivité, etc. Dans le contexte diagnostique, des seuils peuvent être utilisés pour indiquer le degré de gravité du trait évalué.

Au niveau des modalités d'évaluation des résultats, on distingue deux principales approches :

1. **Évaluation normative**. Elle consiste à juger du résultat du sujet examiné en le comparant aux résultats d'autres sujets de la même population. Son but est de différencier les individus entre eux. Pour y parvenir, les tâches qui discriminent les sujets seront privilégiées et des normes de comparaison interindividuelle seront établies. La principale fonction du test est de sélectionner les candidats sur la base de leurs résultats et l'accent est mis sur la variation et la différenciation des résultats individuels.
2. **Évaluation critériée.** Elle consiste à juger du résultat du sujet examiné en différenciant de façon précise les différents aspects de sa performance. Dans ce contexte, la différenciation entre individus perd de son importance : c'est la différenciation du rendement ou de la performance qui importe. Par exemple, deux individus pourraient mériter le même résultat global, mais en présentant des profils de réussite fort différents. Pour parvenir à différencier les réussites, la priorité sera mise sur la spécification de domaine du trait à évaluer.

La relation entre le test et sa fonction est parfois tellement forte que l'on parlera de « test critérié » autant que d'évaluation critériée, de « test normatif » autant que d'évaluation normative et ainsi de suite. Par exemple, on parlera de test normé à propos d'un test qui fournit des informations sur le degré de maîtrise de la langue seconde d'un individu par rapport à l'ensemble de la population. Les résultats à un test normé de ce genre pourraient être utilisés pour le classement des sujets examinés dans différents programmes d'apprentissage de la langue seconde : débutant, intermédiaire, enrichi. Par contre, un test diagnostique critérié pourrait être utilisé pour diriger les mêmes candidats vers des formations de rattrapage pour mettre à niveau leurs compétences de la langue seconde. De plus, une série de tests à visée formative permettrait de surveiller la progression en cours d'apprentissage et d'effectuer des ajustements au besoin. Pour une grande entreprise qui vise à hisser 80 % de ses employés à un niveau de bilinguisme de niveau intermédiaire, une telle combinaison de tests normatifs, diagnostiques, critériés et formatifs peut se traduire par une plus grande efficacité des formations et des économies substantielles.

Le choix du type d'épreuve conditionne la méthodologie utilisée pour le développement et l'utilisation du test. Des techniques particulières doivent être utilisées pour obtenir des tests produisant les résultats aux propriétés métriques dont on a besoin. Par exemple, le temps d'administration du test est déterminé en partie par l'usage anticipé des résultats. Dans l'exemple précédent, nous pourrions envisager d'avoir recours à un test à durée limitée si le but était de sélectionner les candidats possédant le niveau le plus élevé de bilinguisme pour une promotion. Une telle limite de temps serait moins justifiée dans le cas d'un test de classement et encore moins pour un test à visée formative. Le temps d'administration n'est qu'une des variables à prendre en considération selon la fonction du test. Le degré de standardisation des conditions d'administration, très important dans le cas de tests à enjeux élevés,

l'est beaucoup moins dans le cas d'un test à visée diagnostique. Ce degré d'interrelation entre toutes les étapes de développement d'un test fait ressortir l'importance d'avoir un cadre de référence précis dès le départ et de le documenter de façon adéquate afin d'en expliquer le rationnel.

On ne peut faire l'épargne d'une réflexion approfondie sur l'usage auquel on destine un test. Au point de départ du travail de construction, un choix doit être opéré entre différentes fonctions possibles et les modalités d'évaluation qui leur conviennent le mieux. Il est illusoire de croire qu'un test « généraliste » puisse répondre simultanément à plusieurs fonctions, comme différencier les contenus d'apprentissage pour chaque individu et ordonner des sujets examinés selon leurs résultats globaux. Dans la section 2, cette question sera approfondie dans le cas du développement d'un test d'acquis scolaires.

1.2. Deuxième étape : La définition de ce que l'on souhaite mesurer

Même lorsque la fonction du test a été précisée à l'étape précédente, son contenu est relativement vague et général à ses débuts : « évaluer la compréhension en lecture à l'école primaire », « apprécier le développement social de 3 à 6 ans », « diagnostiquer les troubles de la mémoire », « sélectionner du personnel de bureau », etc. Qu'entend-on par « troubles de la mémoire », « développement social » et « compréhension de la lecture » ? Ces contenus sont encore beaucoup trop vagues pour permettre réellement de débuter la construction d'un test.

La deuxième étape du modèle de Downing consiste à effectuer un travail d'approfondissement des concepts et d'opérationnalisation de ceux-ci. Il s'agit de définir avec précision les caractéristiques psychologiques ou éducatives que le test devra mesurer. Sur la base de cette définition, des items pourront alors être construits. C'est sur ce travail préalable de définition et d'analyse de ce que l'on veut mesurer que reposera la validation du contenu du test (voir chapitre 4, section 2).

Mais comment passer d'une intention vague à la définition opérationnelle d'un concept ? Lorsque l'on veut sonder une population de citoyens pour connaître son opinion, il importe d'en connaître les principales caractéristiques et la manière dont celles-ci se distribuent afin d'en tenir compte lors de l'échantillonnage des sujets. Il en va ainsi pour l'analyse du domaine que l'on souhaite tester. Cette analyse du domaine prépare la spécification de domaine du test lui-même à partir des caractéristiques propres au trait que nous souhaitons évaluer. Selon les domaines, le concepteur de tests peut avoir recours à plusieurs modèles d'analyse détaillés ci-après :

1. **L'analyse de contenu d'entretiens**. Lorsque le praticien n'a pas d'idées précises à propos des caractéristiques permettant de discriminer les individus qui seront évalués par le test, il est intéressant de commencer par interroger des personnes appartenant à la population visée par ce test. L'interview, libre ou semi-structurée, permet de recueillir un grand nombre d'informations qui seront sélectionnées et classées au moyen d'une analyse de contenu. Par exemple, Hunt et McKenna (1992) ont procédé de la sorte pour mettre au point un questionnaire de qualité de vie destiné à des patients dépressifs. Cinq psychiatres ont interviewé 30 patients dépressifs à propos de différentes facettes de leur vie quotidienne. Une analyse de contenu des entretiens a permis de

mettre en évidence un certain nombre de propositions caractéristiques permettant d'apprécier la qualité de vie des patients dépressifs. Ces propositions ont ensuite servi à spécifier le domaine pour construire les items du questionnaire.

2. **L'observation directe des comportements**. Dans certains cas, plutôt que d'interroger les personnes, il est préférable de les observer dans leur milieu de vie ou de travail. Cette méthode a été utilisée par Binet pour construire le tout premier test d'intelligence de l'histoire. Au début du XX^e^ siècle, Binet ne pouvait s'appuyer que sur un modèle rudimentaire et vague de l'intelligence. Dès 1900, il commença donc à observer les handicapés mentaux adultes de l'Hôpital Sainte-Anne et les enfants d'une école d'un quartier populaire de Paris afin de mettre en évidence les comportements permettant de distinguer les individus sans handicap intellectuel des individus handicapés mentaux. Les items de l'échelle métrique d'intelligence de 1905 sont issus de ce travail d'observation.
3. **La méthode des incidents critiques**. L'origine de cette méthode est attribuée à Flanagan (1954). Elle est particulièrement utile pour construire des outils d'évaluation des performances professionnelles. Elle consiste à demander à des responsables de décrire des situations de travail où les employés, sous leurs ordres, ont agi de manière particulièrement efficace ou, au contraire, inefficace. Partant de cette description, certains comportements « critiques » peuvent être mis en évidence et servir à construire des échelles d'évaluation. Selon Jaeger (1994), une analyse de domaine d'emploi est une condition sine qua non pour la conception et l'élaboration de tests d'agrément ou de certification en milieu professionnel. Les *Standards for educational and psychological testing* précisent en effet : « Une analyse de poste doit inclure une analyse des comportements de travail importants requis pour une performance réussie et leur importance relative » (1985, cité dans Jaeger, 1994).
4. **La référence à un modèle théorique**. À la différence des autres méthodes, celle-ci ne part pas de l'expérience ou de l'observation, mais d'un modèle de la réalité construit au cours de recherches antérieures. Depuis le début des années 1980, les développements de la psychologie ont conduit à la création de nombreux modèles théoriques utilisables par les constructeurs de tests. Des tests destinés au diagnostic des troubles de la lecture ont, par exemple, été créés sur base de modèles décrivant les processus impliqués dans l'activité de lecture (par exemple, de Partz, 1994 ; Mousty *et al.*, 1994). D'autres outils ont également été construits en référence à des modèles théoriques pour évaluer des caractéristiques aussi diverses que le calcul, la motivation, la mémoire, la créativité, l'intelligence, etc. L'analyse du domaine doit bien faire ressortir dans quelle mesure la conception du test est fidèle au modèle théorique de référence. Par exemple, un test d'intelligence, fondé sur la théorie du développement cognitif de Piaget, devrait être fort différent d'un test d'intelligence, basé sur le modèle de la structure de l'intellect de Guilford. Il en va de même pour de nombreux concepts théoriques de la psychologie, tels que la motivation, la créativité, le leadership. En éducation, les modèles théoriques de l'apprentissage et de l'enseignement peuvent faire partie de l'analyse du domaine au même titre que les programmes d'études, le curriculum ou les standards d'apprentissage.
5. **Les objectifs de formation**. Lorsqu'il s'agit d'évaluer des apprentissages scolaires, la démarche la plus fréquente consiste à analyser l'importante documentation

portant sur les objectifs de formation des programmes d'études, le curriculum et les standards de progression des apprentissages tels que définis par les autorités gouvernementales ou les juridictions scolaires locales ou nationales.

L'analyse de domaine est à la base de la spécification de domaine du test et permet de déterminer le contenu des questions et tâches qui permettront de se prononcer sur les résultats au test. La section 2 du présent chapitre présente en détail différents modèles de spécification de domaine, ainsi que d'autres méthodes permettant de préciser encore davantage les caractéristiques que doit posséder un test dans le cas bien précis de l'évaluation d'acquis scolaires.

1.3. Troisième étape : La création des items

Georges Gallup (1947), fondateur du célèbre institut de sondage du même nom, affirmait :

> Trop d'attention a été accordée à la constitution des échantillons et trop peu à la création des questions […] Des différences dans la construction des questions conduisent souvent à des résultats qui présentent de plus grandes variations que celles habituellement observées en fonction des différentes techniques d'échantillonnage. (p. 383)

Cette constatation garde toute son actualité et peut être généralisée aux questions construites pour les tests psychologiques et les tests d'acquis scolaires. Souvent, les praticiens ne suivent aucune méthodologie particulière pour construire les items. Ayant en tête ce qu'ils souhaitent mesurer, ils se fient à leur intuition pour préparer les questions. Pourtant, il est indispensable d'avoir un projet et un plan précis avant de se lancer dans la production d'items :

1. **Quel format d'items choisir ? Pourquoi ?** Le choix d'un format ne doit pas être arbitraire. Il découle d'un ensemble de contraintes concernant les objectifs du test et les conditions matérielles de création, de passation et de cotation de celui-ci. En conséquence, il n'y a pas de bon format d'item dans l'absolu. Un format est bon s'il est adéquat au but et à la situation d'évaluation. La section 3 du présent chapitre aborde de manière détaillée la question du choix du format d'item et des règles de construction des questions fermées et ouvertes.
2. **Quel doit être le niveau de difficulté des items ?** Dans le cas de tests d'aptitudes ou de rendement scolaire, le choix du niveau de difficulté des items dépend de l'objectif du test. Ce niveau variera selon que le test est normé ou critérié, sommatif ou formatif. En d'autres termes, c'est la nature des informations que nous désirons recueillir qui va déterminer le niveau de difficulté des items à produire. Un raisonnement similaire s'applique aux tests non cognitifs. Par exemple, une question d'un test de personnalité qui ferait l'objet d'un assentiment voisinant 100 % ne sera pas approprié, car il ne permettra pas de différencier les répondants entre eux. Les tests qui visent à différencier les individus doivent en effet comprendre des items ayant une certaine variance.
3. **Combien faut-il créer d'items ?** Le nombre d'items à créer dépend de plusieurs facteurs. Le premier facteur est la durée du test. Selon que l'on prépare un test court, pouvant être passé en 10 minutes, ou un test diagnostique se

déroulant sur plusieurs séances d'examen, le nombre d'items à créer variera considérablement. Un second facteur à prendre en compte est le niveau désiré de fidélité. Un test long fournira des résultats plus fidèles qu'un test court. Par ailleurs, si le test comporte plusieurs sous-scores, il sera nécessaire d'assurer la fidélité de ceux-ci en prévoyant suffisamment d'items dans chacune des sous-échelles du test. Enfin, un dernier facteur à prendre en considération est l'élimination, quasi inévitable, de certains items après leur évaluation par des experts et leur mise à l'essai. Si on veut que la version finale du test contienne assez d'items, il faudra donc en créer plus que le strict nécessaire. Si, par exemple, le test final doit contenir 20 items, on en créera 30 et l'on retiendra les 20 meilleurs de ceux-ci. Habituellement, un surplus de 30 à 50 % d'items doit être prévu pour conserver un nombre suffisant d'items après la mise à l'essai.

4. **Un item unique peut-il suffire ?** Les caractéristiques que nous souhaitons mesurer sont souvent complexes et nécessitent d'être mesurées à l'aide de plusieurs items qui permettent d'en saisir les différentes facettes. C'est le cas d'un trait de personnalité comme l'extraversion ou de la maîtrise d'un apprentissage scolaire comme la multiplication des fractions. Dans ces deux cas, une mesure, qui se baserait sur un seul item, soulèverait des problèmes de validité et de fidélité des résultats. L'item choisi ne pourrait pas couvrir toute l'étendue de la caractéristique à mesurer. L'erreur de mesure serait également plus importante que celle observée avec une mesure de la même caractéristique basée sur plusieurs items. Par conséquent, l'usage d'un item unique est généralement déconseillé. Pourtant, ce choix peut parfois être pertinent. C'est le cas lorsque la caractéristique mesurée est unidimensionnelle, d'étendue réduite et définie de manière précise. Si, par exemple, nous souhaitons connaître le degré de satisfaction des individus par rapport à leur salaire ou l'intensité actuelle de leur douleur lombaire, il n'est pas nécessaire d'évaluer ces caractéristiques à l'aide de plusieurs items. Des questions uniques telles que « Êtes-vous satisfait de votre salaire actuel ? » et « Quelle est l'intensité actuelle de votre douleur lombaire ? », associées à des échelles bipolaires correctement labélisées et précédées de consignes claires, permettent de récolter des informations valides et fidèles. Allen *et al.* (2022) ont identifié plusieurs avantages de tels items uniques. Pour les répondants, ils apparaissent pertinents, sans ambiguïté, exempts de jugement et faciles à compléter. Ils permettent aussi d'économiser du temps de passation et de traitement. Par ailleurs, ils évitent le sentiment de lassitude qu'on peut parfois observer avec des échelles comprenant de nombreux items redondants. Il faut souligner que l'évaluation de la validité et de la fidélité des mesures basées sur un seul item requiert des méthodes sensiblement différentes de celles utilisées pour les échelles de mesure basées sur plusieurs items.

1.4. Quatrième étape : L'évaluation des items et leur mise à l'essai

Avant d'aborder la question de l'évaluation des items, il importe de préciser ce que nous entendons par « item ». La définition de l'item a considérablement évolué pour inclure des formats d'items qui ne se limitent pas aux diverses questions à réponses courtes (QRC) et aux questions à choix multiple (QCM). La définition courante

d'item de l'IMS (Instructional Managerial System, 2002, cité dans Vale, 2006) fait référence à des ensembles plus vastes et plus complexes, tels que les « testlets » (Wainer *et al.*, 2007) et autres formes d'items rendues possibles par le recours aux médias sonores et visuels et à la technologie (ordinateurs, tablettes, etc.). La définition actuelle d'item est donc élargie. Il s'agit du : « bloc de base qui contient une ou plusieurs questions et réponses » (IMS, 2002, cité dans Vale, 2006), ces blocs représentant les « objets » les plus petits qui soient interchangeables dans une banque d'items (Vale, 2006). Lorsque le terme « item » est utilisé, il faut donc le comprendre dorénavant dans sa conception élargie.

Dans ce nouveau contexte, une définition précise de ce qu'on souhaite mesurer et une méthodologie rigoureuse de construction des items sont devenues des conditions nécessaires pour obtenir des items valides appropriés à la fonction du test. Pour garantir les propriétés métriques des items, une évaluation minutieuse de ceux-ci doit également être réalisée. Deux démarches complémentaires sont habituellement suivies pour réaliser cette tâche.

1. **Une évaluation des items par des juges/experts**. Ceux-ci sont chargés d'apprécier la conformité des items aux exigences définies lors de la seconde étape du processus de construction du test. Les méthodes d'évaluation des items par des juges sont détaillées dans la section 2 du chapitre 4.
2. **Une mise à l'essai des items suivie d'une analyse des items.** La mise à l'essai complète l'appréciation des items par les juges. Cette dernière évaluation reste en effet subjective malgré la rigueur méthodologique avec laquelle elle peut être réalisée. La mise à l'essai permet de recueillir des données empiriques directement de la population à laquelle est destiné le test et de procéder à une analyse statistique des items (voir chapitre 5).

La mise à l'essai consiste à faire passer tous les items à un échantillon de la population. Cet échantillon ne doit pas nécessairement être représentatif (pour une discussion de cette notion, voir chapitre 6), ni de très grande taille. Sa taille dépend en fait de l'hétérogénéité de la population visée par le test et de la grandeur de la population de référence. Par exemple, si un questionnaire de stress est destiné à évaluer uniquement des pilotes d'avion, une mise à l'essai sur un échantillon de 50 pilotes permettra généralement une évaluation satisfaisante des items, car la population des pilotes d'avion est plus homogène et de plus petite taille que la population en général. Par contre, si la population est plus hétérogène, un échantillon de 200 à 300 personnes peut être nécessaire pour réaliser une mise à l'essai valable sur un échantillon dont les caractéristiques épousent une distribution proportionnelle à celle de la population. Par exemple, la mise à l'essai des items de la version française du WISC-V (*Wechsler Intelligence Scale for Children* – version 5) a été réalisée sur un échantillon de 224 enfants. Ce test est destiné à évaluer tous les enfants français entre 6 et 16 ans. Dans ce cas, l'échantillon de la mise à l'essai doit être de plus grande taille, car il doit inclure des enfants des deux sexes, de différents âges et de différents milieux sociaux. On ne vise toutefois pas à ce qu'un tel échantillon soit parfaitement représentatif de la population. Il doit avant tout refléter l'hétérogénéité de celle-ci. Un échantillon trop homogène risque en effet de masquer certains items problématiques. Par exemple, si les items d'un questionnaire de dépression destiné à des personnes âgées sont prétestés sur un échantillon qui ne comprend que des retraités possédant un diplôme d'études supérieures, certains problèmes risquent de passer inaperçus.

L'inclusion de personnes âgées possédant le seul diplôme d'études primaires aurait permis de mettre en évidence des questions dont le vocabulaire trop complexe peut entraîner des erreurs de compréhension.

Les résultats d'une mise à l'essai sont analysés tant d'un point de vue qualitatif que quantitatif. En particulier, les commentaires des répondants à propos des items peuvent se révéler précieux pour comprendre des résultats aberrants et pour remédier à certains problèmes de formulation des questions. Pour la mise à l'essai de tests d'apprentissage en milieu scolaire, des « laboratoires cognitifs » peuvent être utilisés pour tester les items au niveau individuel ou de la salle de classe avant une mise à l'essai formelle. De même, les problèmes de manipulation du matériel, d'enregistrement des réponses, de temps de passation, de cotation des réponses peuvent être repérés à cette occasion. Ces problèmes, en apparence mineurs, doivent retenir toute l'attention du constructeur, car ils peuvent diminuer considérablement la validité des résultats d'un test. C'est le cas d'un espace trop petit pour noter les réponses, d'un livret de test difficile à manipuler ou de l'affichage irrégulier d'une page-écran sur différents modèles d'ordinateur.

En plus de ces vérifications qualitatives, la mise à l'essai permet de réaliser différentes analyses statistiques des résultats. Celles-ci sont détaillées dans le chapitre 5 consacré à l'analyse des items. Ces analyses portent, entre autres, sur la difficulté des items, leur discrimination, leur fonctionnement différentiel pour différents groupes d'examinés. Sur la base de ces analyses et des observations qualitatives, les meilleurs items seront finalement sélectionnés et serviront à construire la version définitive du test.

1.5. Cinquième étape : L'analyse des propriétés métriques de la version finale du test

Une fois les meilleurs items sélectionnés et la version définitive du test constituée, il reste à déterminer les propriétés métriques des résultats. Celles qui doivent retenir l'attention du constructeur varient en fonction de la nature du test. S'il s'agit d'un test normé, il sera nécessaire d'établir des normes et de présenter celles-ci selon une échelle aisément compréhensible par les praticiens. S'il s'agit d'un test critérié, il faudra préciser des scores de référence utiles, tels qu'un seuil de réussite ou des seuils de performance correspondants à différents niveaux de rendement. Si les résultats du test doivent être mis en relation avec ceux d'autres tests, il y aura lieu de mettre en équivalence les échelles de mesure concernées. Les techniques nécessaires pour déterminer les normes, les scores de référence et les équivalences sont présentées en détail dans le chapitre 6.

Par ailleurs, une investigation approfondie de la validité et de la fidélité des résultats de la version finale du test devra toujours être réalisée. Le constructeur doit rassembler des preuves de la validité des inférences permises par les résultats au test. Par exemple, s'il propose aux praticiens de calculer et d'interpréter différents sous-scores au test, il sera nécessaire de prouver la pertinence de tels sous-scores quant à l'interprétation qui en est faite (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014, p. 20). Les fondements et la méthodologie de telles études de validité sont explicités dans le chapitre 4.

Il faut souligner que l'évaluation de la validité des résultats à un test n'est pas du seul ressort du constructeur. Elle est partagée par l'utilisateur du test. En fait, la validité n'est jamais une qualité acquise une fois pour toutes. Chaque nouvelle inférence qu'un praticien veut réaliser à partir des résultats doit faire l'objet d'une validation spécifique. Par exemple, si un test de mémoire a été créé pour évaluer les compétences mnésiques des enfants et des adolescents, la pertinence de l'usage de ce test avec des adultes devra être démontrée sur base de données empiriques si on souhaite interpréter ces résultats pour cette nouvelle population d'intérêt.

Le constructeur devra également apporter des informations à propos de la fidélité des résultats. Il peut choisir parmi une variété d'indicateurs tels que le coefficient de fidélité et les autres mesures liées à celui-ci et les communiquer aux praticiens et utilisateurs des résultats à ce test, ce qui inclut, sans s'y restreindre, l'erreur de mesure de scores, les intervalles de confiance, l'erreur de mesure des différences entre scores, etc. Les techniques nécessaires pour calculer ces valeurs relatives à la fidélité sont présentées de manière détaillée dans le chapitre 3.

Lorsqu'un test n'est pas réservé au seul usage de son constructeur, mais est destiné à être diffusé, la rédaction d'une documentation destinée aux utilisateurs est essentielle (American Educational Research Association *et al.*, 2014, pp. 67-70). Cette documentation doit présenter un rapport technique sur les données métriques citées ci-dessus (normes, coefficient de fidélité…) ainsi que les bases théoriques du test, les fonctions pour lesquelles il a été créé et les qualifications requises pour pouvoir l'utiliser et en interpréter correctement les résultats. Le constructeur d'un test a une responsabilité méthodologique, mais également éthique. L'instrument qu'il a créé va en effet servir à évaluer des personnes et à prendre des décisions à leur propos. Les informations communiquées dans le manuel d'accompagnement doivent permettre de garantir un usage correct des résultats au test dans le respect des principes déontologiques et des normes de pratique reconnues. En plus des références déjà mentionnées, il existe plusieurs autres références en ce domaine, notamment un numéro spécial de la revue *Mesure et évaluation en éducation (volume 20, n° 2, 1997)* entièrement consacré à ces questions dans le domaine de l'éducation. Pour une perspective plus générale concernant l'éthique, Hadji (2012) aborde plusieurs questions philosophiques formant la base d'une éthique en évaluation.

2. La construction d'un test d'acquis scolaires

La construction de tests d'acquis scolaires se démarque de la construction de la plupart des autres instruments de mesure sur plusieurs points qui sont abordés dans cette section. Utilisés et développés par les enseignants, mais aussi par des spécialistes des matières et des programmes, ils sont employés tant dans la salle de classe pour attester des apprentissages des élèves que dans l'ensemble d'une juridiction scolaire à des fins de régulation du système éducatif. La fonction du test dans ces deux situations est spécifique et doit être prise en considération. Le niveau d'exigence ne saurait être le même pour un examen préparé par un enseignant en salle de classe que pour un test standardisé à grande échelle développé par un comité d'experts d'une grande juridiction scolaire.

Si les enseignants sont des professionnels de l'apprentissage et de l'enseignement, leur formation en matière de construction d'instruments de mesure est,

par contre, plus limitée. On ne peut donc pas s'attendre à ce qu'ils suivent, dans la conception et la rédaction des examens qu'ils destinent à leurs élèves, les mêmes étapes que l'on retrouve dans les tests standardisés, commerciaux ou à grande échelle. Les conditions et les conséquences des examens préparés par les enseignants ne sont pas comparables.

Quelle confiance et quelle importance accorder alors à l'évaluation des acquis scolaires par les enseignants ? Pour répondre adéquatement à cette question, il faut prendre en considération bien plus que les critères habituels de rigueur statistique et métrique. Il faut prendre également en compte le contexte particulier de l'évaluation scolaire. En effet, celle-ci se caractérise par des prises d'information fréquentes et diversifiées s'échelonnant sur de longues périodes. Elle porte sur l'apprentissage, une caractéristique de l'élève en constante évolution d'autant plus que chez les élèves plus jeunes, le développement cognitif n'est pas terminé. Ceci fait en sorte que l'enseignant doit être en mesure de faire de nombreuses observations et prises d'information qui doivent être regroupées et triangulées (CTREQ, 2018) pour lui permettre de formuler un jugement sur les acquis scolaires, possibilité qui n'est pas à la portée d'une seule prise d'information. Dans la pratique, ce jugement est généralement nuancé et assez valide surtout lorsqu'il est bien aligné sur les apprentissages qu'il est censé mesurer.

L'un des plus grands défis de l'évaluation des acquis scolaires réside dans le fait de faire tenir une grande quantité d'informations dans une simple note. La possibilité pour l'enseignant de faire de nombreuses prises d'information, d'observer les apprentissages de façon continue et individuellement à l'occasion, la possibilité de reprendre des examens aux résultats douteux ou insatisfaisants et la connaissance approfondie qu'il a pu acquérir de chaque élève au cours de ses nombreuses interactions avec celui-ci, sont à la fois sources d'opportunités, mais aussi de défis particuliers. La quantité d'information recueillie est à la fois un avantage par rapport à une seule épreuve, mais présente quelques inconvénients lorsqu'il s'agit de regrouper des observations et des résultats qui peuvent être fluctuants et pas toujours cohérents ou stables. Dans le contexte des examens scolaires, il est possible d'avoir des exigences de rigueur, mais il serait utopique de s'attendre à ce que celles-ci soient de même nature que celles des épreuves psychométriques ou de toute épreuve unique donnant lieu à des scores numériques ou profils de scores standardisés.

2.1. Alignement

Une fois admis le principe que la qualité des examens de rendement scolaire ne peut être régie par les mêmes critères que ceux de la majorité des instruments de mesure, la question consiste alors à identifier quels devraient être ces critères particuliers. Selon l'OCDE (2013), ainsi que Donaldson (2013) et Looney (2011), l'alignement de l'enseignement et de l'évaluation sur le programme de formation constitue l'un des principaux enjeux de l'évaluation scolaire. L'alignement est le degré d'appariement entre le test et le contenu de la matière tel qu'identifié au moyen des standards académiques (LaMarca, 2001). Ce critère de l'alignement s'avère particulièrement pertinent, qu'il s'agisse d'évaluation en salle de classe ou d'examens à grande échelle et à enjeux élevés. À la différence des tests psychométriques qui peuvent porter sur un trait individuel stable (par exemple, le niveau d'intelligence, le type de motivation, etc.), les examens de rendement scolaire mesurent des apprentissages qui sont appelés à

évoluer rapidement. Un bon alignement assure que tous les enseignants ou formateurs qui ont pour mission de faire apprendre rament dans la même direction et si possible, en cadence. Un défaut dans l'alignement peut créer des difficultés d'apprentissage ultérieures qui seront d'autant plus difficiles à surmonter que l'apprenant manifeste déjà des signes de difficultés au niveau cognitif ou au niveau motivationnel. Dans le contexte particulier de l'examen de rendement scolaire, il est primordial de s'assurer que ce qui est enseigné et appris s'appuie sur des prérequis essentiels pour assurer la progression prévue des apprentissages et éviter le cumul de retards scolaires.

L'enseignant doit donc s'assurer que son enseignement s'aligne sur le programme de formation et que l'évaluation des apprentissages en est le reflet fidèle. L'alignement est la condition essentielle qui fait en sorte que chaque enseignant, qui accueille une nouvelle cohorte d'élèves, puisse compter sur des préacquis minimaux avec lesquels poursuivre la progression des élèves. Pour réussir l'alignement, il importe donc que tous les formateurs suivent le programme de formation et l'interprètent de manière univoque. Pour y arriver, les programmes de formation se doivent de communiquer les attentes de résultats d'apprentissage le plus clairement possible, que ce soit en termes de contenu ou de niveaux de rendement.

Les objectifs d'apprentissage des programmes de formation sont donc des outils de communication précieux à l'intention des enseignants et des examinés. À la différence d'un test d'intelligence pour lequel un enfant ou un adulte n'a besoin d'aucune préparation, le test d'acquis scolaires nécessite que l'élève soit non seulement préparé, mais aussi correctement informé de ce qu'il aura à faire. En effet, en cas de mauvaise performance à un examen, il faut faire la distinction entre l'absence de réussite qui résulte de difficultés d'apprentissage, d'une absence de réussite qui provient d'un manque de préparation faute d'une communication adéquate des attentes à l'élève. Fournier et Laveault (1994) ont démontré que les élèves qui obtiennent les meilleurs résultats à un examen de rendement scolaire sont également ceux qui réussissent le mieux à anticiper sur quoi l'examen va porter. Laveault et Miles (2008) ont également montré que les élèves qui peuvent différencier correctement les niveaux de qualité de productions écrites savent mieux comment mettre en pratique les critères d'évaluation à propos de leur propre production.

Les résultats scolaires dépendent donc de plusieurs facteurs et les causes de désalignement sont multiples :

- Le contenu enseigné est inapproprié ;
- Le niveau d'exigence est inadapté : trop ou pas assez élevé ;
- La matière est mal assimilée ou insuffisamment intégrée à cause d'une pédagogie peu différenciée ou de stratégies d'enseignement inefficaces ;
- L'évaluation faite par l'enseignant n'est pas pertinente.

Si ces nombreux facteurs sont une cause légitime de préoccupation et pourraient nous faire remettre en question la pertinence de l'évaluation du rendement scolaire, la contrepartie positive de l'alignement est non négligeable. Un enseignement aligné de manière constructive met à profit l'effet puissant de l'évaluation sur les expériences d'apprentissage des élèves. Dans un tel contexte, l'évaluation peut devenir réellement formative tant pour l'apprentissage des élèves que pour le développement continu des enseignants par un effet de feedback inversé des résultats des élèves vers le formateur (Laveault et Allal, 2016).

La rigueur de l'évaluation scolaire est donc à rechercher non seulement dans la pertinence des observations réalisées sur les élèves au moyen des examens, mais aussi dans la qualité de la communication des attentes d'apprentissage. La validité des résultats aux examens de rendement scolaire dépend donc de facteurs externes liés à la communication des attentes, et ce à plusieurs niveaux :

- **Entre les enseignants et les juridictions scolaires**. Les concepteurs de programme ont la responsabilité de faire en sorte que les attentes d'apprentissage s'appuient sur des modèles éprouvés et soient claires et précises afin d'être bien comprises des enseignants.
- **Entre enseignants**. C'est en travaillant collectivement en tant que communauté d'apprentissage professionnel que les enseignants peuvent développer une représentation commune des attentes d'apprentissage et des niveaux de rendement. Lors de la préparation d'examens ou lors de la correction des copies des élèves, le travail en équipe permet aux enseignants d'aplanir les divergences d'interprétation. La modération sociale (modération fondée sur les standards ou harmonisation) permet de limiter la variation des interprétations individuelles (Laveault & Yerly, 2017 ; Wyatt-Smith *et al.*, 2010).
- **Entre enseignants et élèves**. L'enseignant doit à son tour transmettre à l'élève les objectifs de la formation. Plusieurs moyens sont à sa disposition pour faire comprendre les attentes. L'évaluation formative des apprentissages est une bonne occasion pour établir une forme de dialogue avec l'apprenant à ce sujet. Les copies types de productions peuvent également aider l'apprenant à se faire une meilleure représentation du résultat qui est attendu de sa part (Laveault & Bourgeois, 2014). Enfin, la capacité de l'apprenant à réfléchir par lui-même sur ses apprentissages peut faire en sorte que ses chances de bien performer à un examen de rendement scolaire dépendent non seulement de sa capacité à apprendre, mais aussi de son habileté à réfléchir sur ce qui est important de préparer en vue de l'examen. Ceci assume que l'examen est pertinent et correctement aligné.

Les tests d'acquis scolaires à grande échelle, organisés par les autorités scolaires, ont aussi pour fonction de permettre aux décideurs, gestionnaires et administrateurs des systèmes d'éducation d'utiliser les résultats d'apprentissage des élèves comme indicateur de l'efficacité du système éducatif pour vérifier, entre autres choses, le degré d'alignement. Celui-ci se mesure par le degré de concordance entre les évaluations faites en classe et les examens uniformes préparés par les autorités scolaires pour l'ensemble des élèves suivant le même programme de formation. Les tests de rendement scolaire peuvent difficilement remplir adéquatement toutes ces fonctions à la fois. C'est ainsi que les tests développés pour assurer la régulation du système éducatif auront des caractéristiques et des propriétés bien différentes de celles qui servent à certifier les apprentissages.

Les tests d'acquis scolaires se démarquent donc par le rôle important que joue la spécification de domaine des objectifs de formation et par la nécessité de transmettre ces objectifs à tous les niveaux du système d'éducation. Le choix d'un modèle de spécification de domaine dépend des fonctions et usages prévus des résultats et demande une attention particulière. C'est l'objet de la section 2.2. La section 2.3 présente plusieurs modèles et taxonomies d'objectifs de formation qui ont pour but de standardiser le vocabulaire servant à décrire ce qui doit être appris et évalué pour améliorer la communication des attentes d'apprentissage.

2.2. Les multiples fonctions de l'examen

Dans l'enseignement, les tests sont appelés à jouer plusieurs rôles. L'instrument de mesure sera construit différemment selon la fonction à laquelle on le destine et il ne se limite pas aux tests de type questionnaire. Voici quelques usages courants des instruments de mesure en contexte scolaire :

1. Dresser un bilan des acquis de l'élève.
2. Prendre une décision sur la promotion de l'élève.
3. Sélectionner les élèves selon certaines caractéristiques particulières afin de former des groupes.
4. Identifier les aspects de la résolution d'un problème source de difficultés individuelles ou de groupe.
5. Identifier les transferts d'apprentissage qui ont ou n'ont pas eu lieu.
6. Préparer une révision de la matière à partir des points pour lesquels certains élèves éprouvent des difficultés.
7. Faire prendre conscience aux élèves de certains points importants et mal maîtrisés de la matière.

Cette liste n'est pas exhaustive. Elle illustre simplement deux grands ensembles de situations où les tests exercent des rôles différents en situation scolaire :

a) « l'évaluation sommative » (situations 1, 2 et 3) ;

b) « l'évaluation formative » (situations 4, 5, 6 et 7).

Dans le premier cas, on recherche un instrument de mesure qui dresse un bilan à partir d'un échantillon de la matière enseignée. Un bon bilan nécessite un échantillonnage du contenu qui soit exhaustif et représentatif. De plus, l'instrument de mesure se doit d'être correctement aligné sur les programmes de formation. L'évaluation sommative devient certificative lorsque les résultats obtenus contribuent à la prise de décision sur la certification (bulletin scolaire, obtention de crédits en vue du diplôme, etc.). Si toute évaluation certificative est forcément sommative, la réciproque n'est pas vraie : une évaluation sommative peut servir occasionnellement à faire une mise au point sur les apprentissages d'un groupe-classe, avec comme objectif, non pas de noter les élèves, mais de déterminer leur degré de maîtrise des objectifs d'apprentissage. Un tel bilan devient une source de rétroaction, tant pour l'enseignant que pour les apprenants et s'approche de l'autre rôle de l'évaluation scolaire : l'évaluation formative.

Dans le cas de l'évaluation formative, la prise d'information porte habituellement sur une cible d'apprentissage très précise ou une difficulté particulière. Scallon (2000) définit ainsi l'évaluation formative :

> Processus d'évaluation continue ayant pour objet d'assurer la progression de chaque individu dans une démarche d'apprentissage, avec l'intention de modifier la situation d'apprentissage ou le rythme de cette progression, pour apporter (s'il y a lieu) des améliorations ou des correctifs appropriés.

Alors que plusieurs objectifs peuvent être pris en compte dans un bilan, l'évaluation formative peut ne porter que sur un seul objectif ou un seul aspect de celui-ci, par exemple l'évaluation de prérequis bien précis. Elle peut s'appuyer sur

une approche instrumentée (Scallon, 1988), mais aussi sur des approches informelles (dialogue, discussion en groupe, rétroactions individuelles, etc.).

Depuis 1999, une nouvelle terminologie initiée par le *Assessment Reform Group* au Royaume-Uni a tendance à se substituer aux expressions « évaluation formative » et « évaluation sommative », introduites par Scriven (1967) en évaluation de programmes puis étendues par Bloom, Hasting et Madaus (1971) au domaine de l'évaluation des apprentissages. Cette nouvelle terminologie recoupe en grande partie le champ des appellations plus anciennes. Dans le cas de l'évaluation formative renommée « évaluation soutien d'apprentissage », l'accent est mis sur le caractère informel de l'évaluation, comme en témoigne la définition adoptée en 2009 à Dunedin (Nouvelle-Zélande) dans le cadre de l'*International Symposium on Assessment for Learning* :

> L'évaluation-soutien d'apprentissage fait partie des pratiques quotidiennes des élèves et des enseignants qui, individuellement et en interaction, recherchent, réfléchissent sur et réagissent à l'information provenant d'échanges, démonstrations et observations afin de favoriser les apprentissages en cours (Allal & Laveault, 2009).

Cette nouvelle terminologie est maintenant suffisamment répandue, tant en anglais qu'en français, pour qu'on en tienne compte dans les publications scientifiques en évaluation scolaire (Laveault, 2013). En anglais, les expressions suivantes sont utilisées : *assessment of learning* (évaluation des apprentissages), *assessment for learning* (évaluation-soutien d'apprentissage) et *assessment as learning* (évaluation en tant qu'apprentissage ; Earl, 2003). Dans ce dernier cas, l'accent est mis sur des pratiques d'autoévaluation et l'utilisation des capacités métacognitives de l'élève pour permettre une meilleure prise de conscience de l'activité d'apprentissage et assurer le suivi nécessaire pour apporter les correctifs. Elle est souvent associée et confondue avec l'évaluation-soutien d'apprentissage, car toutes deux visent à accroître l'autonomie des apprenants et leur capacité à s'autoévaluer de manière constructive.

Le tableau 1.1 décrit les trois fonctions précédentes et ce qui les différencie en rapport avec les quatre étapes essentielles de leur déploiement : l'intention, la prise d'information, l'évaluation-jugement et la décision. Il est à noter que les cellules de ce tableau ne sont pas parfaitement étanches. Par exemple, un enseignant peut choisir de faire de « micro-bilans » pour vérifier la progression des élèves et s'assurer qu'ils ne prennent aucun retard. Un exemple de micro-bilan est présenté à la figure 1.8 (section 2.4). Ces micro-bilans peuvent, à la limite, servir à réguler les méthodes d'enseignement si les résultats des élèves sont pris en considération dans la planification des leçons et s'ils permettent de fournir une rétroaction sur les apprentissages. Par contre, s'ils ne servent qu'à faire partie du calcul d'une note finale, il s'agit alors d'une évaluation essentiellement sommative. Ce qui différencie l'évaluation-soutien d'apprentissage de l'évaluation en tant qu'apprentissage dépend en grande partie du degré d'implication de l'apprenant. Un feedback correctif de la part de l'enseignant ne contribue pas nécessairement à développer l'autonomie des apprentissages même s'il peut aider l'élève à progresser. Il est difficile dans un tel cas de parler d'évaluation en tant qu'apprentissage. Par contre, lorsque le feedback de l'enseignant stimule l'élève à réviser sa production et à la modifier en lui fournissant des clés d'amélioration, nous sommes davantage dans une perspective où l'évaluation-soutien d'apprentissage fournit à l'élève l'occasion de s'autoévaluer et de rectifier son travail par lui-même.

Une autre caractéristique qui ressort du tableau 1.1, tient à ce que ni l'évaluation-soutien d'apprentissage ni l'évaluation en tant qu'apprentissage ne font appel à des instruments formels de mesure dans la grande majorité des situations. C'est pourquoi, dans les sections suivantes, nous aborderons essentiellement l'évaluation sommative des apprentissages, car l'évaluation-soutien d'apprentissage relève autant, sinon plus de considérations pédagogiques et didactiques, que de la conception d'instruments de mesure. De nombreux ouvrages et articles sont consacrés à l'évaluation-soutien d'apprentissage, à l'évaluation formative ainsi qu'aux défis posés par son implantation réussie en éducation (Laveault & Allal, 2016 ; Scallon, 1988, 2000 ; Stiggins *et al.*, 2004).

Tableau 1.1. Fonctions de l'évaluation et qualités attendues des instruments de mesure

	Évaluation des apprentissages Évaluation sommative	**Évaluation soutien d'apprentissage Évaluation formative**	**Évaluation en tant qu'apprentissage**
	Assessment of learning	***Assessment for learning***	***Assessment as learning***
Intention	Établir un bilan, accorder une note, certifier, attester d'un niveau de maîtrise ou de rendement.	Identifier les difficultés, repérer les erreurs systématiques, fournir un feedback descriptif et détaillé.	Prise de conscience et suivi par l'élève de ses activités d'apprentissage.
Prise d'information	Calcul de scores à des tests, examens ou à des grilles d'évaluation, données principalement quantitatives.	Données qualitatives, profils individuels, observations, dossier d'apprentissage, dossier anecdotique, dialogues entre élèves ainsi qu'entre enseignant et élèves.	Auto-observation, évaluation avec les paires, coévaluation avec l'enseignant.
Jugement	Jugement global par l'enseignant. Normatif (par rapport au groupe de référence) ou critérié (par rapport aux objectifs d'un programme de formation ou à un seuil de performance à atteindre).	Interprétation en rapport avec le but poursuivi, les exigences de la tâche et les cibles d'apprentissage.	Utilisation appropriée des rétroactions reçues de l'enseignant et/ou d'autres élèves. Recherche autonome de solutions.
Décision	Décision administrative par l'enseignant : réussite ou échec, progression au niveau suivant, recommandation ou non au diplôme.	Décision pédagogique par l'enseignant avec implication plus ou moins grande de l'apprenant quant aux régulations de l'enseignement et de l'apprentissage à mettre en place pour un impact positif.	Décision individuelle par l'élève quant au choix de nouvelles stratégies d'apprentissage ou aux modifications à leur apporter. Au besoin, ajuste les cibles à atteindre et recherche l'aide nécessaire pour progresser.

2.3. L'ÉVALUATION SOMMATIVE

Pour réaliser l'évaluation sommative des apprentissages des élèves, il faut que celle-ci reflète les objectifs du programme d'études et de l'enseignement en salle de classe. C'est ce qui a été décrit dans la section 2.1 sur l'alignement. Mais comment parvient-on à réaliser cet alignement ? Plusieurs modèles de spécification de domaine ont été développés à cet effet, notamment l'évaluation fondée sur les objectifs. Dans ce contexte, l'objectif est conçu comme « une communication d'intention décrivant l'apprentissage qui est attendu de celui à qui il s'adresse » (Morissette, 1984). L'objectif d'apprentissage est donc un maillon important non seulement de l'alignement, mais aussi de la transmission des attentes à tous les niveaux d'un système d'éducation.

2.3.1 *La mesure fondée sur les objectifs*

Les programmes d'études comportent généralement plusieurs catégories d'objectifs. Ceux-ci peuvent être regroupés selon leur degré de spécificité (objectif global, général, spécifique) ou selon leur position dans une séquence d'apprentissage (objectif intermédiaire ou terminal). Quelle que soit la catégorie à laquelle il appartient, l'objectif possède des caractéristiques essentielles et des caractéristiques accessoires (tableau 1.2).

Lors de la rédaction d'un objectif, les deux caractéristiques essentielles sont :

- un verbe d'action et un seul ;
- un contenu (complément d'objet) et un seul.

Tableau 1.2. Formulation des objectifs d'apprentissage

	Obligatoire	Optionnel	
Verbe	• un seul verbe • un verbe d'action • doit décrire un comportement univoque	Contexte	• ce qui est ou n'est pas disponible
Contenu	• un seul contenu par objectif • doit être un élément ou un sous-élément d'un programme	Critères d'évaluation	• condition d'acceptation de la performance • seuil de performance

Le verbe d'action doit décrire un comportement observable directement (par exemple, cocher, souligner, écrire, lancer, etc.) ou indirectement (par exemple, identifier, choisir, etc.). Il ne doit y avoir qu'un seul verbe par objectif, sinon les attentes exprimées peuvent donner lieu à confusion. Prenons l'exemple de l'objectif suivant : « Identifier et nommer les capitales provinciales du Canada ». La présence de deux verbes obscurcit les attentes en ce qui concerne les apprentissages des élèves. Sera-t-on satisfait lorsque l'élève saura nommer les capitales du Canada ou encore lorsqu'il pourra les identifier à partir d'une liste ou d'une carte géographique ? Pour considérer cet objectif comme atteint, faudra-t-il que l'élève manifeste les deux comportements (identifier et nommer) ou un seul des deux (identifier ou nommer) ? L'objectif manque de précision, non seulement à cause de l'ambiguïté créée par la présence de deux verbes, mais aussi parce que nous ignorons les conditions de réalisation de la

performance et le seuil de réussite permettant de déterminer quand l'objectif peut être considéré comme atteint. Pour accroître la spécificité des objectifs, on ajoutera généralement les composantes suivantes :

- le contexte dans lequel sera réalisée la performance attendue ;
- le critère d'acceptation de la performance ;
- le seuil d'acceptation de la performance.

On ne s'attend pas à retrouver ces caractéristiques accessoires parmi les objectifs généraux. Par contre, elles sont essentielles aux objectifs dits spécifiques. La figure 1.1 fournit un exemple d'un objectif spécifique comportant toutes ces composantes accessoires.

Le contexte décrit dans quelles conditions l'élève réalisera sa performance et ce qui sera à sa disposition. Dans le cas de l'exemple de la figure 1.1, il s'agit d'un atlas. Dans le cas d'autres objectifs, il pourrait s'agir d'une calculatrice, d'un dictionnaire ou d'une grammaire. Le critère d'acceptation de la performance décrit le niveau de qualité de la performance attendue. Dans notre exemple, les coordonnées devront être relevées avec une précision d'un degré. Une erreur supérieure à un degré invaliderait la réponse en entier. Enfin, le seuil de réussite fournit un critère quantitatif pour considérer l'objectif comme atteint. Il établit combien de fois l'élève doit répéter sa performance au critère d'acceptation fixé pour que sa maîtrise du contenu de l'objectif soit attestée. Les seuils les plus courants oscillent généralement entre 80 % et 100 %. Dans le cas de l'objectif de la figure 1.1, ce seuil est de 90 %. Qu'est-ce qui constituerait un seuil de réussite acceptable pour l'objectif « identifier les capitales provinciales du Canada ? », 80 % ? 90 % ? Cela pourrait dépendre des élèves à qui s'adresse cet objectif et du programme d'études préconisé : le seuil pourrait être moindre pour des élèves belges que pour des élèves canadiens, par exemple.

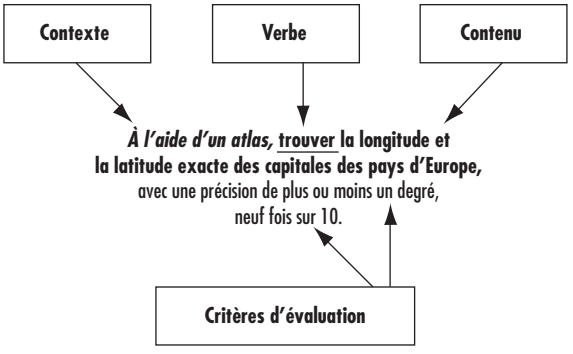


Figure 1.1 — Exemple de formulation d'un objectif

Il existe plusieurs façons de déterminer un seuil de réussite. Cette question sera abordée plus en détail dans le chapitre 6, section 4. Pour l'instant, mentionnons que les composantes accessoires des objectifs sont parfois précisées dans les programmes d'études en fonction des niveaux d'enseignement. Si elles ne sont pas

précisées, l'enseignant peut utiliser son jugement professionnel pour les déduire à partir d'informations complémentaires (par exemple, des textes des manuels obligatoires ou d'autres directives provenant des autorités scolaires). Le contexte, le seuil et les conditions d'acceptation de la performance sont également des moyens de graduer les attentes en termes d'exigences et d'établir une progression et une continuité dans les apprentissages.

2.3.2 *Le modèle de Deno et Jenkins et les taxonomies d'objectifs*

Les objectifs spécifiques nous permettent de préciser les tâches qui peuvent servir à l'évaluation des apprentissages. Toutefois, ils sont peu pratiques pour considérer un programme d'études dans son ensemble ou pour décrire les capacités cognitives auxquelles ces tâches font appel. Lorsqu'il s'agit de planification à long terme de l'enseignement et d'intégration des matières, les objectifs spécifiques peuvent devenir très nombreux et encombrants. L'attention doit alors se porter sur l'organisation de grands pans de la matière et sur le niveau global d'approfondissement des apprentissages visés.

Deno et Jenkins (1969) ont élaboré un modèle qui tient compte de la spécificité nécessaire des objectifs à différents niveaux d'intervention. Il s'agit d'un modèle hiérarchique où chaque niveau supérieur englobe les objectifs des niveaux inférieurs :

- Le niveau A est celui des « objectifs globaux ». Il sert à préciser les choix politiques, institutionnels, les grandes lignes du projet éducatif et de la mission de l'enseignement. Ce niveau est peu élaboré et décrit des politiques gouvernementales et les missions du système d'éducation.
- Le niveau B cherche à préciser les objectifs globaux en situant le degré d'approfondissement des capacités (au niveau cognitif) ou le degré d'intériorisation (au niveau affectif) des attentes de l'objectif : c'est le niveau des « objectifs généraux ». Ils sont une première indication du degré d'approfondissement visé. Ils sont particulièrement utiles pour dresser les grandes lignes d'un programme d'études et articuler entre eux des objectifs qui peuvent, par les processus en jeu et leur contenu, être fort différents. Ils sont habituellement conçus et élaborés par les concepteurs de programme et les spécialistes de l'apprentissage scolaire comme les didacticiens des matières.
- Au niveau C, les intentions se précisent à un point tel qu'on peut y indiquer les conditions précises d'évaluation : comportements attendus, contenus précis, conditions de réalisation de la performance et conditions d'acceptation de la performance. C'est le niveau des « objectifs spécifiques ». Ce niveau intéresse les personnes chargées de l'enseignement et de l'évaluation, telles que les enseignants eux-mêmes et les conseillers pédagogiques.
- Au niveau D, on retrouve les « tâches d'examen » de même que les situations entraînant l'observation de performances complexes. C'est le niveau le plus spécifique des quatre niveaux du modèle. Ce n'est pas à proprement parler un objectif, mais, comme le mentionne Ebel (1956), la tâche est la meilleure manière de connaître comment se traduisent les objectifs pédagogiques dans les faits. À ce niveau, un exemple concret d'item est généré, soit pour illustrer l'objectif spécifique, soit pour faire partie de l'examen.

La figure 1.2 fournit un exemple d'objectifs pour chacun des quatre niveaux (A à D) du modèle de Deno et Jenkins, allant de l'objectif global à la tâche d'examen. L'alignement de l'évaluation sur le programme d'études de géographie se lit de haut en bas (validité déductive) et de bas en haut (validité inductive).

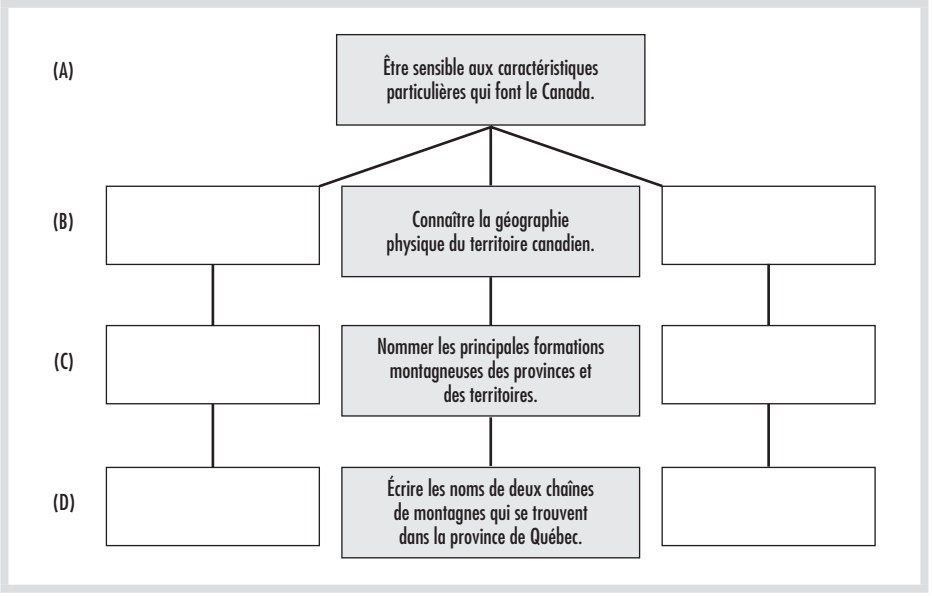


Figure 1.2 — Le modèle de Deno et Jenkins

C'est au niveau des objectifs généraux (niveau B) qu'on retrouve les taxonomies d'objectifs qui décrivent les processus cognitifs ciblés pour l'approfondissement de la matière. Même s'il s'agit d'objectifs généraux, il est important de communiquer, sans ambiguïté, les processus en jeu. Bloom, Engelhart, Furst, Hill, et Krathwohl (1956) avaient remarqué que les examens de rendement scolaire portaient principalement sur la restitution de connaissances et pas suffisamment sur les aptitudes de niveau supérieur. De plus, ils avaient observé de nombreuses divergences dans la compréhension et la définition des connaissances et des capacités cognitives de haut niveau. C'est dans cette perspective qu'a été développée la première taxonomie des objectifs cognitifs par Bloom *et al.* (1956). Elle fait la distinction entre six niveaux d'habileté et d'acquisition de connaissances. Ces six niveaux hiérarchiques sont décrits à la figure 1.3 (connaissances) et à la figure 1.4 (habiletés) et chacun se développe en une ou plusieurs sous-catégories.

La taxonomie des objectifs cognitifs joue un double rôle : (1) au niveau des programmes d'études et (2) au niveau de l'évaluation des apprentissages :

- Au niveau des programmes d'études, la taxonomie apporte plus de rigueur dans la définition de ce que l'on entend généralement par « connaissance », « compréhension », etc. De plus, elle permet de s'assurer que les attentes vis-à-vis des apprentissages des élèves sont conformes à leurs capacités et à leur développement cognitif. On peut ainsi établir une progression des habiletés

intellectuelles impliquées dans l'apprentissage de mêmes contenus, mais à des niveaux scolaires différents. Par exemple, « établir une classification du contenu de son herbier à partir d'un modèle fourni par l'enseignant » constitue un objectif cognitif différent de celui qui consiste à « élaborer une classification originale du contenu de son herbier à partir des échantillons de plantes recueillies ». Le premier objectif porte sur l'application du modèle de l'enseignant (figure 1.4, catégorie 3.00), alors que le second repose davantage sur la synthèse (élaboration d'un plan d'action, figure 1.4, catégorie 5.20).

- Au niveau de l'évaluation des apprentissages, la taxonomie permet de s'assurer que les processus cognitifs activés lors de l'apprentissage seront mesurés lors de l'examen. En plus de faire en sorte que les examens reflètent les niveaux cognitifs attendus, les objectifs généraux précisent l'interprétation à donner aux objectifs spécifiques. Prenons une situation concrète assez répandue. Supposons que nous demandions à un étudiant de « fournir un exemple de renforcement positif ». S'il s'agit d'un objectif de connaissance, il suffira à l'étudiant de répéter un exemple qu'il a entendu en classe ou lu dans le manuel obligatoire du cours. Si, par contre, il s'agit d'un objectif de compréhension, nous nous attendons à ce qu'il fournisse un exemple original. La réitération d'un exemple connu ne serait pas suffisante pour parler de compréhension. De ce dernier exemple, nous pouvons conclure qu'une même tâche peut être employée pour mesurer des niveaux taxonomiques fort différents. La condition d'acceptation de la performance permet de s'assurer que la question d'examen mesure bien le niveau taxonomique visé. Pour que les choses soient claires pour l'étudiant, il faudra que l'énoncé de la question soit sans équivoque à propos de cette condition d'acceptation. Par exemple : « Écrivez un exemple original de renforcement positif. Les exemples du manuel de cours ou du professeur ne seront pas acceptés ».

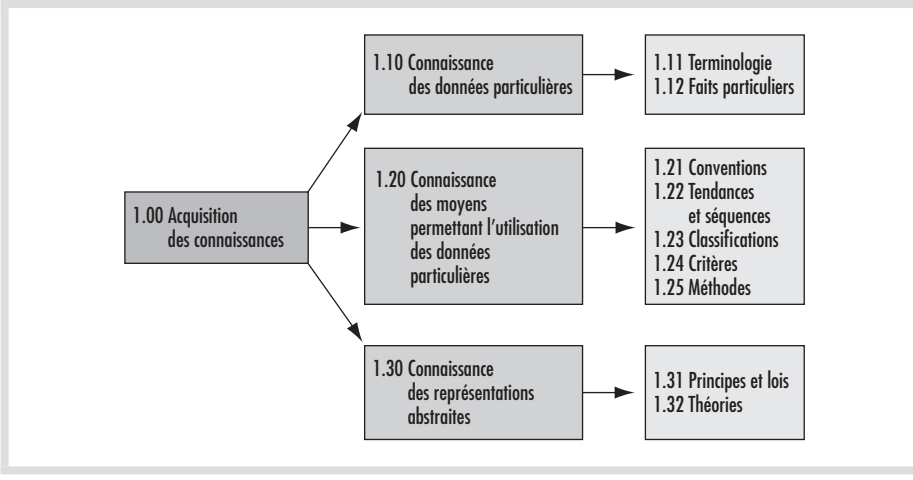


Figure 1.3 — Taxonomie des objectifs cognitifs : les connaissances

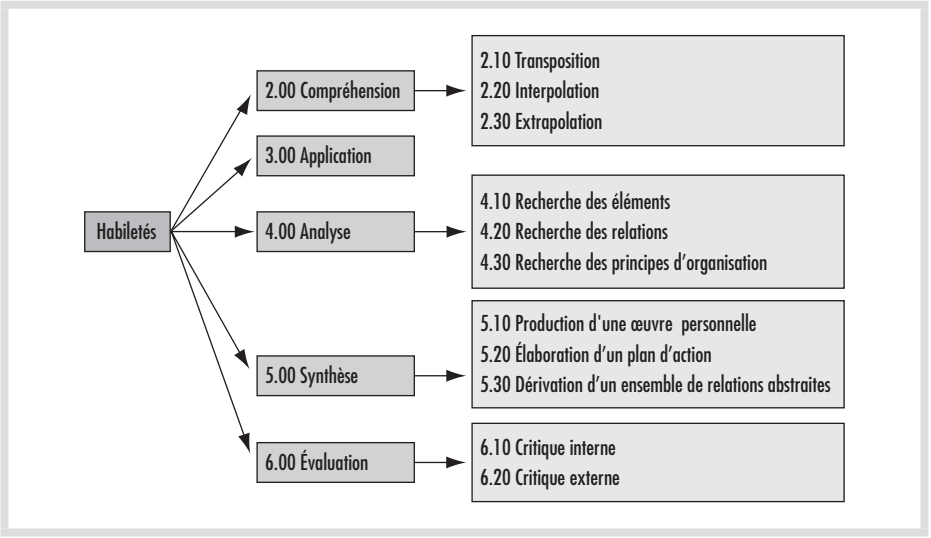


Figure 1.4 — Taxonomie des objectifs cognitifs : les habiletés

Afin de tenir compte des progrès tant dans le domaine de l'évaluation que dans celui des théories cognitives de l'apprentissage, une mise à jour de la taxonomie de Bloom a été réalisée par Anderson et Krathwohl (2001). Plusieurs autres taxonomies ont également été développées[1]. La figure 1.5 présente, côte à côte, la taxonomie originale de Bloom, la taxonomie révisée d'Anderson et Krathwohl (2001) et celle de Stiggins *et al.* (2004).

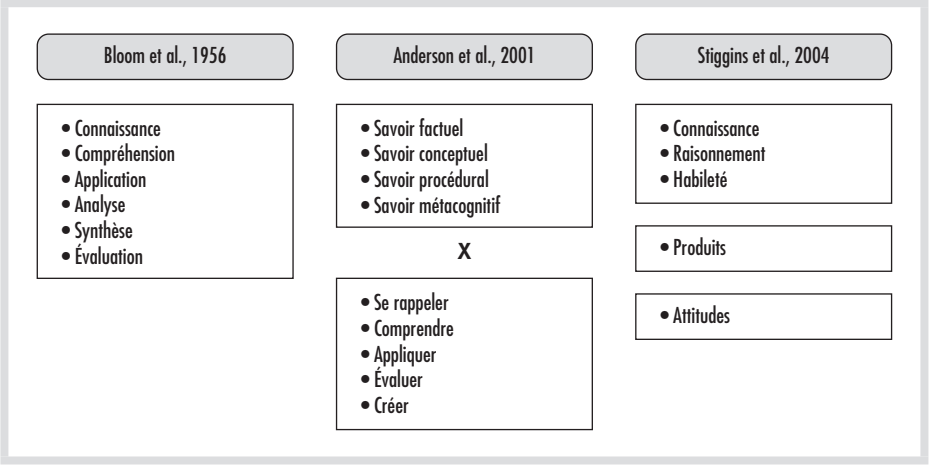


Figure 1.5 — Trois exemples de taxonomies d'objectifs cognitifs

1 Le dictionnaire de Legendre (2005) dénombre pas moins de 31 taxonomies dans les domaines affectif, cognitif, conceptuel, expérientiel, moral, perceptuel, psychomoteur et social.

Chacune des taxonomies comporte ses particularités, ses propres avantages et inconvénients. La taxonomie de Bloom est largement utilisée parce qu'elle est la plus simple. Elle est unidimensionnelle et est familière aux enseignants depuis longtemps. La révision apportée par Anderson et Krathwhol (2001) introduit plusieurs éléments nouveaux qui tiennent compte des développements récents dans le domaine des théories de l'apprentissage, de l'enseignement et des modèles récents d'évaluation scolaire. Ils la rendent de ce fait même d'autant plus complexe que, comme l'indique son titre *A taxonomy for learning, teaching and assesssment*, son champ d'action va bien au-delà de l'évaluation. C'est donc un outil qui répond d'abord aux besoins des concepteurs de programmes de formation, des spécialistes du curriculum, des didactiques des matières et enfin, aux professionnels de l'évaluation des apprentissages. Tel que l'indique la figure 1.5, elle croise deux dimensions : les catégories de savoir et les catégories des processus cognitifs. De plus, chacune de ces catégories se subdivise en sous-catégories. Par exemple, la catégorie « savoir conceptuel » se subdivise en trois sous-catégories : « connaissance des classifications », « connaissance des principes et des généralisations » et « connaissance des théories, modèles et structures ». Il en va de même pour les processus cognitifs. Par exemple, la catégorie « se rappeler » se subdivise en deux sous-catégories : « reconnaître » comme dans une question à choix multiples où il suffit de choisir la bonne réponse parmi plusieurs et « récupérer » comme dans une question à réponse courte où il faut retrouver l'information en mémoire.

En comparaison avec la taxonomie révisée, celle de Stiggins *et al.* (2004) est plus simple. Elle a été conçue spécifiquement pour l'enseignant dans le cadre tant d'une évaluation formative que sommative. C'est ainsi que « les cibles d'apprentissage », pour utiliser l'expression utilisée dans cette taxonomie, incluent les nouvelles catégories suivantes :

- **le niveau d'habileté et d'adresse** (*skill*) avec laquelle une performance complexe est exécutée, tant au niveau cognitif que psychomoteur ;
- **l'attitude** (*disposition*) vis-à-vis de l'expérience d'apprentissage, comme le plaisir à converser en espagnol ;
- **l'évaluation des productions**, car les examens, les questionnaires et les épreuves écrites ne sont pas les seuls moyens pour évaluer si une cible d'apprentissage a été atteinte. Par exemple, pour évaluer une production en arts visuels, une prestation musicale, un poème, il faut fréquemment faire appel à des grilles d'observation ou à des listes de vérification qui serviront non seulement à noter celles-ci, mais également à fournir une rétroaction à l'apprenant.

Ces trois catégories sont particulièrement pertinentes dans le contexte de l'évaluation formative. L'évaluation des productions est parfois le seul moyen mis à disposition pour évaluer les performances attendues et la mobilisation de compétences complexes. Les autres catégories du modèle de Stiggins rejoignent les types de savoirs et de processus cognitifs présents dans les taxonomies précédentes.

2.3.3 *Objectifs terminaux et objectifs intermédiaires*

Il est parfois nécessaire d'aborder l'articulation des objectifs en termes de séquence d'apprentissage. La taxonomie des objectifs permet de décrire une hiérarchisation des processus cognitifs en jeu dans les objectifs, mais ne fournit aucun renseignement

sur la progression des objectifs d'apprentissage dans le temps. Lorsqu'on souhaite préciser l'enchaînement de plusieurs objectifs dans un programme d'études, on peut distinguer les objectifs terminaux des objectifs intermédiaires[2]. L'objectif terminal décrit la finalité ultime d'un apprentissage, son point d'arrivée. L'objectif intermédiaire énumère les étapes nécessaires qui jalonnent le cheminement de l'élève. Sans la maîtrise de ces jalons, l'atteinte de l'objectif terminal est compromise. Par contre, lorsque ce dernier est atteint, on peut conclure que les objectifs intermédiaires ont bien été maîtrisés.

Les objectifs terminaux conviennent particulièrement à l'évaluation sommative. Ils permettent de couvrir les objectifs pour lesquels la formation est censée être achevée et pour lesquels un bilan est approprié. Il est normal qu'un bilan porte sur les apprentissages complétés plutôt que sur ceux qui sont en voie de réalisation. Enfin, lorsqu'il s'agit d'établir un bilan, il est généralement trop tard pour se demander à quel moment de l'apprentissage l'étudiant a éprouvé des difficultés. Par contre, cette dernière information peut être utile dans le cas d'une évaluation formative ou encore de ce qu'il est convenu d'appeler une évaluation « micro-sommative » (Scallon, 1992). Afin de mieux comprendre les raisons d'une difficulté au niveau d'un objectif terminal, il peut alors être utile de s'assurer que tous les objectifs intermédiaires prérequis sont bien maîtrisés. Le degré de maîtrise de chaque objectif intermédiaire peut renseigner sur les moyens de corriger une difficulté liée à l'absence de maîtrise d'un objectif terminal.

Suite aux progrès récents des théories cognitives et des modèles d'enseignement des langues et des mathématiques, un grand nombre de juridictions scolaires des principaux pays industrialisés se sont dotées de normes d'apprentissage (en anglais : *standards*)[3] décrivant les séquences attendues et les résultats d'apprentissage aux étapes déterminantes du parcours scolaire. Le *Dictionnary of Educational Reform* en donne la définition suivante, du point de vue américain :

> Les normes d'apprentissage sont des descriptions concises de ce que les élèves sont censés savoir et être capables de faire à un stade précis de leur éducation. Les normes d'apprentissage décrivent les objectifs pédagogiques – c'est-à-dire ce que les élèves devraient avoir appris à la fin d'un cours, d'un niveau scolaire ou d'une série d'années scolaires – mais elles ne décrivent aucune pratique d'enseignement, programme ou méthode d'évaluation en particulier.[4]

Les standards de progression des apprentissages sont souvent fondés en tout ou en partie sur des modèles théoriques de l'apprentissage et de la didactique des matières. Cependant, il n'existe pas de standards de progression dans toutes les disciplines. Ils se

2 La terminologie utilisée peut varier. Par exemple, les programmes de l'Ontario utilisent les termes « attentes » et « contenus ». Les attentes sont clairement des objectifs terminaux et les contenus des objectifs intermédiaires ou des prérequis nécessaires pour réaliser les attentes. Le programme ne va pas plus loin et ne suggère pas d'ordre précis dans l'acquisition des contenus.

3 Les traductions de « standard » en français sont variées. Par exemple, au Québec, on utilisera l'expression « échelle de niveaux de compétence ». Ces échelles sont accessibles à l'adresse suivante : http://www.education.gouv.qc.ca/fileadmin/site_web/documents/dpse/evaluation/echellesduprimaire.pdf

4 Traduit de l'entrée *Learning standards* de la référence : *The Glossary of Educational Reform* accessible sur internet à l'adresse suivante : https://www.edglossary.org/learning-standards/

retrouvent principalement dans les disciplines de l'apprentissage de la langue (maternelle ou seconde) et des mathématiques. Hendrickson *et al.* (2010) soulignent les différences suivantes dans l'analyse de domaine des disciplines suivantes :

1. **Langues.** Une caractéristique importante du domaine des langues est que les compétences ne sont pas indépendantes les unes des autres. Par conséquent, tout item/tâche donné peut susciter des preuves d'apprentissage à l'appui d'une ou plusieurs affirmations.
2. **Sciences**. Les pratiques scientifiques ne sont pas aussi interdépendantes que les compétences linguistiques. En tant que tels, les items peuvent être rédigés pour obtenir des preuves d'apprentissage spécifiques et l'interdépendance entre ces preuves peut être limitée dans la prise en compte des spécifications du test.
3. **Histoire.** Les contenus et les compétences en histoire peuvent être considérés séparément les uns des autres, mais les affirmations sur les capacités de réflexion historique, fondées sur les compétences, ne peuvent pas être considérées comme complètement indépendantes.

Ces normes exercent une fonction similaire aux objectifs intermédiaires. Tout comme les programmes de formation peuvent varier d'un pays à l'autre ou même à l'intérieur d'un pays, les normes d'apprentissage peuvent également varier. Même si elles ne sont associées à aucune pratique d'enseignement ou méthode d'évaluation, elles peuvent être prises en considération dans la spécification de domaine des épreuves à grande échelle administrées par une juridiction scolaire.

2.3.4 *Échantillons d'items et objectifs d'apprentissage*

Certains instruments de mesure, en particulier les examens, doivent être administrés à période fixe afin de dresser un bilan des apprentissages de l'élève. Cette évaluation répond à une exigence administrative. Ceci ne signifie pas que l'enseignant ne soit pas intéressé de temps à autre à effectuer un bilan des apprentissages de ses élèves pour son propre compte. Mais ce bilan se ferait probablement de façon fort différente. Par exemple, l'enseignant pourrait décider d'éliminer de certains bilans les items qu'il considère comme réussis depuis longtemps par une grande majorité des élèves. Pour certifier un cycle d'apprentissage, cependant, la couverture de la matière devra être exhaustive, même si elle porte sur des points pour lesquels l'enseignant est déjà assez bien informé quant à la réussite des élèves.

Le bilan se doit d'être représentatif. Le type de représentativité peut différer selon l'usage qui sera fait du bilan en question. Lorsqu'il s'agit de certification, cette définition doit être stricte. L'enseignant dispose de moins de marge quant à l'univers des situations qu'il peut échantillonner pour son examen. Afin d'assurer la comparabilité des résultats entre classes, les enseignants de cinquième primaire, par exemple, devront tirer leurs questions d'examen d'un même ensemble spécifié par le programme d'études. Pour chaque enseignant, ce ne seront pas les mêmes questions, mais elles devraient, dans la mesure du possible, constituer des ensembles parallèles comparables et congruents avec le programme commun à tous les élèves.

Un bilan peut reposer sur plusieurs objectifs et, pour chacun de ces objectifs, il existe un très grand nombre de possibilités de questions ou de tâches. L'échantillonnage est un des outils à la disposition de l'enseignant pour construire son instrument de

mesure. Tout comme l'échantillonnage des élèves (voir chapitre 6, section 2.2.2), l'échantillonnage des items peut prendre plusieurs formes :

1. **Échantillonnage aléatoire simple**. Chaque item peut être considéré comme tiré d'un univers presque illimité de possibilités, même si, dans la pratique, ils ne sont pas vraiment tirés de la population en entier. Dans la pratique, la population des objectifs d'intérêt peut faire l'objet de critères d'exclusion parce que l'atteinte de ces objectifs ne peut être vérifiée au moyen d'un examen papier-crayon ou pour des contraintes de temps ou d'horaire. Ceci étant admis, si plusieurs enseignants devaient produire un échantillon de questions à partir des mêmes objectifs, il est possible de considérer que les fluctuations entre les différents échantillons devraient être l'effet du hasard.
2. **Échantillonnage stratifié**. Dans ce type d'échantillonnage, on cherche à restreindre les fluctuations du hasard en le balisant. Cette méthode est particulièrement utilisée lorsque le nombre d'items à choisir au départ est relativement faible et qu'il faut s'assurer qu'ils se retrouveront dans des proportions similaires au domaine d'où ils ont été tirés. Par exemple, on peut faire en sorte que la distribution des items choisis respecte une certaine répartition proportionnelle des processus cognitifs évalués. Des enseignants générant des items d'examen, à partir de ces balises, devraient en arriver non pas aux mêmes questions, mais à une répartition équivalente des questions selon la nature des processus cognitifs. Une autre strate pourrait consister à faire en sorte que le nombre de questions de l'examen reflète une proportion équivalente à l'importance de cet objectif dans le programme.
3. **Échantillonnage par grappes.** Parfois, le nombre d'objectifs à évaluer est si élevé qu'il faut en faire l'échantillonnage comme pour les examens de fin d'année. Les objectifs sont considérés comme des grappes d'items et les items ne peuvent provenir d'autres objectifs que ceux qui ont été retenus.
4. **Échantillonnage hiérarchique.** L'échantillonnage se fait en deux temps : (1) d'abord les objectifs et ensuite (2) les questions à l'intérieur des objectifs. Cette approche est nécessaire lorsqu'un très grand nombre d'items peut être généré par un seul objectif. Ainsi, l'on s'assure que, partant des mêmes objectifs, la sélection des items produira des examens équivalents dont les fluctuations proviendront de l'échantillonnage d'items nichés dans chacun des objectifs.

Ces méthodes d'échantillonnage sont décrites au moyen des quatre schémas de la figure 1.6. Elles permettent une certaine forme d'encadrement dans l'assemblage des items et des tâches devant faire partie d'une épreuve destinée à faire le bilan des apprentissages. Ainsi, l'échantillonnage des items, à partir des objectifs, favorise l'alignement de l'épreuve avec l'enseignement et le programme de formation, en plus de réduire les variations possibles qui pourraient exister entre enseignants évaluant les mêmes objectifs.

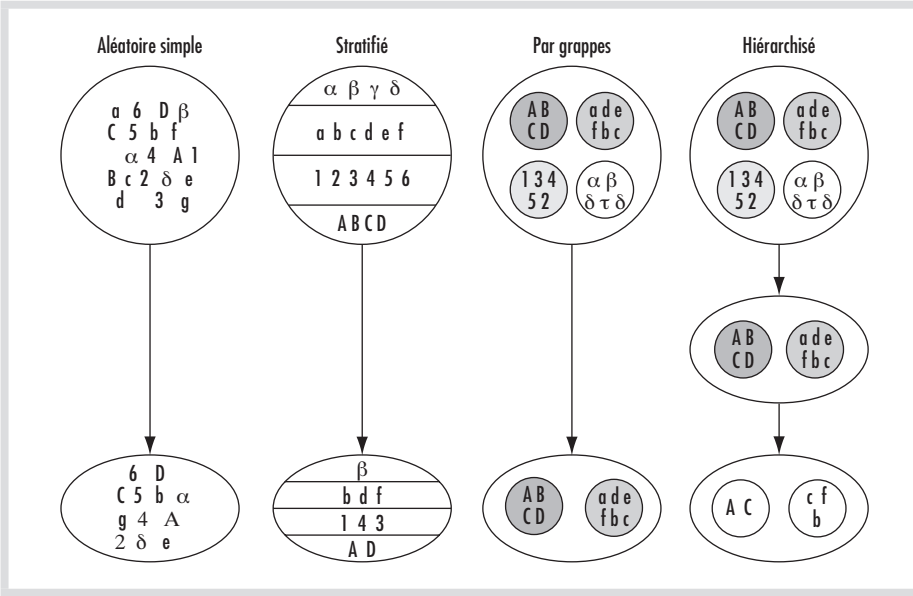


Figure 1.6 — Techniques d'échantillonnage des questions

L'outil le plus courant pour échantillonner les items est le tableau de spécification. Il est utilisé depuis longtemps pour s'assurer que l'échantillonnage des items est véritablement aligné sur le programme de formation et est représentatif de la situation qui a prévalu en salle de classe. Selon Owen (2018), le tableau réunit deux types de spécifications :

- **la spécification basée sur le contenu**, qui regroupe les items en fonction des informations sur le contenu de la matière pondéré en fonction de son importance ;
- **La spécification fondée sur les processus**, qui réunit les items sollicitant les mêmes processus cognitifs tels que les connaissances et les capacités de haut niveau selon la taxonomie choisie et le poids de ces processus dans le programme et le curriculum.

Le tableau de spécification prend souvent la forme d'un tableau de contingence à double entrée, l'une étant constituée du contenu, l'autre du niveau taxonomique des objectifs mesurés. Il permet d'assurer que la proportion des items d'examen corresponde étroitement à l'importance relative du contenu et du niveau taxonomique des objectifs du programme d'études. Le tableau 1.3 présente un exemple de tableau de spécification pour un examen de géographie.

Tableau 1.3. Exemple de tableau de spécification : nombre et pourcentage d'items dans chaque catégorie

Contenu \ Niveau taxonomique	Connaissance	Compréhension	Total
Géographie humaine	10 — 20 %	5 — 10 %	15 — 30 %
Géographie politique	10 — 20 %	5 — 10 %	15 — 30 %
Géographie physique	10 — 20 %	10 — 20 %	20 — 40 %
Total	30 — 60 %	20 — 40 %	50 — 100 %

Le tableau de spécification correspond à un échantillonnage stratifié. En ce qui concerne l'examen de géographie repris au tableau 1.3, la stratification s'est effectuée en tenant compte du contenu (géographie humaine, politique ou physique) ainsi que du niveau taxonomique (connaissance, compréhension). En principe, la répartition des items d'examen, selon ces deux caractéristiques, est censée refléter l'importance consacrée en classe en termes de temps d'étude ou d'enseignement. Si 10 % du temps en classe a été consacré à la compréhension de la géographie politique, 10 % des 50 questions d'examen (5 questions) devraient porter sur cette matière. À défaut de trouver autant de questions, il est toujours possible d'ajuster la pondération de l'examen de manière à rendre le score total plus représentatif de l'importance de ces questions dans le curriculum de l'élève. Plutôt que cinq questions à un point chacune, ce pourrait être une question de deux points et une autre de trois points sur la géographie politique.

D'autres caractéristiques que le niveau taxonomique ou le contenu peuvent être employées pour établir un tableau de spécification. Le type de production (convergente ou divergente), le format d'items (choix de réponses ou réponse élaborée) peuvent également être pris en considération. Néanmoins, l'exemple précédent est sans doute plus représentatif de ce qui se passe en contexte scolaire. En effet, l'organisation habituelle des programmes d'études favorise ce genre de stratification contenu/processus.

Même s'il permet d'introduire plus de rigueur dans l'assemblage des items d'examen, le modèle traditionnel de tableau de spécification connaît des limites qui lui valent plusieurs critiques :

- il est difficile de situer dans un tableau de spécification des items impliquant simultanément plus d'un seul contenu ou plus d'un processus cognitif ;
- il est peu adapté à l'évaluation de performances ou de productions complexes ;
- il ne fournit pas suffisamment de détails pour générer des questions de test bien précises. Pour ce faire, il doit être accompagné de modèles de spécification d'items (voir section 2.5).

2.4. Évaluation critériée

Selon Popham (1992), la spécification du domaine a relativement peu d'importance dans le contexte d'une évaluation normative. Une description globale suffit généralement à différencier les résultats des candidats (par exemple, un concours d'admission à un programme enrichi de formation). L'apparition des examens à enjeux

élevés a conduit le monde de l'éducation à exiger plus de détails et de transparence dans la préparation des épreuves, d'où l'importance accrue de l'évaluation à référence critériée. La caractéristique distinctive de la mesure critériée est la clarté avec laquelle est décrit le domaine mesuré. Cette description prend généralement la forme de spécifications de test détaillées (à la fois « spécification d'item » pour régir la rédaction de tâches d'évaluation et « spécification de domaine » pour régir la composition globale du test).

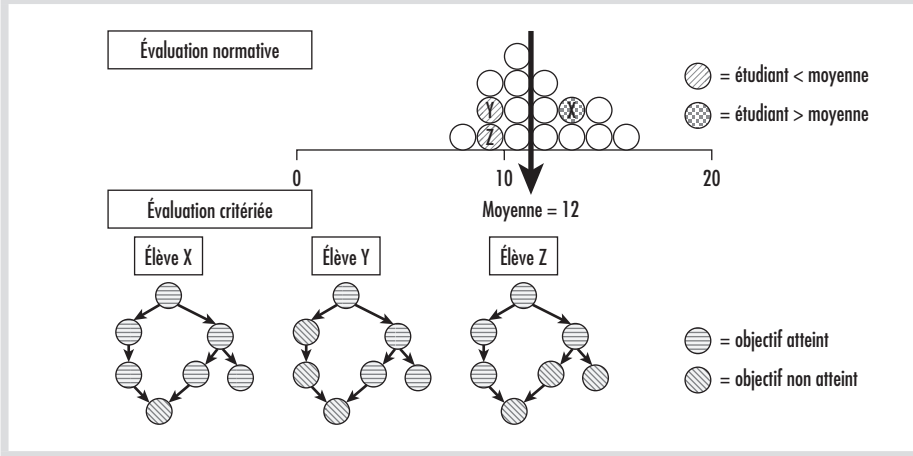


Figure 1.7 — Schémas illustrant l'évaluation normative et l'évaluation critériée

La figure 1.7 illustre le commentaire précédent de Popham dans le contexte scolaire. L'objet de l'évaluation normative est le score individuel (représenté par un cercle). L'évaluation normative différencie ces scores entre eux, notamment par rapport à leur position à la moyenne d'un groupe. Par contre, l'évaluation critériée a pour but de différencier non pas les individus, mais les apprentissages. Dans le schéma de l'évaluation critériée, les cercles représentent l'atteinte ou non d'une cible d'apprentissage à l'intérieur d'une progression prévue dans le cadre d'un enseignement aligné sur un programme d'études. L'élève X est celui qui est parvenu le plus loin dans cette progression. Seul l'objectif terminal n'a pas été atteint. Par contre, les élèves Y et Z, même s'ils ont atteint le même nombre d'objectifs, ne présentent pas le même profil. Pour l'évaluation normative, les élèves Y et Z ne se différencient pas, car leurs scores totaux sont identiques. Il en va tout autrement de l'évaluation critériée qui fait ressortir les différences dans le parcours d'apprentissage de ces deux élèves : chacun est parvenu à une étape différente de son apprentissage, comme illustré dans le schéma.

La mesure critériée regroupe un ensemble de procédures permettant une prise d'information détaillée à propos de l'apprentissage d'un apprenant. Ces procédures ont en commun de mieux définir et de mieux contrôler les critères quantitatifs et qualitatifs de la performance, tels que :

- les aspects de la présentation d'une tâche ;
- les conditions de réalisation d'une tâche ;
- les niveaux d'exigence pour la réalisation d'une tâche.

La mesure critériée permet d'affiner la prise d'information de l'enseignant à propos des apprentissages des élèves et le rend ainsi plus apte à comprendre les raisons de leurs difficultés. La figure 1.8 fournit un exemple de l'utilité de la mesure critériée pour la prise d'information. On y retrouve les résultats bruts des élèves en pourcentages pour chaque objectif d'apprentissage *1* à *k*. Les notes supérieures au seuil de réussite sont indiquées en caractères blancs sur fond noir et les notes inférieures sont représentées en caractères noirs sur fond blanc. Elles forment les balises d'une progression par objectifs. Le seuil de réussite ou de maîtrise a été fixé à 80 %. La lecture verticale permet d'identifier les objectifs qui posent problème et la lecture horizontale, les élèves en difficulté. De ce tableau, il est possible de faire ressortir trois types de cas de « non-maîtrise » des objectifs (valeurs encerclées) :

1. **Le cas de Jean.** Jean n'a pas atteint le seuil de maîtrise de l'objectif 5 et il est le seul du groupe avec Éric à ne pas l'avoir atteint. Cependant, l'objectif *k* est bien maîtrisé, ce qui semble indiquer qu'il s'agit d'une difficulté temporaire et que Jean poursuit normalement sa progression. Il faudrait tout de même s'assurer que les apprentissages prévus à l'objectif 5 ont bien été repris au cas où ils s'avéreraient nécessaires ultérieurement.
2. **Le cas de Lucie.** Lucie n'a pas atteint le seuil de réussite de l'objectif *k*, le dernier en date. C'est aussi le cas de la majorité des élèves. Comme le cas de Lucie n'est pas isolé, cette situation réclame une régulation destinée à l'ensemble du groupe.
3. **Le cas d'Éric.** Cet élève éprouve des difficultés persistantes à atteindre le seuil de réussite et de toute évidence, il accumule les retards. La non-maîtrise de l'objectif 4 n'est que le résultat d'une longue série de difficultés. De toute évidence, Éric a besoin d'un enseignement correctif et d'un plan d'intervention, non seulement pour maîtriser l'objectif d'apprentissage 4, mais aussi tous les autres objectifs du programme.

Élèves	Objectifs						
	1	2	3	4	5	…	K
Jean	82	87	95	88	(70) [1]		83
Louis	87	96	90	87	83		76
Lucie	89	87	86	90	80		(72) [2]
Karina	85	84	80	87	82		80
Julie	86	82	85	85	81		70
Éric	70	82	68	(67) [3]	70		60
Stéphane	81	81	84	85	83		67

80 Note supérieure au seuil de réussite

(72) Note inférieure au seuil de réussite

Figure 1.8 — Balises d'une progression par objectifs

2.4.1 *Analyse et spécification de domaine*

Le modèle de Deno et Jenkins permet un alignement des tâches d'examen sur les objectifs du programme d'études, mais il ne prend pas en compte plusieurs facteurs qui pourraient en réduire la marge d'interprétation, notamment lorsque vient le moment de concevoir des tâches ou des items. Pour pallier ce genre de difficultés, de nouveaux modèles ont été créés pour rendre plus précise la spécification de domaine du test et des items. Nous aborderons les trois suivants :

- **La Conception d'évaluation centrée sur les preuves** (*Evidence-Centered Design : ECD design*). L'approche ECD met l'accent sur l'importance d'utiliser les tâches appropriées pour susciter les preuves nécessaires pour appuyer les affirmations que l'on souhaite faire sur l'apprentissage des élèves. Elle mise notamment sur l'importance de l'alignement entre les tâches d'évaluation, les items et les preuves d'apprentissage et un système d'inférences pour appuyer ces affirmations.
- **L'ingénierie inverse** (*reverse engineering*). Il existe des situations où les apprentissages sont évalués année après année sans référence à un schéma organisateur particulier, tel que des normes d'apprentissage ou les objectifs d'un programme d'études. Le contenu des épreuves est généré par des experts à partir de ce qu'ils considèrent comme pertinent et prioritaire. L'analyse de ces items peut présenter un intérêt particulier. L'ingénierie inverse consiste à induire, à partir d'items et tâches connues, les caractéristiques des résultats d'apprentissage et des tâches qui ressortent comme importantes.
- **La planification à rebours** (*backward planning*). Dans le modèle *Understanding by design* (Wiggins & McTighe, 2005), la planification des activités d'enseignement et d'apprentissage se fait à rebours. Au lieu d'élaborer l'évaluation une fois l'enseignement terminé, l'enseignant planifie d'abord l'évaluation pour ensuite faire en sorte que l'enseignement mette en place ce dont l'élève a besoin pour réussir les tâches considérées comme prioritaires dans le plan. Cette approche n'est pas à confondre avec le phénomène *teach to the test* qui consiste à préparer les élèves indûment à des examens standardisés à enjeux élevés.

Ces trois modèles s'arriment parfaitement avec les théories les plus récentes sur la validité (voir chapitre 4, section 1). Tous trois visent à identifier et à spécifier les preuves nécessaires pour évaluer un apprentissage et faire les inférences possibles à partir de résultats probants.

2.4.1.1. CONCEPTION D'ÉVALUATION CENTRÉE SUR LES PREUVES

La conception d'évaluation centrée sur les preuves (ECD)[5] a été développée principalement par Robert Mislevy. De nombreuses publications traitent de ce modèle et la

5 Nous utiliserons l'acronyme anglais ECD dans ce livre (*Evidence-Centered Design*) car il est couramment utilisé dans les publications.

Introduction aux théories des tests

Un manuel opérationnel rassemblant les informations les plus récentes sur les modèles de la théorie classique des scores, de la généralisabilité et de la réponse aux items :

- spécification de domaine, alignement et développement d'un instrument de mesure
- rédaction d'items et adaptation d'un test en plusieurs langues
- distribution de scores et statistiques de base
- modèles récents sur la fidélité et la validité des résultats
- analyse des items et étude de biais
- transformation et interprétation des scores
- banque d'items et testing sur ordinateur

EN LIGNE : questionnaires de compréhension en fin de chapitres et matériel complémentaire d'apprentissage

Design **cerise.be**

Ph.D. en psychologie et professeur émérite en mesure et évaluation à l'Université d'Ottawa, **Dany Laveault** intervient à titre de chercheur, expert et conférencier auprès de nombreuses associations professionnelles et juridictions scolaires.

Docteur en psychologie et professeur émérite à l'Université de Louvain où il a enseigné la psychométrie et les méthodes de l'évaluation, **Jacques Grégoire** a une longue expérience du développement et de l'adaptation de tests psychologiques et éducatifs. Il est consultant en psychométrie auprès d'organismes publics et privés.

978-2-8073-5107-3

9 782807 351073

39,90 €

www.**deboecksuperieur**.com

Dans le cadre du nouveau Système Européen de Transfert de Crédits (E.C.T.S.), ce manuel couvre en France les niveaux : Licence 1-2-3, Master 1-2.
En Belgique : Bachelier 1-2-3, Master 1-2
En Suisse : Bachelor, Master
Au Canada : Baccalauréat, Maîtrise

L 1-2-3
M 1-2
D