BUREAU
POUR
L'ENSEIGNEMENT
DE LA
LANGUE
ET DE LA
CIVILISATION
FRANCAISES
A L'ETRANGER

LES TESTS DE PROGRES

DANS LA CLASSE DE FRANCAIS.

Jean-Claude MOTHE

BUREAU
POUR
L'ENSEIGNEMENT
DE LA
LANGUE
ET DE LA
CIVILISATION
FRANCAISES
A L'ETRANGER

LES TESTS DE PROGRES

DANS LA CLASSE DE FRANCAIS.

Jean-Claude MOTHE

1.0. Notion d'objectivité.

La caractéristique primordiale qui différencie les tests des moyens traditionnels d'évaluation des connaissances est leur objectivité. L'objectivité n'est pas l'unique qualité exigible d'un test, ni même la plus importante, mais les autres caractères spécifiques des tests en dérivent.

Quand on oppose objectivité et subjectivité dans une évaluation, il faut comprendre qu'il n'est question que de l'attribution des notes ou notation. Comme le remarque A.E.G. Pilliner[1], la subjectivité peut en effet intervenir au niveau de chacune des trois étapes communes à toute épreuve d'évaluation :

- la construction des questions : quelle que soit la procédure, elle est toujours subjective à ce niveau (choix des questions, manière de les poser, etc.)[2].

- la réponse à ces questions par le candidat : là aussi, toute procédure est subjective (choix des réponses).

- la notation de ces réponses : c'est là qu'existe la différence entre subjectivité et objectivité, la notation étant subjective quand l'examinateur doit décider lui-même de l'adéquation de la réponse, objective quand il n'a pas à en décider, la décision ayant été prise d'avance au niveau de la construction des questions.

La distinction entre un examen subjectif et un examen objectif réside donc uniquement dans la manière dont les notes sont attribuées.

1.1.0. Subjectivité des moyens traditionnels d'évaluation.

Les épreuves traditionnelles d'évaluation, dans les examens, les concours, les divers contrôles scolaires, sont conçues, par leur nature même et par la manière dont elles sont notées, de telle sorte que leur notation est très fortement soumise à la subjectivité de l'examinateur.

1.1.1. Nature des épreuves.

Pour s'en tenir au domaine des langues étrangères, les épreuves traditionnelles les plus courantes appartiennent aux types suivants :

- épreuves écrites ou orales de traduction d'un texte, authentique ou fabriqué, soit dans le sens langue cible - langue source (version), soit en sens inverse (thème).

.../...

---

(1) : A.E.G. Pilliner : "Subjective and Objective Testing", in Language Testing Symposium. A psycholinguistic approach - Ed. by Alan Davies - Oxford University Press, London, 1968.

(2) : Voir cependant G. de LANDSHEERE : Le test de closure, Mesure de lisibilité et de la compréhension. Bruxelles, Labor Nathan, 1973.

- épreuves écrites ou orales de questions diverses posées à partir d'un support quelconque (texte, image, situation...) ou sans support (questions de grammaire, de civilisation, d'histoire littéraire, etc...) et nécessitant des réponses relativement longues.

- épreuves écrites de rédaction d'un texte d'une certaine longueur, généralement indéterminée : construction de phrases, résumé de texte, dialogue, essai, narration, dissertation, commentaire de texte, etc...

- épreuves orales de conversation avec l'examinateur ou d'élaboration d'un discours oral suivi : récit, exposé, commentaire de texte...

- épreuves écrites de dictée d'un texte oral.

- épreuves orales de lecture à haute voix d'un texte écrit.

Toutes ces tâches présentent, à des degrés plus ou moins élevés, deux caractéristiques communes qui rendent leur notation nécessairement subjective :

a) Elles donnent lieu à des réponses d'une diversité infinie, cette diversité étant due à :

- des consignes imprécises (raconter, décrire, commenter, apprécier...) laissant aux élèves la liberté de construire leur réponse à leur guise.

- une absence de modèle unique de référence, évident pour les tâches d'expression à consignes extrêmement vagues par nécessité (rédaction, exposé...) mais réel même pour les tâches dont les consignes paraissent plus précises : une traduction peut toujours, pour peu que le texte ne soit pas extrêmement bref, s'effectuer de plusieurs manières différentes également acceptables ; la lecture à voix haute d'un texte quelconque donnera lieu de toutes façons, même avec des locuteurs natifs, à autant d'interprétations différentes que de lecteurs, ou plutôt de lectures ; même un texte dicté admet le plus souvent des variantes également acceptables. La diversité s'amplifie encore dans les épreuves (rédaction par exemple) laissant aux élèves la possibilité d'opter entre plusieurs "sujets" (l'avantage étant de donner aux candidats plus de chances de trouver un thème facilitant).

- une très sensible longueur de la réponse, allant généralement de pair avec le petit nombre des tâches exigées et destinée à le compenser ; pour certaines de ces épreuves (résumé, narration, essai, dissertation, exposé oral, commentaire de texte), la dimension de la réponse constitue même souvent un important élément intervenant dans la notation.

.../...

Il est évident que la diversité de ces réponses rend impossible leur comparaison et leur classement en fonction de critères objectifs.

b) Il s'agit de tâches complexes, dont la notation ne peut pas être fonction d'un critère simple car elles font intervenir des savoir-faire nombreux et différenciés, qui ne vont pas nécessairement de pair chez un sujet déterminé.

L'exemple de la composition écrite est naturellement privilégié et souvent cité à propos de l'évaluation subjective, car les qualités dont on tient compte pour l'établissement de la note sont particulièrement nombreuses et diversifiées. On en a, pour les compositions en langue maternelle, dénombré jusqu'à 17[1] :

1. Lisibilité
2. Esthétique
3. Présentation
4. Exactitude de l'orthographe
5. Exactitude morphologique
6. Exactitude syntaxique
7. Structure de l'exposé
8. Richesse d'idées
9. Pertinence des idées
10. Précision d'information
11. Exhaustivité
12. Concision
13. Propriété du langage
14. Style
15. Originalité
16. Maturité
17. Imagination

Il est vrai qu'une analyse factorielle montre que ces qualités se regroupent en fait en 4 ou 5 catégories[2]. Mais les éléments d'appréciation restent nombreux et radicalement différents.

Il est facile de montrer qu'une conversation, un exposé, une traduction font également intervenir plusieurs types très différents d'habileté. Même l'épreuve de la dictée n'isole pas un savoir-faire simple : elle fait également intervenir la discrimination auditive, la compréhension orale d'éléments lexicaux et de structures grammaticales, la vitesse

.../...

---

(1) : C. Remondino : Etude factorielle sur la notation des compositions scolaires portant sur la langue maternelle, in Le Travail Humain, XXII, Janvier-Juin 1959, p. 27-40, cité par Gilbert de Landsheere. Evaluation continue et examens. Précis de docimologie. Labor, Bruxelles/Nathan, Paris, 1972, p. 123.

(2) : cf. ci-dessous 5.3.2.

d'écriture, la capacité de se relire et de corriger ses propres erreurs, sans omettre la diversité complexe des savoir-faire qu'implique déjà la capacité d'orthographier conformément aux normes de la langue française.

Dans quelle proportion l'examinateur devra-t-il tenir compte, dans sa notation, de ces différents éléments, en supposant qu'il ait pris conscience de leur diversité et en ait dressé une liste exhaustive ? Doit-il les noter séparément en attribuant à chacun d'eux un coefficient, et en ce cas comment éviter l'arbitraire ? Ou doit-il se fier à son impression générale, nécessairement subjective ?

Le plus souvent, dans les examens traditionnels, on ne donne aux examinateurs aucune instruction à ce sujet, sinon des indications très vagues, comme : "On tiendra compte de l'orthographe et de la ponctuation.. On accordera une importance particulière à l'enchaînement logique des idées"... etc. Généralement, on les laisse libres de décider comme ils l'entendent :

- du nombre d'éléments dont ils auront à tenir compte pour l'établissement de la note.
- du choix de ces éléments.
- de l'importance relative qu'il convient de leur accorder.
- de la méthode d'évaluation, synthétique et impressionniste, ou analytique.

1.1.2 Modalités de notation.

Il n'est pas étonnant dans ces conditions que les notations fassent montre d'une grande inconstance. De multiples expériences, dont celles d'Henri Piéron[1] sont parmi les plus systématiques, ont fait la preuve de l'ampleur des différences inter-individuelles et intra-individuelles que peuvent atteindre des notes attribuées à une même copie d'élève par des examinateurs différents ou par le même examinateur à des moments différents. Ces différences, qu'on a quelquefois tendance à exagérer, sont réelles. Beaucoup plus larges pour des tâches scolaires comme la composition philosophique, elles restent sensibles même pour des épreuves de mathématiques ou de physique, les langues vivantes occupant dans cette échelle une place intermédiaire, d'ailleurs variable selon le type d'épreuves.

a) Différences inter-individuelles.

Plusieurs éléments de la notation sont affectés par ces écarts :

- la moyenne : un juge peut être d'une façon générale plus ou moins

.../...

(1) : G. de Landsheere, ouvrage cité, p. 153 et suivantes.

sévère et, pour une même série de copies corrigées par plusieurs examinateurs, la moyenne des notes varie sensiblement en fonction du degré de générosité ou de sévérité de chacun.

- la dispersion : dans une échelle d'évaluation, certains examinateurs utilisent pour leur notation la totalité de l'échelle ; d'autres, choisissant plus souvent la solution de facilité des valeurs centrales, se maintiendront d'une façon constante dans une "fourchette" réduite et n'utiliseront qu'un petit nombre de degrés moyens.

- le classement : on pourrait croire que les deux causes d'écarts précédentes seraient assez aisément réduites en attribuant à chaque juge, après essais, un coefficient de dispersion et un coefficient de sévérité, et en utilisant une procédure de modération qui permettrait de ne considérer comme pertinent dans le classement que le rang. De tels systèmes existent en fait(1), mais les expériences montrent que, pas plus que la note, le rang ne se révèle stable quand on passe d'un juge à l'autre. Le meilleur système de modération est la correction multiple, d'autant moins économique qu'elle tend, par une augmentation du nombre des examinateurs, vers une plus grande approximation de l'objectivité.

b) Différences intra-individuelles.

Ce sont celles qui existent entre des notations effectuées par le même juge à un certain intervalle de temps. H. Piéron(1), entre autres, a montré leur importance.

Ces variations sont dues à toutes sortes de facteurs plus ou moins bien définis : variations de santé et d'humeur, évolution, dans la compétence, modification de l'échelle d'évaluation utilisée, variations contextuelles (un même travail pourra être surévalué ou sous-évalué s'il est corrigé immédiatement après un autre travail très médiocre ou excellent.)

c) Contamination.

D'autres facteurs de subjectivité interviennent dans la notation, et la résistance à ces facteurs, dus à une contamination d'éléments extérieurs à ce qu'on veut mesurer, est également variable selon les examinateurs :

- l'effet de halo, qui fait varier affectivement le jugement de l'examinateur selon un préjugé favorable ou défavorable, dû par exemple à l'allure extérieure du sujet, à son écriture, etc.

- la stéréotypie, qui influence surtout l'examinateur connaissant depuis un certain temps les examinés, et qui fait surévaluer la performance d'un élève qui a obtenu dans le passé de bons résultats, et sous-évaluer celle d'un élève jugé médiocre.

.../...

1.2.0. Les tests objectifs.

C'est de la recherche de l'objectivité dans la notation que sont nés les tests. Il serait illusoire de croire que les tests ont résolu tous les problèmes de l'objectivité, encore moins tous les problèmes de l'évaluation. Pour évaluer certaines tâches, l'objectivité ne sera jamais possible. Il reste qu'on peut essayer de réduire la subjectivité et, dans certains cas, la supprimer complètement. Pour cela, il est indispensable de modifier la nature et la forme des tâches exigées des sujets.

1.2.1. Questions contraignantes et standardisées.

Atteindre l'objectivité de notation implique que l'on supprime le rôle de l'équation personnelle de l'examinateur : la note obtenue devra rester identique quel que soit le correcteur. C'est remettre en cause l'une des exigences des moyens traditionnels d'évaluation, celle de la nécessité de la compétence de l'examinateur : cette compétence, élément important de son équation personnelle, doit s'effacer comme les autres éléments. Le résultat obtenu à un test devra rester constant même si la compétence de l'examinateur est nulle en la matière et à la limite même si le correcteur humain est remplacé par une machine.

Il faut donc bannir tout jugement au niveau de la correction, ce qui ne peut se faire que si deux conditions sont remplies :

- une dichotomie rigoureuse dans l'évaluation d'une réponse : celle-ci ne doit pouvoir être considérée que comme tout à fait correcte ou indiscutablement erronée, à l'exclusion de toute possibilité intermédiaire (cotation 0 ou 1).

- une standardisation non moins rigoureuse de la tâche à accomplir, avec une détermination précise des conditions de réussite : la réponse considérée comme correcte aura été établie à l'avance au stade de la construction de l'épreuve et c'est cette réponse, à l'exclusion de toute autre, qui sera exigée de tous les sujets comme condition de succès ; au cas où plusieurs bonnes réponses, ou certaines variations, seraient acceptables, la liste exhaustive en aura été dressée par le constructeur.

Pour que ces conditions soient remplies, il faut donc que la question ait été conçue comme étroitement contraignante, de façon à limiter au minimum (et si possible à une seule unité) le nombre des réponses acceptables.

1.2.2. Brièveté des réponses.

Si les questions du type traditionnel ne peuvent être considérées comme contraignantes, c'est avant tout parce que les réponses attendues

.../...

des sujets sont trop longues, donc trop complexes (ce qui exclut l'alternative simple "correct/erroné") et trop diversifiées.

Pour que les questions soient contraignantes, il est donc nécessaire qu'elles n'exigent que des réponses les plus brèves possibles. Dans un test de langue, cette longueur minimale sera par exemple celle d'un mot graphique, d'un segment de mot, d'un syntagme bref, etc., ou bien la réponse sera non-linguistique (cocher une case, par exemple), l'essentiel étant que la brièveté soit suffisante pour que toutes les possibilités acceptables puissent être énumérées à l'avance, ou réduites à une seule.

Il faut préciser que cela n'implique pas que ce type de réponses soit le seul possible, ni surtout le seul souhaitable. C'est simplement le seul qui soit objectif. Plus la dimension de la réponse exigée s'écartera de cette longueur minimale, plus l'objectivité de notation sera difficile à atteindre.

1.2.3. Multiplicité des réponses.

Il est évident qu'une réponse aussi brève n'aurait pas de signification si elle restait isolée : le but de la mesure, quel qu'il soit, ne sera pas atteint si on ne dispose pas d'un nombre suffisant d'éléments d'appréciation. L'inconvénient des réponses brèves ne peut être évité que si elles sont nombreuses.

La multiplicité des réponses améliore d'ailleurs la validité d'un test[1] en ce qu'elle permet un sondage plus étendu et plus diversifié des capacités des sujets : à quantité totale égale, des tâches brèves et nombreuses auront sur des tâches longues et peu nombreuses l'avantage d'être à la fois moins aléatoires et plus représentatives de l'ensemble du champ exploré.

1.2.4. Définition des tests de langue.

Epreuve à questions nombreuses, contraignantes, standardisées et nécessitant une réponse brève, telle pourrait être la définition formelle des tests, par opposition aux questions peu nombreuses, imprécises et demandant une réponse longue et diversifiée des épreuves de type traditionnel.

En fait, si en anglais le mot "test" désigne n'importe quelle sorte d'épreuve, en français le terme est rarement réservé, du moins en ce qui concerne les tests de langue, à la catégorie d'épreuves qui correspondent à cette définition et qu'on appelle en anglais "objective tests". Il est pris le plus souvent en un sens plus large, englobant des épreuves où la

.../...

(1) : cf. ci-dessous 3.0.

brièveté et la standardisation des réponses ne sont pas exigées, par exemple les épreuves d'expression, comme dans le test C.G.M. 62[1].

Il semble donc que le terme s'applique en français à toute épreuve conçue pour approcher le plus possible les meilleures conditions d'objectivité et, plus généralement, de fidélité.

1.3. Notion de fidélité.

L'objectivité de notation n'est en effet que l'élément essentiel de ce qu'on appelle la fidélité d'un test, c'est-à-dire la qualité qui fait que les résultats obtenus restent stables et constants :

- qu'on fasse passer l'épreuve tel jour ou tel autre,
- qu'on la note, tel jour ou tel autre,
- qu'elle soit notée par tel ou tel examinateur.

Il existe des moyens d'évaluer la fidélité d'une épreuve. Le moyen le plus simple, celui qu'on pratique par exemple dans les sciences physiques, est de répéter la mesure, pour vérifier la constance des résultats. On peut utiliser ce moyen avec un test de langue, en faisant passer deux fois le même test aux mêmes sujets (test-retest avec le même test) et en calculant le coefficient de corrélation entre les deux épreuves[2].

Mais il se pose alors un problème que les sciences physiques ignorent et qui est celui de l'intervalle entre les deux passations :

- si cet intervalle est trop court, malgré l'effet d'apprentissage, les sujets auront tendance à confirmer leur première réponse, qu'ils auront retenue, et la fidélité de l'épreuve sera surestimée.

- si cet intervalle est trop long, les capacités des sujets se modifieront (en progression si l'apprentissage continue pendant l'intervalle, en régression si l'apprentissage est suspendu) et la fidélité sera sous-estimée.

Pour éluder cette difficulté, on administre, au lieu du même test, deux tests équivalents (test-retest avec tests parallèles), ou les deux moitiés d'un même test (items pairs/items impairs). Encore faut-il présumer que ces deux tests ou ces deux moitiés de test sont équivalents. Le coefficient de fidélité d'un test[2] ne peut donc être qu'approximatif. Pour l'améliorer, le constructeur s'efforcera d'éliminer dans la mesure

.../...

(1) : voir ci-dessous 5.3.1.

(2) : voir ci-dessous 6.3.

du possible les facteurs d'inconstance en prenant certaines précautions :

- au niveau de la construction du test, en multipliant les items pour augmenter le nombre de tâches différentes ; ce nombre n'est pourtant pas indéfiniment extensible : il conviendra au contraire de limiter la durée totale de l'épreuve afin de ne pas atteindre le seuil de fatigue au-delà duquel diminueraient à la fois la fidélité (ce seuil de fatigue pouvant varier selon les jours et les heures chez un individu donné) et la validité (la dimension de l'endurance à la fatigue venant interférer avec ce qu'on cherche à mesurer et biaisant les résultats. Cf. chapitre 3).

- toujours au niveau de la construction, en standardisant au maximum les tâches à accomplir, que ces tâches soient exactement les mêmes pour tous les sujets quand c'est possible ou, lorsque c'est impossible (tests d'expression), qu'elles soient de difficulté comparable.

- au niveau de la passation, en vérifiant que les conditions de passation sont identiques pour tous les sujets : s'il s'agit par exemple d'un test d'audition nécessitant l'écoute d'un enregistrement, il faut éviter que les sujets placés près du magnétophone se trouvent dans des conditions matérielles d'écoute meilleures ou moins bonnes que les autres ; s'il s'agit d'une épreuve en temps limité, ce qui est le cas lorsque la vitesse fait partie des capacités à mesurer, il faut veiller à ce que la durée de l'épreuve soit la même pour tous. D'une façon générale, il faut tenir compte du fait que l'état de santé, la motivation, la réaction affective à l'égard du test constituent autant de variables individuelles des sujets qui tendent à diminuer la fidélité d'un test en affectant les performances. Si la première de ces variables échappe à toute correction (excepté la répétition différée de l'épreuve), il n'est pas impossible de réduire les deux autres, et spécialement les réactions affectives négatives, qui diminuent avec l'habitude des tests, donc avec l'augmentation de leur fréquence (aux dépens, il est vrai, de la motivation).

- au niveau de la correction, par la standardisation maximale, ce qui fait d'ailleurs difficulté pour les tests d'expression[1].

La notion de fidélité s'applique donc uniquement à la précision de l'instrument de mesure, comme l'indiquent les termes de constance, de stabilité, d'objectivité dont on fait usage pour expliciter ce qu'elle recouvre. Mais en admettant que cette qualité soit atteinte d'une façon satisfaisante, cela ne garantirait en aucune façon que le test mesure

.../...

(1) : cf. ci-dessous 5.3.

effectivement ce qu'on voudrait qu'il mesure. C'est là qu'intervient la notion de validité, incomparablement plus importante que celle de fidélité car, à choisir, un instrument qui mesure sans précision ce qu'on veut mesurer est préférable à un instrument qui mesurerait avec précision... autre chose.

C'est pourtant par l'importance qu'il accorde à la recherche de la fidélité que, comme nous l'avons vu, un test se définit formellement en opposition à d'autres épreuves d'objectif comparable. Pour préciser cette définition formelle des tests de langue, il vaut mieux, avant d'aborder les problèmes de la validité, cesser de se placer sur ce plan de généralité, car la forme que peuvent revêtir les tests se diversifie en fonction des objectifs qu'on leur assigne.

2.0. Objectifs des tests de langue.

Une évaluation comporte toujours deux objectifs, dont l'importance relative est variable : un diagnostic et un pronostic[1].

Même dans les deux cas extrêmes, où l'un de ces deux objectifs semble effacer complètement l'autre, l'effacement n'est jamais total. Dans le cas des tests d'aptitude par exemple, où le pronostic est primordial, on ne peut éliminer de l'opération d'évaluation l'aspect diagnostique, du moment que la prédiction s'appuie nécessairement sur l'observation de comportements dans une large mesure acquis. Inversement, lorsqu'un examen ou un test paraît destiné uniquement à mesurer un apprentissage passé, cette mesure ne peut être conçue qu'en fonction de décisions à venir, qui seront par exemple une orientation différenciée ou une modification de l'apprentissage.

.../...

---

(1) Ces deux termes sont pris ici au sens large, "pronostic" au sens de prédiction d'événements à venir, "diagnostic" au sens de reconnaissance d'événements passés. On peut restreindre la signification du terme de "diagnostic" en la rapprochant de sa valeur dans la terminologie médicale et en y incluant les sèmes "recherche des causes" et "interprétation qualitative". Il faut alors distinguer comme le fait G. DE Landsheere (ouvrage cité, p. 14-15), trois rôles de l'évaluation, le rôle pronostique, le rôle diagnostique proprement dit ("Pourquoi un apprentissage parfait ne s'est-il pas produit ? Quelles matières ou techniques l'étudiant domine-t-il insuffisamment, quels sont les processus mentaux en cause ?") et un rôle de "jaugeage" ("a) Contrôle des acquisitions ; b) Evaluation du progrès, cas où l'on compare l'élève à lui-même ; c) Situation de l'élève à un moment donné" dans un ensemble d'élèves plus ou moins étendu). En ce sens, un test de classement, par exemple, ne jouerait aucun rôle diagnostique.

Cependant, le degré de validité de la mesure sera plus grand si les objectifs sont nettement établis, et en particulier si l'objectif primordial est défini comme diagnostique ou pronostique. Cette définition est très souvent absente dans les examens de type traditionnel comme le baccalauréat français, dont l'objectif est hybride, puisqu'il est à la fois l'examen de contrôle de la formation secondaire et un brevet d'aptitude à la formation universitaire.

C'est en fonction de ce critère qu'on pourra distinguer deux grandes catégories de tests, distinction applicable aux tests de langue, les tests de pronostic et les tests de diagnostic.

2.1. Les tests de pronostic.

Tournés vers l'avenir et utilisés à des fins avant tout prédictives, ils peuvent être divisés en deux sous-catégories :

- les tests d'aptitude (aptitude tests), qui doivent fournir une indication statistique des probabilités de succès dans un type déterminé d'apprentissage, par exemple, celui d'une langue étrangère. Sous réserve de ce qui a été dit plus haut, ces tests ne sont pas censés tenir compte des acquisitions passées.

- les tests de niveau (proficiency tests), qui évaluent les acquisitions des sujets, mais avec l'intention explicite d'utiliser les résultats obtenus pour procéder à un certain classement parmi ces sujets. C'est le cas des tests de sélection, qui aboutissent à une répartition en deux classes par rapport à un seuil, qui détermine le succès ou l'échec : le niveau de tel sujet sera, selon qu'il atteint ou non ce seuil, considéré comme suffisant ou insuffisant pour suivre un certain type de cours, obtenir un certain avantage, remplir certaines fonctions . C'est également le cas des tests d'orientation ou de classement, qui visent à une répartition en deux ou plusieurs classes en fonction de programmes différents de perfectionnement (qualification professionnelle, recyclage, répartition entre groupes de débutants, moyens, avancés, etc.).

On voit que les tests de niveau supposent la détermination préalable d'une limite (plusieurs même dans le cas de certains tests d'orientation) séparant les différents niveaux[1].

.../...

(1) : Sur les test d'aptitude et de niveau, voir Frédéric François et Emmanuel Companys - Les tests de langue - B.E.L.C. - Paris, multigr., 1969.

2.2. Les tests de diagnostic.

Contrairement aux tests de pronostic qui sont d'abord conçus comme un point de départ, les tests de diagnostic, bien que tournés dans une certaine mesure vers l'avenir comme nous l'avons vu, sont surtout un point d'aboutissement, un bilan.

Les tests de niveau évaluent les acquisitions des sujets en fonction d'un programme futur, et indépendamment de tout programme passé, les tests de diagnostic s'appliquent au contraire à des sujets ayant en principe suivi le même programme, ce programme passé étant connu.

On peut également distinguer deux catégories de tests à l'intérieur de la classe des tests de diagnostic, bien que cette distinction ne soit pas aussi nette et profonde qu'entre tests d'aptitude et de niveau :

- les tests de contrôle (achievement ou attainment tests) mesurent le degré d'acquisition d'un programme déterminé, à l'issue d'une période d'enseignement considérée comme un cycle complet (année scolaire, scolarité), sans que cette mesure soit nécessairement fondée sur un cours ou une méthode unique : seul est commun le programme supposé couvert dans la période en question.

- les tests de progrès (progress tests), comme les tests de contrôle, sont construits à partir d'un programme ; mais ils mesurent le degré d'acquisition d'une unité de cours particulière, dont les dimensions peuvent être extrêmement réduites[1].

.../...

---

(1) : Dans certaines terminologies, par exemple celle d'Antoine Beck ("les tests dans l'enseignement des langues vivantes", Cahiers d'allemand n° 1, septembre 1970) ce qui est appelé ici "test de niveau" est nommé "test de qualification" ou "d' efficience", l'expression "test de niveau étant considérée comme synonyme de "test de contrôle" (ou "d'acquisition") et l'appellation "test de diagnostic" réservée aux tests de progrès.

Jean-Guy SAVARD (Bibliographie analytique de tests de langue. Presses de l'Université Laval, Québec, 1969), réserve également la dénomination de "test de diagnostic" aux tests "dont le but est de déceler les faiblesses du sujet en vue de lui donner, par la suite, l'enseignement conectif approprié", donc aux tests de progrès. Il réserve en outre le nom de "test de pronostic" aux tests d'aptitude, et réunit tests de niveau et de contrôle sous le nom de tests de rendement.

Ce qui distingue ces deux derniers types de tests est donc apparemment davantage une différence de degré qu'une différence de nature. Alors que les tests de contrôle sont conçus à une grande échelle et doivent donc être soigneusement expérimentés et standardisés, les tests de progrès, bien que certains auteurs de manuels en aient prévu pour accompagner leur matériel didactique, sont plutôt construits par le professeur lui-même pour les élèves de sa classe, et lui permettent de voir dans quelle mesure les objectifs de telle unité de cours qu'il a enseignée ont effectivement été atteints et d'ajuster son enseignement en conséquence.

L'évaluation des acquisitions par les tests de diagnostic peut s'appliquer à plusieurs objets différents (à la fois ou séparément) :

- aux sujets eux-mêmes, que de tels tests permettent de différencier sur le plan de la facilité ou de la vitesse d'acquisition, des capacités de rétention et de transfert, de la motivation, de la persévérance dans l'effort, etc., et de classer d'une manière d'autant plus fine que l'évaluation est plus diversifiée (tests de contrôle), ou de comparer par rapport à un stade antérieur d'apprentissage et/ou par rapport aux objectifs du cours (tests de progrès).

- aux méthodes ou aux manuels utilisés, dont on peut ainsi comparer à travers les résultats obtenus sur des populations réputées de niveaux et d'aptitude similaires, l'efficacité et le degré d'adaptation à la population considérée (tests de contrôle).

- aux enseignants, dont on peut évaluer l'efficacité dans les mêmes conditions (tests de contrôle).

- aux processus d'apprentissage particuliers dont les composantes sont le groupe-classe, le professeur et l'unité de cours (tests de progrès).

En fait, il y a réellement une différence de nature entre les tests de contrôle, qui se placent après (donc en dehors de) l'apprentissage, et les tests de progrès, qui y sont intégrés et interagissent avec lui : c'est que l'évaluation par les tests de contrôle est sommative, tandis que l'évaluation par les tests de progrès est formative[1]. Les tests formatifs n'ont pas pour but de comparer ou classer les sujets entre eux

.../...

(1) : Expressions crées par Michael SCRIVEN, in R. STAKER, Ed., Perspectives of of Curriculum Evaluation, Chicago, Rand McNally, 1967.

Voir Benjamin S. BLOOM, J. Thomas HASTINGS, George F. MADAUS - Handbook on formative and summative evaluation of student learning - Mc Graw Hill, New York, London, 1971.

(ni de comparer des méthodes ou des professeurs), mais de déterminer a) dans quelle mesure chacun d'eux a acquis ce qu'il avait à acquérir ; b) ce qu'il lui reste à acquérir. L'évaluation formative des résultats obtenus par un sujet ne s'effectue pas par comparaison avec les résultats obtenus par les autres sujets, mais par rapport aux objectifs de l'apprentissage. Cette distinction importante place les tests de progrès (les tests de diagnostic au sens restreint) tout à fait à part des autres tests à bien des points de vue.

2.3. Tests de classe et tests standardisés.

La seule catégorie de tests qui puissent être construits par l'enseignant lui-même est celle des tests de progrès. Les tests des autres catégories ne peuvent en effet avoir d'intérêt pratique que s'ils sont conçus pour être appliqués à un grand nombre de sujets :

- les tests de pronostic (aptitude, niveau) ne peuvent jouer leur rôle de prédicteur s'ils ne sont pas construits en fonction de données statistiques, et ces données seraient peu sûres si elles ne tenaient pas compte des résultats obtenus par un grand nombre de sujets.

- les tests de contrôle sont administrés d'une manière uniforme à des sujets dont l'apprentissage ne s'est pas effectué dans les mêmes conditions (les méthodes, les professeurs, les établissements pouvant être différents).

Pour tous ces types de tests, l'exigence de l'application à une population numériquement importante implique une construction par une équipe de spécialistes, dont il est d'ailleurs inconcevable que des professeurs ne fassent pas partie si l'objet du test est une discipline scolaire, mais qui doit comporter aussi des psychologues et des statisticiens. Même les tests de progrès, s'ils sont réalisés par exemple par les auteurs d'une méthode pour l'ensemble des élèves qui sont appelés à l'utiliser, tombent alors dans le même cas.

L'importance numérique de la population visée influe considérablement sur la conception et la forme d'un test, et on peut opposer sur de nombreux points les tests de classe, ou exercices-tests (angl. classroom tests), à l'ensemble des tests standardisés destinés à une population nombreuse :

- les tests de classe, construits, administrés et notés par la même personne peuvent faire l'économie des instructions très précises que les constructeurs d'un test standardisé doivent élaborer à l'intention des administrateurs et des correcteurs (à moins que ces derniers soient remplacés par des machines). .../...

- les objectifs des tests de classe sont étroitement dérivés de ceux du cours, et le professeur a une idée précise du niveau de ses élèves, ce qui n'est pas le cas du constructeur d'un test standardisé. Disposant d'un corpus beaucoup mieux défini, le constructeur se heurtera d'une manière bien moins aigüe au problème de l'échantillonnage des questions et de leur représentativité.

- un test standardisé s'adressant à de nombreux sujets, l'économie[1] de passation et de correction y devient une qualité importante dont il convient de tenir compte. C'est pour ce type de tests que la correction mécanique constitue, outre ses garanties d'objectivité, un avantage appréciable sur le plan de l'économie (temps de correction). Si les exigences d'économie de correction vont dans le même sens que les exigences de fidélité, elles risquent par contre de nuire à la validité du test (risque de choisir les questions parce qu'elles permettent la correction mécanique plutôt que parce qu'elles sont représentatives), et c'est également le cas du souci d'économie de passation (risque d'abréger à l'excès la longueur des épreuves). Ces problèmes sont étrangers au constructeur d'un test de classe, qui n'a que peu de sujets à examiner et pour lequel il n'est pas question de correction mécanique pour des raisons de coût financier.

- inversement, si le petit nombre de sujets permet au constructeur de test de classe de perdre quelques minutes pour la correction de chaque feuille de réponse, le souci de rentabilité l'empêche de consacrer à la construction même du test un temps excessif. Or c'est au niveau de la construction que le test standardisé se révèle peu économique, le grand nombre des sujets et l'ampleur du programme (ou, dans le cas des tests de pronostic, la difficulté qui représente la validité prédictive) exigent une préparation et une expérimentation minutieuses.

- les tests standardisés, administrés en des occasions exceptionnelles, sont contraints à une évaluation isolée, rarement corrigée par une seconde évaluation à court terme, de sujets inconnus du constructeur, Les tests de classe peuvent être fréquents, les résultats qu'ils provoquent ne donnent pas lieu à une évaluation définitive impliquant de graves conséquences, le constructeur connaît très vite les sujets auxquels le test est destiné, et c'est en fait beaucoup moins les sujets eux-mêmes que son propre enseignement que le professeur a besoin d'évaluer.

- les tests standardisés ont la plupart du temps en tête de leurs objectifs celui d'effectuer parmi les sujets un classement, c'est-à-dire une répartition en un certain nombre de classes (deux ou plus) ; ils doivent donc posséder une puissance discriminative suffisante, ce qui

.../...

(1) Le mot d'économie recouvre ici celle d'économie de temps plus encore que celle d'économie financière.

signifie qu'ils doivent être conçus de telle sorte qu'ils soient faciles pour les meilleurs candidats et difficiles pour les moins bons ; la recherche de cette finesse discriminative constitue dans la préparation de ces tests l'objet d'une partie importante de la pré-expérimentation. Généralement, sauf cas particuliers, les tests de classe n'ont pas cette exigence : le professeur préparant ses items en fonction des buts qu'il s'était assignés, son espoir est qu'ils soient réussis par le plus grand nombre possible de ses élèves. Un item trop facile ou trop difficile d'un test standardisé doit être éliminé parce que peu discriminateur donc inutile ; au contraire, un item qui se révèlera très facile ou trop difficile dans un test de classe fournira au professeur des informations utiles sur ce qui a été acquis ou non par ses élèves. C'est que la plupart des tests de classe entrent dans la catégorie des tests se référant à un critère ("criterion-referenced tests"), tandis que la plupart des tests standardisés sont des tests se référant à une norme ("norm-referenced tests")(1).

- les tests standardisés sont généralement administrés au dehors de la classe, et c'est l'évaluation qui en est la fonction essentielle, si ce n'est la seule. Les tests de classe s'insèrent, eux, dans un processus d'apprentissage, et c'est sans doute pour eux la fonction primordiale : ils suivent les objectifs du cours et réagissent sur eux, ils permettent aux élèves de constater leurs progrès, qu'ils peuvent stimuler, et leurs lacunes, qu'ils peuvent concourir à combler, ils aident le professeur à mieux ajuster son enseignement, bref leur rôle est celui d'une évaluation formative.

On trouve chez les auteurs américains, en particulier dans le manuel sur le testing des langues de Rebecca Valette(2), une distinction entre le test de classe et le "quiz" : "La distinction entre le test et le quiz est une question de dimension et d'utilisation plutôt que de contenu". Tandis que le test de classe est annoncé à l'avance, couvre une unité spécifique du cours, qui a fait l'objet de plusieurs séances en classe, et peut durer une séance entière, "le principe du quiz est la brièveté. Contrairement au test, il peut être donné à l'improviste...

.../...

(1) : cf. Edward David ALLEN, Rebecca M. VALETTE- Modern Language Classroom Techniques. A Handbook. Harcourt Brace Jovanovich, New-York, 1972, p. 36

(2) : Rebecca M. VALETTE - Modern Language Testing. A Handbook. Harcourt, Brace and World, New-York, 1967, p. 7.

On peut dire aux élèves de s'attendre à un quiz à chaque séance, même si certains jours, on n'en donne pas... La valeur du quiz réside surtout dans son effet positif sur l'apprentissage et dans la pratique qu'il donne, dans l'art de passer des tests... La valeur du test réside dans l'exhaustivité du contrôle des acquisitions du matériel à étudier". On peut sans grand inconvénient considérer le "quiz" comme un cas particulier de test de progrès, et ses caractéristiques comme celles d'un test de classe de dimensions restreintes.

Il est normal, et d'ailleurs heureux, que les tests de classe soient beaucoup moins difficiles à élaborer que les tests standardisés : les problèmes de fidélité, d'économie, de puissance discriminative ne se posent pas du tout dans les mêmes termes. Mais, pas plus que les tests standardisés d'ailleurs, ils n'échappent au problème crucial de la validité.

3.0. Notion de validité.

La validité est la qualité à la fois la plus importante et la plus difficile à atteindre pour n'importe quelle épreuve d'évaluation. C'est la qualité qui fait que l'épreuve mesure effectivement ce qu'elle est censée mesurer et non autre chose. C'est donc d'abord une qualité qui n'est pas absolue ni intrinsèque au test, mais relative à son objectif. Comme le dit R. Lado[1] : "la validité n'est pas générale, mais spécifique. Si un test de prononcitation mesure la prononciation et rien d'autre, c'est un test valide de prononciation, ce ne serait pas un test valide de grammaire ou de vocabulaire parce qu'il ne teste pas la grammaire ou le vocabulaire".

Pour un test de progrès, la notion de validité peut être débarrassée de certains des aspects qu'elle recouvre dans d'autres formes de tests, c'est-à-dire :

- la validité prédictive, qui exprime une correspondance entre la réussite à un test prédicteur, et un succès ultérieur plus général (avec toutes les difficultés que représentent d'une part le critère d'évaluation de ce succès ultérieur, d'autre part le calcul de sa corrélation avec la réussite au test) : cet aspect de la validité, capital pour un test de pronostic, est de peu d'importance quand on n'accorde pas au test de valeur prédictive.

.../...

(1) : Robert LADO : Language Testing. The construction and use of foreign Language tests. Longmans, London, 1961, p. 30

- la validité concourante, ou validité de corrélation, qui n'est autre qu'une validité prédictive établie indirectement, puisqu'elle est la qualité d'un test qui donne des résultats comparables à ceux obtenus avec un autre test déjà reconnu comme prédictivement valide.

La validité d'un test de progrès se limite essentiellement à ce qu'on appelle sa validité de contenu, qui est déjà bien difficile à atteindre, ne serait-ce que parce que, contrairement à la validité prédictive ou concourante, elle ne peut se calculer par des coefficients.

On parlera de validité de contenu lorsqu'il y a correspondance d'une part entre le contenu du test et ses objectifs, d'autre part (du moins pour les tests de diagnostic), entre le contenu et les objectifs du test et ceux de l'apprentissage, chacune de ces correspondances pesant le double problème de l'adéquation et de la représentativité de la tâche imposée par le test.

3.1. Adéquation du contenu du test à ses objectifs.

La remarque de Lado citée ci-dessus éclaire bien ce que signifie une telle adéquation : un test à contenu phonétique ne saurait avoir des objectifs lexicaux. Il est pourtant moins aisé d'atteindre l'adéquation du test à ses objectifs que ne pourrait le laisser croire un exemple aussi évident.

Supposons par exemple que, pour vérifier si un groupe d'élèves anglophones a bien assimilé l'opposition entre les adjectifs possessifs masculins et féminins à la 3ème personne du singulier (opposition son/sa tout à fait différente de l'opposition his/her en anglais), on donne un test qui aurait par exemple la forme suivante :

Consigne : complètez comme il convient les phrases suivantes en écrivant dans le blanc l'un de ces deux mots : son, sa.

Items :

1. C'est le frère de Marie ? - Non, c'est .... cousin.
2. Pierre est dans .... chambre.
3. Paul lit .... livre, etc.

L'objectif de ce test est visiblement de vérifier si les élèves sont capables de surmonter les interférences dues à la règle anglaise, et de répondre correctement "son cousin" bien qu'il s'agisse du cousin de Marie (anglais "her cousin") et "sa chambre" bien qu'il s'agisse de la chambre de Pierre (anglais "his bedroom").

.../...

Il est facile de voir pourtant que, tel qu'il est conçu, ce test manque son objectif : des réponses erronées aux items 1 et 2 peuvent être interprétées soit comme une méconnaissance de la règle grammaticale en question, soit comme une lacune purement lexicale (genre des substantifs "cousin" et "chambre") ; une réponse érronée à l'item 3 ne peut même recevoir que la seconde interprétation.

Pour que l'adéquation du contenu à cet objectif précis puisse être réalisée, il faudrait modifier les items de telle sorte que l'information sur le genre du substantif, considérée dans ce test comme non pertinente pour l'évaluation, soit fournie d'une façon quelconque. Par exemple :

1. C'est le cousin de Marie ? - Oui, c'est bien ..... cousin.
2. Pierre aime bien ..... petite chambre.

La question ne se poserait pas si le test portait par exemple sur la distribution de sa et son féminins, car l'information nécessaire est apportée par le contexte (initiale du mot suivant).

1. Marie n'aime pas .... école.
2. Pierre a amené .... petite amie.

On peut trouver des erreurs d'adéquation de ce genre dans des tests publiés. Par exemple ces deux items extraits d'un sous-test grammatical d'un test de classement[1] :

Consigne : choisissez la forme convenable de l'article.

Items :

- Voici ..... garage.
  1. le
  2. la
  3. l'
  4. les
  5. - (2)

- ..... Brésil est un grand pays.
  1. Le
  2. La
  3. L'
  4. Les
  5. - (2)

Il est vrai qu'il s'agit ici d'un test de classement, et non d'un test de diagnostic, et que l'objectif de ces items est de vérifier la connaissance de toutes les règles applicables aux articles, et non d'une seule. Mais on peut douter de la validité d'un item portant sur la connaissance du genre du substantif "Brésil".

.../...

(1) : Test Laval. Formule A. Test de classement. Français langue seconde — Procédé de l'Université Laval, Québec. 1971

(2) : Ce signe veut dire qu'aucun mot n'est nécessaire dans ce contexte.

3.2. Représentativité du contenu du test par rapport à ses objectifs.

Supposons maintenant que, pour vérifier si la discrimination auditive entre les différents phonèmes vocaliques du français est suffisante chez un certain groupe d'étudiants, on prépare un test dont tous les items porteraient sur les oppositions entre les 3 voyelles d'aperture minimale, /i/, /y/ et /u/. Ce test peut être parfaitement valide par rapport à ce jeu spécifique d'oppositions, il ne le sera pas par rapport à l'objectif supposé, qui était beaucoup plus étendu.

On dira alors que le contenu du test n'est pas représentatif par rapport à son objectif : les résultats obtenus sur ce contenu partiel n'autorisent pas d'inférences sur les résultats qui auraient pu être obtenus sur le contenu plus vaste supposé par l'objectif. La déficience de la validité sera moins grave que dans les exemples précédents, dans la mesure où les résultats obtenus ne seront pas faux, mais seulement partiels. Toute généralisation serait pourtant hasardeuse.

Contrairement à ce qui se passe pour l'adéquation du contenu à l'objectif, le problème de la représentativité de cet échantillon réduit qu'est nécessairement le contenu d'un test (on ne peut pas tout tester, et le test n'est qu'un sondage) se pose évidemment d'une manière d'autant plus aigüe que ce qu'on cherche à mesurer est plus vaste et complexe. Il sera plus difficile à résoudre pour un test de niveau que pour un test de progrès, pour un test "de français" que pour un test de compréhension écrite du français, pour un test visant à contrôler les acquisitions d'une année scolaire que pour un test mensuel, pour un test standardisé que pour un test de classe, pour un test de compréhension lexicale que pour un test de discrimination phonétique.

On ne pourra considérer un tel échantillon comme représentatif que si on a pu au préalable :

- délimiter le champ dont on aura à extraire l'échantillon ; c'est facile lorsqu'il s'agit d'un ensemble fini d'éléments, par exemple de l'ensemble des oppositions phonologiques du français standard, ou des mots du français fondamental (encore que la polysémie et les contraintes syntagmatiques viennent singulièrement compliquer le problème) ; c'est impossible lorsque le champ est ouvert et que le nombre des éléments est infini (les structures grammaticales ou les oppositions lexicales du français).

- isoler les uns des autres les éléments de l'ensemble (ces éléments pouvant être des relations entre les éléments d'un autre ensemble), ce qui est aisé lorsque ces éléments sont discrets (oppositions phonologiques), mais arbitraire lorsqu'il s'agit de degrés sur un continuum (schémas intonatifs).

- choisir parmi ces éléments plus ou moins bien isolés d'un ensemble plus ou moins bien défini ceux qui feront partie de l'échantillon. Sauf si l'inventaire de ces éléments est très restreint, on se heurte alors au problème des critères de choix, qu'on ne peut résoudre que d'une façon arbitraire, et empirique.

Un début de solution consiste à distinguer dans l'ensemble en question plusieurs sous - ensembles, et à prendre garde que chacun de ces sous-ensembles soit représenté dans le test. Pour reprendre un exemple commode, il est certain que le contenu d'un test pourtant sur la connaissance des mots du français fondamental constituera un échantillon moins représentatif s'il n'est composé que de substantifs du Français fondamental. Mais on voit que non seulement le problème précédent n'est pas résolu entièrement (selon quels critères choisis parmi les éléments des sous-ensembles "verbes" ou "adjectifs du Français fondamental" ?), mais qu'il s'en crée d'autres : d'abord, le critère qui permet une partition en sous-ensembles ne s'impose pas toujours d'une façon aussi évidente que celui des catégories grammaticales pour les éléments lexicaux, il peut y en avoir d'autres (par exemple distinction entre mots retenus dans le F.F.1 pour leur fréquence et mots disponibles) qui définissent autant de partitions différentes se recoupant les uns les autres et pouvant aboutir, pour peu que l'ensemble soit vaste et complexe, à un nombre très élevé de sous-ensembles qui ne pourront pas être tous représentés dans le contenu du test. Surtout, cette division en sous-ensembles pose le problème délicat du dosage : quelle part accorder à chacun d'eux, et en fonction de quels nouveaux critères effectuer une pondération ? (dans le F.F.1, consacrera-t-on davantage d'items aux substantifs parce qu'ils sont plus nombreux, ou aux verbes parce qu'ils sont plus fréquents ?). Il faut sans doute tenir compte de critères comme la fréquence, la productivité, la disponibilité, et les contradictions inévitables entre tous ces critères ne peuvent se résoudre, elles aussi, que d'une façon empirique et arbitraire.

5.3. Adéquation et représentativité du contenu et des objectifs d'un test par rapport au contenu et aux objectifs de l'apprentissage.

Ce problème ne se pose en fait que pour les tests de diagnostic, donc pour les tests de progrès. Les tests de niveau se heurtent à une difficulté qui n'est pas moins grande, mais qui est différente : c'est celle de l'adéquation et de la représentativité du contenu, et des objectifs du test, par rapport au contenu supposé d'un apprentissage non connu, et aux objectifs de performance[1] qui ont été définis comme seuil de succès par le constructeur du test.

.../...

(1) voir ci-dessous 5.5.

Pour ~~les tests de~~ diagnostic, leur contenu doit être adéquat à celui de l'apprentissage : on ne peut par exemple tester validement la capacité de traduction de sujets qui n'ont pas appris à traduire, ou l'orthographe d'élèves qui n'ont pas encore appris à écrire. Il doit également en être représentatif : il ne serait pas valide de fonder un test de compréhension lexicale à un niveau avancé sur un corpus de mots excluant systématiquement le contenu du F.F.1, ou ne retenant que lui.

Surtout, un test de diagnostic, et spécialement un test de progrès, doit être adéquat et représentatif des objectifs de l'apprentissage, ce qui suppose que ceux-ci ont été définis avec suffisamment de précision. En effet, un test de diagnostic ne peut par exemple se contenter de vérifier si ce qui a été enseigné a effectivement été acquis, car l'enseignement n'est pas une fin en soi, il doit autant que possible permettre de se rendre compte si le but de l'apprentissage a été atteint.

Ainsi, si l'on a enseigné, à travers un certain nombre d'exemples et d'exercices, une structure grammaticale quelconque, un test qui se contenterait de reprendre dans ses items des énoncés extraits de ces exemples ou de ces exercices ne permettrait pas de vérifier si l'unité didactique a atteint son objectif, qui n'était pas de faire mémoriser un nombre limité d'énoncés mais de faire assimiler une règle permettant à l'élève d'en comprendre et d'en produire un nombre illimité. Pour être valide, le test devrait donc permettre de contrôler si cette capacité de transfert a bien été acquise, ce qui ne peut se faire qu'à travers des énoncés nouveaux comportant des éléments inconnus qui n'ont pas fait l'objet de l'apprentissage(1).

Les buts de l'apprentissage d'une langue étrangère ne se limitent d'ailleurs pas à développer chez les élèves cette capacité de transfert dans une situation étroitement contrôlée par le professeur : ils comportent un ~~stade plus~~ complexe et plus élevé encore qui est la capacité de communication en situation authentique(2). Idéalement, cette capacité devrait aussi pouvoir faire l'objet d'une évaluation malgré les difficultés énormes auxquelles se heurte une évaluation de ce genre(3).

En fait, les tests de progrès n'étant pas isolés mais constituant une série, il n'est pas nécessaire qu'un test de la série soit à lui seul adéquat et représentatif de l'ensemble des objectifs de l'apprentissage, même s'il ne s'agit que des objectifs d'une unité en cours restreinte.

.../...

---

(1) : Le Centre de Linguistique Appliquée de Besançon a introduit dans certains de ses tests de niveau des items portant sur la compréhension d'éléments lexicaux supposés inconnus des sujets, mais placés dans des contextes éclairants.

(2) : Voir ci-dessous 5.5. la taxonomie des objectifs de performance dans l'apprentissage d'une langue étrangère.

(3) : ci-dessous 5.2.7. et 5.3.

Certains de ces objectifs, surtout ceux qui se trouvent dans les niveaux les plus élevés de la taxonomie, sont en effet des objectifs à long terme, qu'il serait inopportun et psychologiquement néfaste d'essayer d'évaluer prématurément. Mais si, comme il est souhaitable, le professeur a pu fixer avec précision ses objectifs de performance, en définissant les comportements terminaux qu'il attend de tout ou partie de ses étudiants, à la fin de l'année ou du cycle scolaire, il peut planifier son programme de tests de telle sorte que la série une fois réalisée soit adéquate et représentative de ces objectifs.(1)

3.4. Compétence et performance.

La grande difficulté, pour un test de langue, vient en fait de la contradiction, apparemment insoluble, qui existe entre le but de l'apprentissage d'une langue étrangère, qui est de développer chez les élèves une compétence nouvelle, et l'impossibilité de tester directement cette compétence, autrement que par l'intermédiaire des performances.

Ceci n'est d'ailleurs que l'aspect, particulier aux tests de langue, du grand problème de la psychométrie, qui est l'impossibilité de mesurer directement un trait psychologique et la nécessité de recourir à un intermédiaire, seul mesurable, qui est un comportement. C'est à partir de l'observation et de la quantification d'opérations qui auront été reconnues au préalable comme émanant de tel trait psychologique qu'on infèrera des conclusions sur ce trait(2).

Pour les tests de langue, la validité sera pour une grande part fonction de la valeur des inductions qu'on pourra tirer sur la compétence linguistique à partir de performances provoquées, et qui seront soumises :

- au choix du type de performances que l'on tentera de provoquer (problèmes de l'adéquation du contenu et de la représentativité de l'échantillon, évoqués ci-dessus).

- à l'interprétation qualitative des performances réalisées effectivement, et en particulier des échecs par rapport à la norme choisie par le constructeur du test.

Les fautes que commettent les élèves apprenant une langue étrangère, aussi bien en situation de classe qu'en situation de testing, sont de deux sortes :

.../...

(1) : Voir Robert F. MAGER - Preparing instructional objectives, Fearon Publishers, Palo Alto (California), 1962.
Traduction française de Georges DECOTE : Comment définir les objectifs pédagogiques, Gauthier-Villars, Paris, 1971.

(2) : cf. Anna BONBOIR. La méthode des tests en pédagogie. P.U.F., Paris, 1972
Introduction : "Problème de la psychométrie".

- celles qu'ils seraient capables de ~~corriger eux-mêmes s'ils en prenaient conscience, que la terminologie de la pédagogie traditionnelle~~ appelle les "fautes d'inadvertance" ou "d'inattention" et qui sont en quelque sorte des lapsus.

- les erreurs proprement dites, que leur auteur serait incapable de corriger spontanément.

Les premières ne sont que des accidents de performance, tels que les locuteurs natifs en connaissent eux-mêmes. Les secondes se placent au niveau de la compétence dont elles trahissent des lacunes. Du point de vue de l'acquisition d'une compétence, l'importance de ces deux sortes de fautes n'est pas du tout la même. Et cependant, sur le plan des performances, donc du testing, elles se traduisent de la même manière.

Faut-il donc conclure que toute évaluation de la compétence est impossible ? Deux remarques peuvent pourtant corriger l'impression que la compétence ne peut être atteinte à travers les performances :

1) On peut considérer les accidents de performance de la même manière que les lapsus, c'est-à-dire comme des actes manqués significatifs. Il arrive souvent dans une situation d'apprentissage qu'un élève laisse échapper une faute qu'habituellemeмt il ne commet pas. Certes, il "sait" qu'il s'agit d'un énoncé inacceptable et il est capable de se reprendre, souvent même spontanément et immédiatement. Mais la réalisation de la faute est le signe que la compétence n'est pas solidement acquise et que, bien qu'existante (puisque l'énoncé correct peut être produit), elle reste fragile et incomplètement autonomisés de la compétence parallèle en langue maternelle, en supposant que la faute commise est due à une interférence.

2) S'il est vrai qu'un échec au niveau de la performance n'est pas la preuve d'une absence de compétence, et qu'inversement une performance réussie ne signifie pas nécessairement que la compétence existe, il n'en est pas de même lorsqu'une performance isolée est remplacée par un faisceau de performances judicieusement choisies par le constructeur du test. La répétition de l'échec ou du succès dans des énoncés mettant en jeu la même règle est probablement significative au niveau de la compétence, dont on peut de la sorte avoir une idée assez précise avec une marge d'erreur raisonnable.

Malgré ces réserves, on ne peut que s'en tenir, pour le testing, aux performances, puisque celles-ci sont seules directement accessibles. C'est au niveau de l'établissement du programme d'enseignement que se posera le problème de fixer les objectifs en termes de performances, et

.../...

non en termes nécessairement vagues de compétence. Il appartiendra alors au professeur d'essayer d'atteindre ces objectifs, et au constructeur de tests de vérifier dans la mesure du possible si ces objectifs ont été atteints.(1)

4.0. Le testing des langues et les théories linguistiques et psychologiques.

L'usage qui vient d'être fait des termes "compétence" et "performance" qui appartiennent à la terminologie chomskyenne nous amène à poser le problème des relations entre le testing (et l'apprentissage) des langues et les théories linguistiques et psychologiques qui en sous-tendent la conception.

Ces relations semblent inévitables du moment que l'apprentissage des langues est l'acquisition d'un comportement linguistique et que le testing en est la mesure. C'est pourtant seulement à partir du structuralisme que de telles relations ont été explicitées, et pour le testing des langues c'est Robert Lado qui en a fourni le modèle le plus élaboré.

4.1. Apport du structuralisme.

Le modèle de Lado(2), construit à partir d'une théorie linguistique, le strudturalisme bloomfieldien, et d'une théorie psychologique, le behaviorisme skinnérien, a fait faire au concept de validité de contenu dans le testing des langues des progrès décisifs, dont certains n'ont été remis en question par la suite que dans la mesure où la nécessité, l'utilité et la possibilité du testing ont elles-mêmes été remises en question.

Les principes formulés par Lado et les structuralistes pour le testing des langues sont généralement des corollaires de leur théorie de l'apprentissage, Lado ayant explicitement insisté sur le rapport de subordination qui lie le testing à l'apprentissage, à deux réserves près :

- la "statégie" du testing peut être différente de celle de l'apprentissage.

.../...

(1) : Voir J.B. CARROLL : "The Psychology of Language Testing", in Language Testing Symposium - ouvrage cité.

(2) : Robert LADO - Language Testing. The construction and use of foreign language tests. Longmans, London, 1961.

- il y a en fait interaction entre les deux, puisque les résultats obtenus à un test de diagnostic réagissent sur l'apprentissage en modifiant le comportement du professeur et celui des étudiants.

Certains de ces principes restent valables, dans la mesure où les principes d'apprentissage auxquels ils correspondent le restent aussi[1] :

a) Eviter la traduction.

Si on admet que la traduction est un exercice difficile, que l'on ne peut introduire qu'à un niveau avancé de l'apprentissage, il faut admettre aussi que tout recours à la traduction est à éliminer du test tant que ce niveau avancé n'est pas atteint.

Il en découle que :

- la traduction ne sera pas utilisée dans un test en tant que tâche explicite (consignes du type : "traduisez" ou du type "choisissez parmi les traductions suivantes").

- on aura soin de ne pas y faire appel comme tâche implicite, c'est-à-dire comme moyen d'effectuer une autre tâche (emploi de la langue maternelle dans les questions ou les réponses proposées).

Bien entendu, ce principe d'élimination de la langue maternelle ne s'applique pas aux consignes du test, qui doivent être comprises parfaitement de tous les sujets (le résultat obtenu par un sujet serait sans signification en cas de mauvaise interprétation des consignes) et qui doivent donc au contraire être données en langue maternelle chaque fois que c'est possible (groupe linguistique homogène).

b) Tester la connaissance de la langue, non les connaissances sur la langue.

Il est peut-être utile pour l'enseignant, dans certains cas, de faire réflèchir ses élèves à propos de la langue cible, en leur fournissant des explications grammaticales ou en utilisant une terminologie métalinguistique. Mais, à moins de former des spécialistes, il n'emploie ces procédés que comme moyens d'approche pédagogique, non comme but de l'apprentissage.

Si le principe structuraliste "Teach the language, not about the language" peut être considéré comme excessif, en ce sens qu'un enseignement sur la langue peut quelquefois faciliter un enseigne-

(1) : Voir aussi D.P. HARRIS : "The linguistics in language testing", in Language Testing Symposium (ouvrage cité).

ment de la langue, il n'en est pas de même du corollaire "Test the language, not about the language" dans la mesure où seule compte en définitive la connaissance de la langue.

Sont donc à rejeter, sous peine de diminuer la validité du test :

- la vérification de connaissances acquises à propos de la langue (exemple : identifier des parties du discours).

- les questions qui exigent une telle connaissance pour qu'une réponse correcte y soit fournie. Exemple : "complètez les phrases suivantes en écrivant dans les blancs l'adjectif démonstratif qui convient".

Il n'est pas nécessaire de bannir totalement des consignes toute terminologie métalinguistique, il suffit de faire en sorte que son éventuelle utilisation ne pénalise pas ceux des sujets qui ignoreraient ou en auraient oublié la signification. Exemple : "complètez les phrases suivantes en écrivant dans les blancs l'un des adjectifs démonstratifs suivants : ce, cet, cette, ces".

c) Eviter le point de vue normatif.

Une langue, c'est ce que disent ceux qui la parlent comme leur langue maternelle, ce n'est pas ce qu'on peut penser qu'ils devraient dire.

Ce principe de l'enseignement d'une langue étrangère, qui tend à remplacer les préoccupations normatives traditionnelles par un point de vue descriptif, s'applique également au testing, et probablement sans soulever de problèmes aussi épineux.

Il est en effet plus facile par exemple de décider qu'on ne pénalisera pas dans un test un énoncé comme "j'en veux pas" ou "c'est quoi ?" que de décider qu'on l'enseignera.

d) Distinguer les différents niveaux de langue.

Les énoncés cités ci-dessus appartiennent à un certain registre de la langue orale. Le testing, comme l'enseignement, aura à tenir compte d'une part des différences entre langue parlée et langue écrite, d'autre part des différents registres de l'un et l'autre codes.

Sans doute, à un niveau avancé d'apprentissage, un programme de tests qui ne mesurerait que la maîtrise de la langue écrite littéraire, ou que celle de la langue orale conversationnelle, n'aurait-il, à moins d'objectifs très précis de performances, qu'une validité limitée.

Les a-priori structuralistes de primauté de l'oral, malgré leur caractère trop systématique et universel de pédagogie centrée sur la

.../...

méthode plutôt que sur l'élève, ont eu le mérite de tirer l'apprentissage et le testing (d'une langue seconde) de l'impérialisme de la langue écrite. Le constructeur d'un test ne manquera pas de tenir compte, pour des raisons d'économie, du champ important d'intersection entre langue orale et langue écrite, mais la frange particulière de chacun des deux codes devra faire l'objet d'un apprentissage, donc d'un contrôle, spécifiques.

e) Ne pas exiger de performances que ne peuvent fournir les locuteurs natifs.

Un test, devrait, chaque fois que c'est possible, être expérimenté avec des locuteurs natifs d'un âge et d'un niveau culturel proches de celui des sujets testés. Cela permettrait d'éviter l'erreur consistant, dans le cas d'un test de français, à éxiger de sujets non francophones des performances que ne pourraient produire ou que ne produiraient pas spontanément des francophones placés dans les mêmes conditions ; par exemple :

- une activité métalinguistique (voir plus haut § b)
- des performances se référant à une norme extérieure aux habitudes linguistiques réelles des locuteurs natifs (voir § c)
- des résolutions d'ambiguïtés, des choix, des oppositions, des exclusions que certains locuteurs natifs n'effectuent pas.

Ce principe, qui peut paraître naturel et d'application simple, pose en fait des problèmes complexes dont la solution n'est pas évidente. Pour prendre l'exemple des oppositions phonologiques en français, exigera-t-on d'élèves étrangers qu'ils sachent discriminer et reproduire correctement l'opposition /ɛ̃/~/œ̃/ que la plupart des Français du Nord ignorent ? ou l'opposition /a/~/ɑ/ complètement absente de tout le Midi et en voie de disparition ailleurs ? ou les oppositions phonologiques /o/~/ɔ/ , /e/~/ɛ/ ; /ø/~/œ/ , qui, au moins dans la moitié méridionale de la France, se réduisent à de simples paires de variantes combinatoires d'un phonème unique ? ou l'opposition /ɲ/~/nj/, neutralisée dans certains idiolectes ?

L'attitude de l'enseignant et celle du constructeur de test peuvent diverger ici. Le professeur pourra en effet décider que ces oppositions doivent être connues de ses élèves, au moins passivement. Mais s'il élabore un test de progrès, devra-t-il sanctionner des "erreurs" que des Français pourraient tout aussi bien commettre ?

.../...

Le problème du choix de la norme, problème difficile de l'enseignement d'une langue seconde, se pose pour le testing en des termes différents.

L'expérimentation d'un test sur des locuteurs natifs permet également d'en faire une validation partielle sur un autre plan : des réponses "erronées" (il serait préférable de dire : non prévues par le constructeur) produites par un locuteur natif peuvent en effet avoir d'autres causes que les variantes géographiques, par exemple :

- question mal posée ou ambigüe.
- question faisant appel à une dimension autre que la compétence linguistique, par exemple l'intelligence, la mémoire, des connaissances autres que linguistiques, l'endurance à la fatigue.
- consignes insuffisantes.

Bien entendu, ce genre d'expérimentation sur des locuteurs natifs suppose :

- que ces locuteurs natifs ne soient pas gênés par l'usage de la langue maternelle des sujets, par exemple dans les consignes du test ;
- qu'ils jouent convenablement leur rôle d'informateurs en tentant honnêtement de répondre de leur mieux aux questions posées ;
- que leur âge et leur niveau d'instruction soient comparables à ceux des sujets testés : un enfant francophone ne saurait être utilisé pour la validation d'un test destiné à des adultes étrangers ;
- qu'il se trouve des locuteurs natifs disponibles, ce qui n'est pas souvent le cas quand il s'agit d'un test de progrès construit par le professeur pour sa classe.

Cette expérimentation, indispensable pour un test de niveau et même pour un test de contrôle, est simplement utile pour un test de progrès.

Quant au second type d'expérimentation préalable prévu par Lado, celle qui s'applique à un échantillon de population considéré comme représentatif de la population qui sera soumise au test définitif, n'offre guère d'intérêt pour un test de progrès construit par le professeur lui-même, du moins en termes de rentabilité. Il en serait autrement dans le cas d'un test de progrès intégré à une méthode commercialisée, où une expérimentation sur un échantillon représentatif, avant la publication du test, permettrait, par une analyse des items du prétest, d'opérer parmi ces items une sélection et un classement. L'ampleur de la population à laquelle est destiné le test définitif rend alors rentable ce raffinement.

.../...

4.2. Limites du structuralisme.

Le modèle de Lado a donc eu le mérite, pour ce qui est de la validité d'un test de langue, d'en préciser la notion et d'en permettre une approche moins hasardeuse en dénonçant certaines erreurs graves que l'on commettait dans le testing des langues vivantes à ses débuts.

Encore convient-il de préciser que les principes énumérés ci-dessus n'ont pas tous été explicités ou développés par Lado, et que très peu l'ont été en rapport direct avec la notion de validité.

C'est justement dans la partie explicite de sa théorie, dans le sous-chapitre qu'il intitule 'Theory of foreign language testing" (p. 22 et suivantes), que se trouve ce qui nous paraît aujourd'hui le plus contestable, et qui peut se réduire à deux affirmations :

1) "Une langue est un système d'habitudes de communication... Ces habitudes engagent les plans de la forme, de la signification et de la distribution à plusieurs niveaux de structure, à savoir ceux de la phrase, de la proposition, du syntagme, du mot, du morphème et du phonème" (p. 22).

2)"Les problèmes de l'apprentissage [d'une langue étrangère] peuvent être prédits et décrits dans la plupart des cas par une comparaison linguistique systématique des structures des deux langues... Les problèmes sont constitués par les éléments et les modèles qui n'ont pas d'équivalent dans la langue maternelle, ou dont les équivalents ont une distribution ou une signification structuralement différente" (p. 24)

Ces affirmations sont sous-tendues :

- d'une part par le schéma skinnerien stimulus - réponse de l'apprentissage d'une langue ;

- d'autre part, par l'idée que, la langue se décomposant par une série d'analyses en éléments de plus en plus simples (de la phrase au phonème), l'apprentissage d'une langue (donc aussi le testing de cet apprentissage) suppose l'isolation de ces éléments et leur restructuration en un nombre fini de structures qui peuvent être enseignées (et testées) séparément ;

- en troisième lieu par l'hypothèse selon laquelle une analyse contrastive entre la langue source et la langue cible permettrait de dresser une liste exhaustive des problèmes posés à des sujets parlant une langue maternelle donnée par l'apprentissage d'une langue étrangère donnée, et d'y concentrer le contrôle, en négligeant les éléments qui, parallèles dans les deux langues, ne posent pas de problèmes particuliers.

.../...

Dans sa critique de "Verbal Behavior" de Skinner, Chomsky[1] a montré que les deux premières des hypothèses ci-dessus ne sont susceptibles de rendre compte ni de l'acquisition d'une compétence linguistique ni du fonctionnement d'une telle compétence, dont relève la créativité "gouvernée par les règles" d'un sujet parlant. On peut donc contester la validité de tests qui procèdent à une atomisation de la langue en une poussière d'éléments et de structures déclarées isolables, dont il suffirait de choisir judicieusement un certain nombre pour obtenir un échantillonnage représentatif de la langue en question.

D'autre part, il semble bien que la linguistique contrastive n'ait pas justifié les espérances que l'enseignement des langues avait placées en elle, peut-être parce qu'elle est liée à l'illusion qu'il existe en toute langue un nombre fini de structures. De toutes façons, aussi bien pour l'apprentissage que pour le testing, elle ne peut être utile que pour des groupes linguistiquement homogènes, ce qui n'est pas toujours le cas. De plus, il n'est pas certain qu'une analyse contrastive entre deux langues (tâche gigantesque qui n'a encore jamais été menée à bien) apporterait beaucoup plus d'indications au professeur et au constructeur de tests que l'analyse empirique d'un corpus de fautes suffisamment étendu. Enfin, il peut paraître contestable de limiter l'enseignement et le contrôle à ce que Lado appelle les "problèmes", en négligeant ce qui est (ou qu'on croit être) acquis par transfert à partir de la langue maternelle et en aboutissant ainsi à faire du testing ce que G. Perren appelle "testing for error" plutôt qu'un "testing for success"[2].

L'échec du behaviorisme, incapable de rendre compte des phénomènes d'acquisition d'une langue seconde, et la remise en question du structuralisme bloomfieldien en linguistique appliquée à l'enseignement des langues, posent les limites du modèle de Lado. Si l'on note que les applications pédagogiques de la théorie chomskyenne, beaucoup plus éloignée, dans sa formalisation, de l'apprentissage que les modèles précédents, n'ont pas encore donné de résultats bien probants, et qu'aucune théorie psycholinguistique n'est encore venue prendre la relève du behaviorisme skinnérien pour rendre compte des mécanismes mystérieux de l'apprentissage d'une langue étrangère, on en vient à se demander s'il est justifiable que l'enseignement et le testing des langues prétendent transposer

.../...

(1) : Noam CHOMSKY? "Un compte-rendu du comportement verbal de B.F. Skinner" (traduction), Langages n° 16 - décembre 1969.

(2) : G.E. PERREN : "Testing ability in English as a second language. English Language Teaching, 1967. Vol. 21 et 22 - cité par Cl. TRUCHOT "Les tests de langue. Réévaluation critique". Les langues Modernes, 65ème année n°2 mars-avril 1971.

une seule théorie linguistique et une seule théorie psycholinguistique quelles qu'elles soient.

Parmi tous les principes que les structuralistes ont cru pouvoir transposer tout naturellement de la linguistique Bloomfieldienne à l'enseignement des langues, beaucoup sont remis en question ou, s'ils continuent à être admis, le sont pour des raisons toutes différentes, en relation avec les objectifs spécifiques de l'enseignement. Par exemple, la priorité que l'on peut donner à la langue orale dans l'enseignement d'une langue vivante ne peut en fait trouver aucune justification dans le fait linguistique que n'importe quelle langue naturelle est d'abord orale et que le code écrit (pour celles qui en possèdent un) n'est qu'une transposition secondaire. Une telle justification, absolue, aprioriste et purement linguistique, n'est pas pertinente en pédagogie. Pour que toute priorité accordée à l'oral soit justifiable dans une situation donnée d'apprentissage, il faut avoir d'abord déterminé les objectifs de cet apprentissage et avoir reconnu que ces objectifs comportaient l'utilisation de la langue apprise comme langue de communication orale. C'est seulement ensuite qu'on pourra tenir compte des faits linguistiques dans la démarche pédagogique, et admettre que la comparaison linguistique entre les deux codes montre que leur apprentissage simultané serait surtout préjudiciable à l'apprentissage du code oral.

Néanmoins, s'il est vrai que la critique chomskyenne a remis en cause le principe même du testing, il reste que les institutions d'enseignement, depuis la simple classe de langue jusqu'aux responsables nationaux des examens et des programmes dans la plupart des pays, se trouvent confrontés aux problèmes de l'évaluation et doivent les résoudre. Il serait abusif de condamner les tests de langue sous le prétexte qu'il n'existe plus de théorie du testing, tout comme il serait injustifié de renoncer à enseigner les langues étrangères sous le même prétexte. Même si leurs fondements théoriques sont aujourd'hui mis en pièces, les méthodes audio-visuelles fonctionnent. De la même façon, l'absence de fondements théoriques du testing des langues, n'empêche pas de construire des tests.(1) Et ce qui est vrai des tests standardisés dont l'objectivité peut être garantie par un appareil scientifique rigoureux, l'est à plus forte raison pour les tests de classe, qui ne peuvent prétendre à cette rigueur, mais qui peuvent empiriquement tenter de se rapprocher le plus possible des conditions idéales de validité.

Les principes structuralistes ne sont pas nécessairement à rejeter en bloc, ni pour l'enseignement des langues, ni pour le testing. Qu'ils soient insuffisants pour rendre compte des mécanismes de l'apprentissage ne signifie pas que l'enseignant n'a pas à en tenir compte du tout : ils apportent une aide précieuse dans l'analyse et la structuration du

savoir à transmettre et dans l'opération de feed-back qui doit nécessairement réguler cette transmission en la contrôlant. Cette analyse et cette restructuration sont encore indispensables, au stade actuel de la méthodologie des langues, à l'apprentissage et à son évaluation. La "révolution" chomskyenne a eu pour effet salutaire d'en montrer les insuffisances. Il appartient au méthodologue, à l'enseignant et au constructeur de test de faire empiriquement la part de ce qui est à conserver du structuralisme et de ce qui lui fait défaut. Si la démarche inspirée du structuralisme reste en définitive à peu près satisfaisante au niveau des débutants, chez lesquels, faute d'une compétence linguistique suffisante, c'est la "créativité qui change les règles" qui prédomine et qu'il faut combattre, il semble que ses déficiences s'aggravent à mesure que l'apprentissage se complète et que la "créativité gouvernée par les règles" se développe : à ce niveau, l'enseignement comme le testing se condamnent au piétinement et à l'échec s'ils ne se libèrent pas du carcan structuraliste.

5.0. Les différentes sortes d'items.

Parmi les nombreuses formes que peut prendre un item dans un test de langue, on peut opérer plusieurs classements différents selon qu'on utilise l'un des critères suivants :

- le canal de communication (oral, écrit)
- le pôle de communication (réception, émission)
- la composante linguistique (phonétique/orthographe, grammaire, lexique)
- le niveau de performance linguistique (automatismes, connaissance, transfert, communication, critique).

5.1. Les quatre skills du comportement linguistique.

Les deux premiers des critères ci-dessus pourraient être définis autrement, mais la référence faite à la communication en fait ressortir l'ambivalence lorsqu'il est question des tests de langue.

Tout test en effet est communication : une communication qui s'établit entre le constructeur du test et le sujet testé. Quels que soient le contenu et l'objet du test, le constructeur devra choisir à la fois :

- celui des deux canaux (oral ou écrit) qui sera le support le plus approprié à l'objectif du test ;
- celui des deux pôles (réception ou émission) où il conviendra de placer le sujet selon le comportement attendu de lui.

.../...

Mais quand le test porte sur une langue, le contenu du test est lui-même communication et il s'agit pour le constructeur d'établir si le sujet, utilisant l'un des deux canaux et placé à l'un des deux pôles, est capable d'assurer la communication. Ce n'est pas un hasard si la combinaison de ces deux critères aboutit à quatre modes de communication qui correspondent aux quatre dimensions psychologiques, aux quatre "skills" du comportement linguistique : "compréhension"(1) orale(2) (audition), production orale (élocution), "compréhension" écrite (lecture), production écrite (écriture).

Pour un test autre qu'un test de langue, le choix du canal de communication se fera pour de simples raisons de commodité pratique, et ne sera dicté que par la situation de test elle-même. Dans la plupart des cas, l'écrit offre sur l'oral des avantages pratiques évidents : sur le plan des questions, la forme écrite ne nécessite aucun matériel coûteux, aucune disposition particulière lors de la passation, aucune précaution pour assurer la bonne interprétation des questions (pour autant que la compréhension des questions ne constitue pas précisément l'objet du test) ; sur le plan des réponses, elle permet une passation collective (d'où économie de temps), une conservation des réponses qui ne nécessite pas de matériel spécial, la possibilité d'une correction objective et même mécanisée. Par contre, dans certains cas particuliers, comme celui où les sujets testés sont des illettrés (par exemple de jeunes enfants), le support écrit ne pouvant être que non-linguistique, toute communication linguistique, question ou réponse, devra être orale.

.../...

---

(1) : Ce terme ambigu doit être débarrassé ici de ses implications sémantiques : il désigne dans cet usage le comportement linguistique dit "interne" ou "passif", par opposition à la production, qui est un comportement "externe" ou "actif". Il serait peut-être plus approprié de parler de "réception" pour recouvrir à la fois la perception et la compréhension proprement dite.

(2) ; L'anglais fait une distinction logique entre le terme "oral", qu'il réserve à la production et le terme "aural", qui s'applique à l'audition. Bien que l'expression "compréhension orale" soit effectivement impropre, strictement parlant, elle est entrée dans l'usage en français, et peut d'ailleurs facilement s'interpréter comme "compréhension de l'oral", de même que "compréhension écrite" signifie "compréhension de l'écrit". Deplus, cela permet de regrouper commodément sous la même appellation les épreuves qui font intervenir le même canal de communication. De même, l'usage également impropre du terme "compréhension" (note 1 précédente) permet, par opposition à "production" de désigner d'un seul mot les épreuves qui placent le sujet à l'un des deux pôles de la communication.

Le pôle de communication affecté au sujet sera lui aussi, dans un test de connaissances ordinaire, choisi en fonction de la situation de test. Le sujet sera placé en position d'émetteur, en situation de production, lorsqu'on attendra de lui une réponse qu'il aura construite (oralement ou par écrit) par ses propres moyens, en remémorant activement les connaissances nécessaires à l'expression de sa réponse ; il met alors en oeuvr sa capacité de rappel. Au contraire, il se trouvera en position de récepteur, en situation de compréhension, lorsqu'il n'aura qu'à sélectionner entre plusieurs réponses qui lui sont fournies par le constructeur du test (question à choix multiples)(1). Cette fois, il ne s'agit pas d'un rappel actif de ses connaissances, mais de la reconnaissance relativement passive d'une réponse qu'il ne produit pas lui-même.

Or la plupart des expériences portant sur les tests de contrôle des connaissances ont montré une haute corrélation entre les résultats obtenus à une épreuve de compréhension (dans laquelle le sujet choisit entre plusieurs réponses proposées) et à une épreuve de production (où le sujet construit lui-même sa réponse) ; le constructeur choisira donc, selon les conditions du test, soit le premier type d'épreuves, dont la correction est plus économique et plus aisée à rendre objective, soit le second, plus économique au niveau de la construction.

Dans un test de langue, les quatre modes définis par les deux canaux et les deux pôles de la communication ont un second statut : ils constituent aussi des dimensions psychologiques à tester, et ce rôle peut se trouver en contradiction avec celui qu'ils jouent dans la forme même du test, d'où des solutions de mixité dans la conception des items.

S'il s'agit par exemple de tester la compréhension orale, il est évident que cette communication que constitue le test sera axée sur la question(2) (qu'il s'agit de "comprendre") et que le canal utilisé pour formuler cette question sera oral. Cela n'empêchera pas qu'au niveau de

.../...

(1) : Sont réunis ici sous la dénomination de questions à choix multiples non seulement les items ainsi nommés par Lado ("multiple-choice items") et qui comportent au moins 3 options, mais également les items à deux options, du type "vrai ou faux" ("true-false items").

(2) : Le terme "question" est pris ici au sens de "stimulus primaire" (voir lexique). Il s'agit de ce qui doit être compris du sujet pour que la réponse qu'il fournit soit correcte : dans le cas d'une épreuve portant sur la compréhension d'un passage, par exemple, la "question" sera le passage lui-même, ou un élément de ce passage, et non le stem de chaque item, qui est dans ce cas un stimulus secondaire.

la réponse, le canal utilisé puisse être l'écrit, et que le mode de communication puisse être celui qui est habituellement considéré comme celui de la production (réponse ouverte, construite par le sujet), si les conditions pratiques du test le rendent préférable. Ce sera alors ce qu'on appelle un test hybride : par exemple, pour tester la compréhension d'un texte oral, des items écrits exigeant une réponse écrite construite par le sujet dans la langue-cible. Un tel test est en fait malgré les apparences un authentique test de compréhension orale, et peut se révéler préférable, tant pour des raisons d'économie que de validité, à un test dit pur, qui fait intervenir le même skill unique aux deux niveaux de la forme et du contenu (dans notre exemple, des items également oraux et des réponses à choix multiples).

Un test hybride de compréhension écrite est également possible. Par contre, un test de production, orale ou écrite, ne peut être que pur, puisque, aussi bien sur le plan du contenu que sur celui de la forme, c'est la réponse seule qui est pertinente à l'exclusion de la question, le pôle de communication où est placé le sujet étant celui de l'émission, et qu'une réponse à choix multiples ne pourrait guère être considérée comme valide dans un test de production.

Le tableau ci-dessous résume les compatibilités (indiquées par le signe +) entre les skills à tester (contenu) et ceux qui interviennent dans la forme du test (support). Les tests purs sont ceux qui figurent dans la diagonale du tableau.

FORME DU TEST
(skill support)

	Compréhension orale = question orale, réponse à choix multiples.	Compréhension écrite = question écrite, réponse à choix multiples.	Production orale = réponse orale, réponse ouverte.	Production écrite = réponse écrite, réponse ouverte.
Compréhension orale = audition	+	-	+	+
Compréhension écrite = lecture	-	+	+	+
Production orale = élocution	-	-	+	-
Production écrite = écriture	-	-	-	+

Remarques.

I. - On appelle aussi quelquefois hybride une épreuve dont le stimulus primaire et le stimulus secondaire empruntant des canaux différents : c'est par exemple le cas d'un test de compréhension d'un texte oral (stimulus primaire) sur lequel sont posées des questions écrites, avec options également écrites (stimulus secondaire). Ce n'est qu'une convention terminologique, mais on peut considérer un tel test comme pur, puisqu'il y a harmonie entre forme et contenu, le stimulus secondaire n'ayant pas d'autre objet que de provoquer la réponse.

II. - Le stimulus d'une épreuve de production peut être non-linguistique (image). Cela n'empêche pas le test d'être pur puisque, s'agissant d'un test de production, c'est la forme de la réponse et non celle de l question qui est pertinente.

III. - La répartition, devenue traditionnelle, des tests de langues en ces 4 catégories, qui correspondent aux 4 skills de l'activité linguistique, ne doit pas être considérée comme absolue et étanche. On peut en effet imaginer des épreuves qui mettent en oeuvre à la fois deux de ces quatre skills. Dans la vie courante d'ailleurs, on trouve moins de situations dans lesquelles les skills de l'activité langagière fonctionnent isolément, que de situations où deux d'entre eux sont mis en même temps à contribution : en situation de dialogue par exemple (la plus fréquente sans doute), chacun des interlocuteurs est à la fois et à tout moment émetteur et récepteur[1]. Or, pour qu'un test de langue soit réellement valide, il faudrait que la situation de testing se rapproche le plus possible des situations authentiques dans lesquelles se manifeste l'activité linguistique. C'est l'une des contradictions les plus difficiles à résoudre des tests de langue : comment concilier l'exigence d'analyse et d'atomisation d'une évaluation diagnotique avec l'exigence de synthèse et de globalité d'une évaluation valide. Cette contradiction est d'ailleurs la même que celle de l'apprentissage d'une langue, en dehors d'une situation naturelle.

Le tableau de la page suivante donne une idée des situations authentiques mettant en oeuvre un ou deux des skills de l'activité linguistique, et des situations de testing qui peuvent s'en rapprocher le plus.

.../...

(1) : Voir Antoine CULIOLI : La formalisation en linguistique. Les Cahiers pour l'Analyse, n° 9, juillet 1968. Article repris et annoté dans Antoine CULIOLI, Catherine FUCHS, Michel PECHEUX. Considérations théoriques à propos du traitement formel du langage. Dunod, Centre de Linguistique quantitative de la Faculté des Sciences de l'Université de Paris, 1970.

Skills mis en oeuvre	Situation authentique	Situation de testing		
		Question		Réponse
		Stimulus primaire	Stimulus secondaire	
Audition	Ecouter une émission radiophonique ou télévisée ; voir un film ; écouter un cours, un discours, une conférence, une histoire.	Texte oral	Questions de compréhension orales ou écrites.	Sélection parmi options proposées.
Elocution	Raconter une histoire, relater un événement, exposer son point de vue (?)	(Stimulus visuel)	(Questions orales ou écrites)	Récit oral, exposé, réponses orales aux questions.
Lecture	Lire le journal, un livre, une affiche.	Texte écrit	Questions de compréhension écrites (ou orales)	Sélection parmi options proposées.
Ecriture	Ecrire une lettre, rédiger un article, un tract, une motion	(stimulus visuel)	(Questions écrites)	Récit écrit, rédaction réponses écrites aux questions.
Audition + Elocution	Dialoguer	(texte oral)	Questions orales, répliques de conversation.	Entretien dirigé, réponses orales aux questions.
Audition + Ecriture	Prendre des notes à un cours, un exposé, une conférence.	(texte oral)	(Questions orales ou écrites)	Résumé écrit, dictée, réponses écrites aux questions.
Audition + lecture	?	--	--	--
Lecture + Ecriture	Répondre à une lettre, prendre des notes sur un texte écrit.	(texte écrit)	Questions écrites	Résumé écrit, compte-rendu, réponses écrites aux questions.
Lecture + Elocution	Lire à haute voix(?), résumer oralement un article, un livre, raconter une histoire lue	Texte écrit	Questions écrites ou orales.	Lecture à haute voix, résumé oral, réponses orales aux questions.
Elocution + Ecriture	?	--	--	--

Correspondance entre les skills de l'activité linguistique, les situations authentiques qui les mettent en oeuvre et les situations de testing qui s'en rapprochent.

5.2.0. Les questionnaires à choix multiples (Q.C.M.)

Un bon nombre des critiques dont les tests font ~~souvent l'objet~~ sont en fait dirigées contre les questionnaires à choix multiples auxquels ils sont quelquefois hâtivement assimilés. Il est vrai que cette forme d'items est la plus spectaculaire et la plus ouvertement en opposition avec les moyens traditionnels d'évaluation, qui accordent à la production une part prépondérante, pour ne pas dire exclusive. Il est vrai aussi que la plupart des tests existants tombent dans l'excès inverse, et ont tendance pour des raisons de commodité pratique à abuser des Q.C.M..

Les critiques les plus couramment formulées à l'encontre de ce type d'items sont de plusieurs sortes.

5.2.1. Q.C.M. et production.

En premier lieu, on leur reproche de ne pas faire appel à la production active. Cette objection est justifiée. Mais il faut répéter que cette forme de questions, qui met en jeu les capacités de reconnaissance des sujets, n'est destinée à tester que les skills de compréhension, non ceux de production.

Encore convient-il de nuancer cette affirmation par deux remarques :

- ainsi qu'il a été dit plus haut, la corrélation qu'on peut observer entre un test à choix multiples et un test parallèle à réponses construites est suffisamment élevée pour qu'on puisse considérer comme possible d'atteindre indirectement, grâce à des questions à choix multiples, au moins un aspect important des skills de production.

- il existe des tests à choix multiples qui ont été conçus spécialement pour tester certains éléments des skills de production, et qui peuvent donncer des résultats intéressants, quoique partiels et insuffisants : c'est le cas par exemple du test de prononciation "papier et crayon" de Lado[1].

Il faut remarquer aussi que les Q.C.M. permettent de tester indirectement certains appects des skills de production qu'il serait impossible d'atteindre autrement : ils contraignent en effet les sujets à un choix précis, là où une question à réponse ouverte leur permettrait d'esquiver (consciemment ou non) la difficulté en fournissant une réponse différente, quoique également acceptable.

.../...

(1) : R. LADO : ouvrage cité pp. 95-104

Exemple : avec un item à réponse ouverte, il est pratiquement impossible de vérifier si un élève est capable de reconnaître comme acceptable (à plus forte raison s'il est capable de produire) un énoncé aussi usuel que (a) "il part demain". Il aura toujours en effet la possibilité d'exprimer la même idée(ou presque) sous la forme (b) "il partira demain" ou (c) "il va partir demain", à moins que le constructeur ne fasse intervenir dans ses consignes des interdictions peu naturelles et exprimées métalinguistiquement. Il n'existe sans doute pas de situation dans laquelle l'énoncé (a), qui est pourtant probablement le plus fréquent, soit le seul possible, à l'exclusion de (b) ou de (c), que privilégient la plupart des méthodes d'enseignement existantes et que les élèves utilisent donc en général plus volontiers. S'il ne permet pas de vérifier l'utilisation spontanée de cette structure, un Q.C.M. pourra au moins permettre de voir, non seulement si elle est comprise (ce qui ne présente aucune difficulté pour peu que le mot "demain" soit connu), mais si elle est reconnue comme acceptable. Par exemple :

Tu ne pourras pas voir Pierre, ...

A. il est parti demain.

B. il vient de partir demain.

C. il part demain.

D. il est en train de partir demain.

E. -

Il s'agit ici d'un type particulier de Q.C.M., où la réponse correcte n'est pas forcément fournie et dont la consigne générale est quelque chose comme : "choisissez la réponse correcte et mettez une croix dans la case correspondant à la bonne réponse. Si aucune des réponses n'est correcte, mettez une croix dans la case E". Dans notre exemple, le sujet qui pense que seules seraient acceptables les réponses "il partira demain" ou "il va partir demain" (que le constructeur aura bien entendu pris soin d'exclure de la liste des options) donnera la réponse E, et le professeur saura en tirer les conséquences. Ce type de Q.C.M. qui nécessite de la part du sujet une réflexion sur la langue, ne peut être utilisé qu'à partir d'un certain niveau d'apprentissage. Il est vrai qu'il n'est guère utile tant que la progression ne comporte que peu de cas de variantes comme celles de notre exemple, dont le contenu sémantique est à peu près identique et qui peuvent se trouver pratiquement dans les mêmes contextes.

Un énoncé comme "il part demain" peut être proposé comme option dans d'autres types de Q.C.M. :

- type vrai/faux. Par exemple :

Consigne. Lisez chacune des phrases suivantes. Si elle est correcte, mettez une croix dans la case "vrai". Sinon, mettez une croix dans la case "faux".

Items

1. Il part aujourd'hui. (vrai)
2. Il part hier. (faux)
3. Il part demain. (vrai) etc.

- Type d'items à plusieurs bonnes réponses possibles. Par exemple :

Consigne. Vous allez lire x séries de 3 phrases de sens voisin. Dans chaque série, vous direz combien il y a de phrases correctes (0, 1, 2 ou 3).

Items

1. A. Il partira demain.
   B. Il va partir demain.
   C. Il part demain. (Réponse : 3)
2. A. Il est parti il y a une heure.
   B. Il vient de partir il y a une heure.
   C. Il part il y a une heure. (Réponse : 2)
3. A. Que Pierre pense ?
   B. Que Pierre pense-t-il ?
   C. Quoi Pense Pierre ? (Réponse : 0)

Ce dernier type d'items, qui est à employer avec les mêmes réserves que le type à bonne réponse non nécessairement proposée, présente de plus l'inconvénient que les réponses érronées des sujets sont susceptibles de plusieurs interprétations différentes.

5.2.2. Q.C.M. et facilité.

On reproche souvent aux questions à choix multiples d'être trop faciles. Il est vrai qu'on obtient, à contenu égal, des scores plus élevés avec des choix multiples qu'avec des réponses ouvertes. Mais :

- un score plus élevé à un questionnaire à choix multiples qu'à un questionnaire à réponse ouverte n'implique pas forcément que les premiers sont plus faciles : il faut en effet tenir compte du rôle du hasard et de la divination (cf. ci-dessous 5.2.3. et 5.3.4.) et au besoin le compenser dans une telle comparaison.

- même si l'on fait la part de la divination dans les Q.C.M., il reste que les connaissances passives auxquelles ils font appel ne peu-

vent être plus restreintes que les connaissances actives qu'exigent les questions à réponse construite, et qu'elles sont presque toujours plus étendues : on est toujours capable de reconnaître ce qu'on est capable de se rémémorer, alors que l'inverse n'est pas vrai.

- les items à chaix multiples peuvent atteindre un très haut degré de difficulté, et il est même aisé de tomber dans l'excès en ce domaine.

- un indice de difficulté élevé est rarement la preuve de la valeur d'un item ; un tel indice prouverait au contraire que l'item est mal adapté au niveau des sujets. Par contre, s'il est vrai que dans de nombreux types de tests un item réussi par un pourcentage élevé de sujets peut être lui aussi considéré comme mal adapté et pauvre en signification, tel n'est pas le cas pour un test de progrès, où un item qui s'avère facile apporte malgré tout au professeur des renseignements utiles.

- quoi qu'il en soit, il n'est pas pertinent de poser le problème de cette comparaison en termes de difficulté relative, mais en termes de validité, car ces deux types de questions ne mesurent pas la même chose.

5.2.3. Q.C.M. et hasard.

L'un des grands griefs qu'on adresse aux Q.C.M. est qu'ils laissent au hasard une part trop importante. Il est exact que le hasard peut jouer un rôle non négligeable dans le chaix des réponses, mais ce rôle peut être circonscrit avec précision et réduit à des proportions raisonnables :

Il faut d'abord remarquer que ce rôle du hasard est dissymétrique : un sujet qui ne connaît pas la réponse correcte peut la trouver quand même si la chance le favorise, tandis qu'un sujet malchanceux qui connaît la bonne réponse n'en choisira pas pour autant une mauvaise (ou s'il le fait, ce n'est pas par hasard).

Le rôle du hasard est d'autant plus important que le nombre de choix possibles est plus restreint. Pour les questions à deux options (type oui/non ou vrai/faux), la probabilité de réussite, pour un sujet qui choisit sa réponse d'une façon purement aléatoire, sera de 0,5 (ou 50%) ; pour un choix entre 4 options, cette probabilité tombera à 0,25 (ou 25%).

L'importance du rôle du hasard est également réduite si l'on augmente le nombre des items. Pour une épreuve de 5 items du type vrai/faux, la probabilité de fournir par hasard de bonnes réponses à toutes les

.../...

questions est de $1/2^5 = 1/32$. Pour un test de 10 items du même type, cette probabilité sera de $1/2^{10}$, c'est-à-dire 1/1024, et s'il s'agit d'un test de 20 items, elle ne sera plus que de $1/2^{20} = 1/1048\ 576$, soit moins d'une chance sur un million. Le calcul des probabilités permet d'ailleurs de se rendre compte qu'à nombre égal d'options, une augmentation du nombre des items diminue la probabilité d'obtenir une note égale ou supérieure à l'espérance mathématique(1).

Le tableau ci-dessous indique les probabilités approximatives, exprimées en pourcentage, d'obtenir, dans un test à 2 choix, une note égale ou supérieure à l'espérance mathématique, selon qu'il s'agit d'un test de 5, 10 ou 20 items, en supposant que les réponses sont données d'une façon purement aléatoire.

Nombre d'items \ Note au moins égale à :	5/10 ou 10/20	3/5/ ou 6/10 ou 12/20	7/10 ou 14/20	4/5 ou 8/10 ou 16/20	9/10 ou 18/20	5/5 ou 10/10 ou 20/20
5	-	50		19		3
10	63	38	17	5,5	1	0,01
20	61	26	6	0,6	0,02	0,0001

Probabilités (en pourcentage) d'atteindre ou dépasser, par des moyens aléatoires, une note x dans un test à 2 choix de 5, 10 20 items. (à droite du trait gras, le rôle du hasard peut être considéré comme négligeable).

(1) : L'espérance mathématique E (x) d'une variable aléatoire x (ici, le score obtenu si les réponses sont données au hasard) est la valeur vers laquelle tendrait la moyenne des x si l'épreuve était répétée un grand nombre de fois. Pour un nombre k d'options, la probabilité p de trouver par hasard la bonne réponse à un item sera : $p = \frac{1}{k}$

Si n est le nombre d'items, l'espérance mathématique de la variable x (nombre de bonnes réponses) sera :

$$E(x) = np = \frac{n}{k}$$

Pour k = 2 et n = 20, E (x) = 10, ce qui signifie que si on administre un test de 20 items, du type vrai/faux à un grand nombre de sujets répondant d'une façon purement aléatoire, la moyenne des scores obtenus sera de 10 sur 20. Pour k = 5 et n = 20, E (x) = 4.

Ce tableau établi d'après le triangle arithmétique de Pascal[1], montre que la probabilité d'obtenir un score élevé d'une manière purement aléatoire diminue rapidement quand le nombre d'items augmente. On voit par exemple qu'avec 5 items un sujet totalement incompétent a de réelles chances d'obtenir une note égale ou supérieure à 4/5 ; avec 10 items, la probabilité d'obtenir une note équivalente (8/10) est encore non négligeable ; la note de 16/20 devient déjà hautement improbable dans de telles conditions avec un test de 20 items.

Encore ces calculs ne sont-ils établis ici qu'avec des tests très courts, et avec le nombre d'options minimal. Le tableau suivant, qui donne (toujours en pourcentage) les probabilités d'atteindre par des moyens aléatoires, dans un test de 10 items (donc encore très bref) à 2, 3, 4 ou 5 options, tous les scores possibles, montre qu'un nombre suffisant d'options diminue considérablement ces probabilités.

Nombre d'options. \ Note au moins égale à : E (x)	1	2	3	4	5	6	7	8	9	10
2 — 5	99,9	99	95	83	63	38	17	5,5	1	0,01
3 — 3,33	98	90	70	44	21	7,6	2	0,4	0,04	0,002
4 — 2,5	94	76	47	22	8	2	0,3	0,04	0,003	0,0001
5 — 2	89	62	32	12	3,3	0,6	0,09	0,008	0,0004	0,00001

Probabilités (en pourcentage) d'atteindre ou dépasser, par des moyens aléatoires, une note x (sur 10) dans un test de 10 items à 2,3,4 ou 5 options (à droite du trait gras, le rôle du hasard peut être considéré comme négligeable).

.../...

(1) : Voir Charles MULLER - Initiation à la statistique linguistique - Larousse, Paris, 1968, p. 29 - Sur l'espérance mathématique, voir p.48

La combinaison de ces deux tableaux (et leur extension à des nombres plus élevés) permet de tirer quelques conclusions sur le rôle du hasard dans les questionnaires à choix multiples :

- les questionnaires à 2 options (type vrai/faux) paraissent trop soumis aux variations aléatoires pour avoir une grande fidélité, à moins qu'ils ne comportent un nombre trés élevé d'items.

- les tests dont le nombre d'items est trop bas tombent dans le même défaut, que le nombre nécessairement limité des choix (il est difficile de trouver 4 distracteurs efficaces pour une bonne réponse) ne peut pas toujours compenser. Un test de classe peut difficilement comporter moins de 30 items. Les tests de contrôle et de niveau doivent en compter beaucoup plus.

- quels que soient le nombre d'options et le nombre d'items, le rôle du hasard ne peut être totalement éliminé. Dans les valeurs basses, proches de l'espérance mathématique, les scores sont peu significatifs : à un test de 100 items à 4 choix, on ne peut guère tirer de conclusions à comparer par exemple un score de 35 à un score de 45. Dans les valeurs élevées, rien ne permet de déterminer combien de réponses justes ont été données au hasard, et la proportion de ces réponses peut varier sensiblement d'un sujet à un autre.

- Cependant, si le produit Nombre d'options/Nombre d'items est suffisamment élevé, on dispose alors, nettement au-dessus de l'espérance mathématique, d'un éventail de scores possibles assez étendu pour rendre faible la probabilité qu'un nombre important de réponses justes aient pu être fournis par le seul jeu du hasard.

- ces précautions sont en tout cas efficaces et moins arbitraires que celles qui consistent à pénaliser les mauvaises réponses selon la formule parfois utilisée :

$$\text{Note définitive} = \text{nombre de rép. justes} - \frac{\text{Nombre de rép. erronées}}{(\text{nombre d'options}) - 1}$$

dans laquelle l'absence de réponse n'est pas considérée comme une réponse erronée. Ce procédé n'a d'autre effet que de pénaliser les sujets impressionnables ou scrupuleux, qui préféreront s'abstenir plutôt que de fournir une réponse dont l'exactitude ne leur est pas absolument certaine.

5.2.4. Q.C.M. et divination : choix des distracteurs.

Ces considérations sur le hasard sont naturellement théoriques. Dans la pratique, il est rare que les choses se passent de la manière

.../...

qu'on vient de supposer pour définir mathématiquement le rôle du hasard proprement dit : en fait, un sujet qui ignore la réponse à une question à choix multiples répond rarement d'une façon purement aléatoire, il essaie de deviner, parmi les réponses proposées, laquelle est la bonne.

On peut par conséquent reprocher aux Q.C.M. une diminution de validité due à l'intervention d'une dimension étrangère à celle que l'on cherche à tester : ils favoriseraient les sujets capables d'induire la bonne réponse à partir d'autre chose. A cette objection s'opposent deux types d'arguments :

Le premier est que rien ne permet de penser que cette faculté de divination est vraiment étrangère à ce que l'on sait des mécanismes de la compréhension. Il n'est pas impossible qu'un locuteur natif, en présence d'un élément linguistique qu'il ne connaît pas, procède pour le comprendre d'une façon analogue, par un choix intuitif entre plusieurs hypothèses que le contexte rend plus ou moins plausibles. Il y a probablement dans les skills de compréhension un aspect de créativité qu'il serait inopportun d'éliminer complètement d'une évaluation de la compréhension : il n'est pas certain qu'il faille s'en tenir aux "connaissances solidement acquises", à l'exclusion de ce qui n'a pas été appris (la créativité en compréhension cosiste précisément en la capacité de comprendre des énoncés qui n'ont encore jamais été entendus) ou de ce qui l'a été d'une manière incertaine. Il est vrai qu'il est probablement préférable de tester cette créativité directement par des items appropriés[1] que d'une façon aussi indirecte, aussi indéterminée et aussi artificielle (les distracteurs ne sauraient être comparés à un contexte situationnel).

En second lieu, il faut souligner que la divination dans le choix des réponses peut être rendu plus difficile par le constructeur du test, au moment de l'élaboration des distracteurs. Le rôle de ces derniers est précisément d'attirer sur eux le choix des sujets qui ne connaissent pas d'emblée la bonne réponse, sans tromper ceux qui la connaissent.

Pour satisfaire la première exigence, un distracteur doit avoir l'air de constituer une réponse plausible pour un sujet qui en est réduit à deviner. Un distracteur trop visiblement erroné n'attire personne, manque son but et devient donc inutile, diminuant ainsi le nombre des choix réels. Bien entendu, il y a là un problème d'adaptation au niveau des sujets, car un distracteur efficace pour des sujets d'un niveau donné peut cesser complètement de l'être à un niveau plus avancé. Les interférences avec la langue maternelle ou une autre langue étrangère peuvent fournir d'excellents distracteurs, variables selon les groupes linguistiques. La linguistique contrastive, ou plus simplement les relevés de fautes, donnent pour cela d'utiles indications.

Mais les distracteurs ne doivent pas être pour autant des traquenards destinés à piéger les sujets qui connaissent la bonne réponse : un distracteur trop efficace aura seulement prouvé que l'item était trop difficile ; un distracteur choisi surtout par les meilleurs sujets[1] révèlera que la question était ambigüe ou mal posée. Naturellement, il faudra veiller à ce que ces distracteurs constituent des réponses réellement inacceptables et que leur choix soit indiscutablement exclu par des locuteurs natifs.

Enfin, il est préférable d'éliminer dans le mesure du possible tout artefact de la présentation même des distracteurs : par exemple, la place qu'occupe la bonne réponse dans la série des choix doit être vraiment aléatoire, et n'obéir à aucune règle visible dans la succession des items (éviter non seulement des séries comme "abcdabcd" ou "dcbadcba", mais encore des règles moins visibles comme "jamais deux fois de suite la bonne réponse à la même place" ou "la bonne réponse doit figurer autant de fois sous a, b, c et d dans l'ensemble du test".)

De même, si dans un item à quatre options, l'une des réponses est nettement différente des 3 autres, soit par sa longueur, soit par tout autre élément formel, elle peut attirer sur elle l'attention des sujets, qui risquent ainsi d'avoir des raisons non-linguistiques de la choisir ou de l'exclure. Ou encore, si tous les distracteurs sont construits à partir de la réponse correcte par simple substitution d'une variante chaque fois différente, un étudiant qui ignore la bonne réponse mais possède un certain sens logique peut facilement la reconstituer. Ainsi par exemple, devant un item comme celui-ci :

Il a donné de l'argent à son fils ? - Oui, ...

a. Il lui en a donné
b. Il le lui a donné
c. Il leur en a donné

où "lui" et "en" se trouvent chacun deux fois, tandis que "le" et "leur" ne sont mentionnés qu'une seule fois, on peut facilement déduire que "lui" et "en" font partie de la bonne réponse, et que celle-ci par conséquent est a. Pour que cet inconvénient soit supprimé, il suffirait

.../...

(1) : ce que pourra révéler une analyse des items (voir ci-dessous 6.4.)

soit d'ajouter une 4ème option qui assurerait la symétrie et interdirait un tel raisonnement:

d. il le leur a donné

(type de distracteur qu'on hésite à proposer sous prétexte qu'il comporte une double erreur), soit d'inverser la dissymétrie pour faire tomber dans le piège le sujet qui ne se fierait qu'à des raisonnements de ce genre, ce qui après tout est bien le but des distracteurs, c'est-à-dire dans notre exemple :

a. il lui en a donné
b. il le lui a donné
c. il le leur a donné

(ici, c'est le distracteur b qu'on amène à choisir).
ou bien, en utilisant un distracteur inacceptable dans n'importe quel contexte :

a. il l'en a donné
b. il lui en a donné
c. il le lui a donné
d. il l'a donné

(où la fréquence de le/l', qui apparaît 3 fois, incite à opter pour toute autre réponse que la bonne réponse b)[1].

Pour ce qui est des items sur passage[2], le choix des distracteurs doit également être fait de telle sorte que la bonne réponse ne puisse être trouvée sans que le passage ait été compris : il faut éviter les distracteurs que le simple bon sens, ou des connaissances extérieures au texte lui-même pourraient permettre de déceler.

Exemple tiré de l'épreuve de compréhension écrite d'un test américain[3] : deux items sur la compréhension d'un même texte écrit :

- L'auteur prétend que les élèves intelligents et travailleurs

.../...

(1) Voir sur cette question l'article de Daniel HIRST. Somme logical defects in objective language tests (or How to get good marks by cheating). Les Langues Modernes, 66ème année, n°4, 1972.

(2) Les items sur passage sont des items portant sur la compréhension d'un texte oral ou écrit, par opposition aux items discrets, qui sont isolés.

(3) MLA - Coopérative Foreign Language Tests. French.Form MA. Educational Testing Service, Princeton, 1963.

qui ne réussissent pas aux examens, manquent, le plus souvent, de

A chance
B facilités
C paresse
D mémoire

- Les professeurs savent bien que les bons élèves

A font quelquefois moins bien aux examens qu'ils ne mériteraient
B réussissent toujours aux examens
C méritent toujours ce qu'ils reçoivent aux examens
D s'inquiètent inutilement au moment des examens

Il n'est pas nécessaire de reproduire le passage qui sert de stimulus primaire pour se rendre compte que (à moins que l'auteur du passage ait le goût du paradoxe, ce qui n'est pas le cas) les seules réponses possibles au premier item sont A et D, et la seule possible au second est A. Il est vrai que pour s'en rendre compte, il faut avoir compris la signification des stems et des options, ce qui suppose des connaissances linguistiques réelles. Mais alors à quoi sert le passage ?

Les items sur passage doivent également éviter que la réponse correcte ne soit que la reprise textuelle d'un extrait du passage, surtout si les distracteurs, eux, n'en reprennent pas certains termes : il suffirait de lire sans comprendre pour déceler la bonne réponse et les distracteurs seraient sans effet.

5.2.5. Q.C.M. et contamination.

Autre objection adressée aux Q.C.M. : les distracteurs auraient sur l'apprentissage un effet négatif, puisqu'ils proposent à l'élève des réponses erronées, qui risquent d'être ainsi, soit acquises alors que l'élève n'y aurait peut-être pas pensé, soit renforcées si elles correspondent à une erreur qu'il aurait tendance à commettre spontanément.

Cette objection, peu applicable aux tests de contrôle et de niveau, que leur caractère isolé et leur place chronologique excluent du processus d'apprentissage, est à prendre en considération dans le cas des tests de progrès, qui s'intégrent, eux, à l'apprentissage, dont ils constituent une étape et sur lequel ils influent.

En fait, aucune recherche systématique ne semble avoir abouti à des résultats permettant de savoir si un distracteur peut influer négativement sur l'apprentissage. Le problème ne se pose d'ailleurs

.../...

que pour certains types de distracteurs, ceux qui proposent des associations syntagmatiques inacceptables ou des "barbarismes".

Un item comme celui-ci :

Avez-vous fait une bonne pêche aujourd'hui ?

A. Non merci, je préfère les poires.
B. Mais oui, j'ai bien fait de prendre ce papier.
C. C'était mieux que la semaine dernière.(1)

ne présente aucun risque de contamination, car les deux distracteurs A et B ne sont des "fautes" que parce qu'ils proposent à la question "stem" des réponses inadéquates. Mais ils constituent des séquences parfaitement conformes à la norme du français.

Il n'en est pas de même d'items tel que :

- C'est un village abandonné depuis longtemps, ...

A. on n'y a personne jamais vu
B. on n'y a jamais personne vu
C. on n'y a jamais vu personne
D. on n'y a personne vu jamais(2)

- Vous avez déjà fini de déjeuner ? - Non, ...

A. je n'ai pas commencé même
B. je n'ai commencé même pas
C. je n'ai commencé pas même
D. je n'ai même pas commencé(2)

où les distracteurs (A, B, D pour le premier ; A, B, C pour le second) sont des séquences inacceptables en français ; ni surtout d'items comme ceux-ci, extraits d'un test de classement standardisé canadien(3) :

- Il ignorait les ....

1. déteaux
2. détaux
3. détailes
4. détails
5. détailles

.../...

(1) Rebecca M. VALETTE - ouvrage cité p. 116

(2) Epreuve de performance grammaticale pour instituteurs yougoslaves, non publié.

(3) Test Laval. Formule A. Test de classement. Français langue seconde. Presses de l'Université Laval, Québec, 1971.

- Je n'oublierai pas les soins que j'ai ... de vous
  1. reçus
  2. recevu
  3. recevus
  4. reçues
  5. reçu

- Qu'on .... agir ainsi, cela me révolte.
  1. peut
  2. peuve
  3. pouve
  4. pourrait
  5. puisse

- Ce procédé a rapport à la chimie.
  C'est un procédé .....
  1. chimeux
  2. chimal
  3. chimique
  4. chimable
  5. chimible

- Cet homme est ridicule.
  C'est un être .....
  1. risable
  2. risible
  3. riseux
  4. risal
  5. risique

Il est probable que ces distracteurs, que les constructeurs de tests standardisés de contrôle et de niveau évitent habituellement, seraient intuitivement écartés d'un test de classe par la majorité des professeurs. Sans tomber dans le purisme un peu superstitieux qui fait dire à certains pédagogues qu'une forme erronée même présentée une seule fois risque d'être plus facilement mémorisée que la forme correcte (métaphore de la mauvaise herbe), on peut effectivement s'interroger sur l'effet que de tels distracteurs peuvent produire :

a) sur l'apprentissage (s'il s'agit d'un test de classe) dans la mesure où les formes correctes (reçu, puisse) sont aberrantes dans le système verbal du français et ne peuvent être que mémorisées isolément : si cette mise en mémoire n'est pas acquise au moment de l'évaluation, elle peut être compromise par la perception de formes analogiques comme

.../...

*recevu ou *peuve, qui ont l'avantage d'être plus conformes à l'ensemble du système (voir/vu = recevoir/*recevu ; ils viennent / qu'il vienne = ils peuvent / *qu'il peuve).

b) sur l'évaluation elle-même (quels que soient la forme et l'objet du test) dans la mesure où certains sujets, qui auraient peut-être spontanément produit la forme correcte s'ils avaient été libres de la construire, risquent d'opter au moment de la passation pour une forme erronée plus logique.

Il ne faut probablement pas exagérer les dangers de contamination des distracteurs sur l'apprentissage : ce n'est pas en les traitant par le mépris ni en feignant de les ignorer qu'on réduira les erreurs, et on peut considérer avec A. Beck que les Q.C.M. ont au contraire une influence bénéfique en ce qu"ils aident les élèves à prendre conscience des formes correctes en les mettant dans l'obligation de choisir, et cette prise de conscience est un chaînon indispensable entre l'automatisation initiale des structures et leur transfert dans l'activité langagière réelle"[1]. Au moins à partir d'un certain niveau d'apprentissage, une telle prise de conscience est en effet nécessaire, et la systématisation de cette prise de conscience doit tôt ou tard se substituer à la pédagogie asptisée des programmes skinnériens , d'où toute erreur est bannie.

Il reste que ce type de distracteurs est à utiliser avec précaution dans un test de progrès : le constructeur doit se limiter strictement aux fautes réellement commises par les élèves ; il est préférable que dans ce cas la bonne réponse soit toujours présente parmi les distracteurs (en écartant donc le type d'items mentionné en 5.2.1.) ; la prise de conscience doit avoir lieu chez l'élève au moment de la correction.

5.2.6. Q.C.M. et économie d'emploi.

Au niveau de la construction du test, les Q.C.M. sont peu économiques, car ils exigent une longue préparation. Outre les problèmes de validité que pose n'importe quelle épreuve de compréhension, le choix des distracteurs ne peut se faire au hasard, ni même d'une façon purement intuitive ; certains items à choix multiples, qui peuvent paraître

.../...

(1) Antoine BECK. Les tests de niveau à l'entrée en classe de seconde. Cahiers d'Allemand n° 4. Didier, janvier 1973

parfaitement valables au premier abord, se révèlent à l'expérience mal adaptés à la population et non valides, par la faute de leurs distracteurs.

Par contre, les Q.C.M. sont extrêmement économiques au niveau de la correction : même si l'on néglige les possibilités de correction mécaniques, qui sont hors de portée de la grande majorité des professeurs, les Q.C.M. offrent l'avantage de pouvoir être corrigés (à la main) très rapidement.

Il suffit au correcteur de prendre une feuille de réponse identique à celles qui, polycopiées, sont distribuées aux élèves (figure 1 ci-après), et d'y pratiquer, à l'aide d'une poinçonneuse ou d'une lame de rasoir, des trous à l'emplacement où devraient se trouver les bonnes réponses (figure 2). La simple superposition de cette grille de correction (ou clé) à une feuille de réponse remplie par un élève (figure 3) fera apparaître au premier coup d'oeil les réponses exactes, que le correcteur n'aura plus qu'à compter (figure 4). Il aura pris, avant cette superposition, la précaution de vérifier qu'il n'y a pas de réponses multiples (dont l'interdiction doit être rappelée aux élèves avant chaque passation) et qu'en cas de modification de réponse par un sujet au cours de la passation, la réponse primitive annulée par le sujet a été bien effacée ; si tel n'est pas le cas, il est bon, avant de superposer la grille de correction sur la feuille de réponse, de barrer bien visiblement les réponses multiples, ainsi que les réponses annulées mal effacées. Autre précaution à prendre sur la grille de correction : perforer, outre les emplacements des bonnes réponses, un ou deux points de repère (par exemple le numéro du premier et du dernier items), pour prévenir tout décalage au moment de la superposition[1].

.../...

(1) Rebecca M. VALETTE - ouvrage cité - pages 11-15

NOM : Note :

Prénom :

	A	B	C	D	E
1.	( )	( )	( )	( )	( )
2.	( )	( )	( )	( )	( )
3.	( )	( )	( )	( )	( )
4.	( )	( )	( )	( )	( )
5.	( )	( )	( )	( )	( )
6.	( )	( )	( )	( )	( )
7.	( )	( )	( )	( )	( )
8.	( )	( )	( )	( )	( )
9.	( )	( )	( )	( )	( )
10.	( )	( )	( )	( )	( )
11.	( )	( )	( )	( )	( )
12.	( )	( )	( )	( )	( )
13.	( )	( )	( )	( )	( )
14.	( )	( )	( )	( )	( )
15.	( )	( )	( )	( )	( )
16.	( )	( )	( )	( )	( )
17.	( )	( )	( )	( )	( )
18.	( )	( )	( )	( )	( )
19.	( )	( )	( )	( )	( )
20.	( )	( )	( )	( )	( )

1. Feuille de réponse.

NOM : Note :

Prénom :

	A	B	C	D	E
	( )	( )	( )	( )	( )
2.	( )	( )	( )	( )	( )
3.	( )	( )	( )	( )	( )
4.	( )	( )	( )	( )	( )
5.	( )	( )	( )	( )	( )
6.	( )	( )	( )	( )	( )
7.	( )	( )	( )	( )	( )
8.	( )	( )	( )	( )	( )
9.	( )	( )	( )	( )	( )
10.	( )	( )	( )	( )	( )
11.	( )	( )	( )	( )	( )
12.	( )	( )	( )	( )	( )
13.	( )	( )	( )	( )	( )
14.	( )	( )	( )	( )	( )
15.	( )	( )	( )	( )	( )
16.	( )	( )	( )	( )	( )
17.	( )	( )	( )	( )	( )
18.	( )	( )	( )	( )	( )
19.	( )	( )	( )	( )	( )
	( )	( )	( )	( )	( )

2. Grille de correction.

NOM : Y Note:

Prénom : Z

	A	B	C	D	E
1.	( )	( )	(x)	( )	( )
2.	(x)	( )	( )	( )	( )
3.	( )	( )	( )	( )	(x)
4.	( )	( )	(x)	( )	( )
5.	( )	(x)	( )	( )	( )
6.	( )	(x)	( )	( )	( )
7.	( )	( )	( )	(x)	( )
8.	(x)	( )	( )	( )	( )
9.	( )	( )	(x)	( )	( )
10.	(x)	( )	( )	( )	( )
11.	( )	( )	( )	(x)	( )
12.	(x)	( )	( )	( )	( )
13.	( )	(x)	( )	( )	( )
14.	(x)	( )	( )	( )	( )
15.	( )	( )	( )	( )	(x)
16.	( )	( )	( )	(x)	( )
17.	( )	( )	(x)	( )	( )
18.	( )	(x)	( )	( )	( )
19.	( )	( )	(x)	( )	( )
20.	( )	( )	( )	( )	(x)

3. Feuille de réponse d'élève.

NOM : Note : 15

Prénom :

	A	B	C	D	E
1.	( )	( )	(x)	( )	( )
2.	(x)	( )	( )	( )	( )
3.	( )	( )	( )	( )	(x)
4.	( )	( )	(x)	( )	( )
5.	( )	( )	( )	( )	( )
6.	( )	(x)	( )	( )	( )
7.	( )	( )	( )	(x)	( )
8.	(x)	( )	( )	( )	( )
9.	( )	( )	(x)	( )	( )
10.	( )	( )	( )	( )	( )
11.	( )	( )	( )	(x)	( )
12.	(x)	( )	( )	( )	( )
13.	( )	(x)	( )	( )	( )
14.	(x)	( )	( )	( )	( )
15.	( )	( )	( )	( )	(x)
16.	( )	( )	( )	( )	( )
17.	( )	( )	(x)	( )	( )
18.	( )	( )	( )	( )	( )
19.	( )	( )	( )	( )	( )
20.	( )	( )	( )	( )	(x)

4. Superposition de la grille.

Du point de vue de l'économie d'emploi, il est donc clair que, pour un test de compréhension, un Q.C.M. sera d'autant moins coûteux que le nombre de sujets testés sera plus nombreux : le temps de préparation est en effet le même quel que soit le nombre de sujets, tandis que le temps de correction est proportionnel à ce nombre. Pour un test de classe, fabriqué par le professeur pour ses élèves, l'investissement en temps que représente la construction d'un Q.C.M. n'est donc pas très rentable, à moins que le même test puisse être utilisé plusieurs fois (avec des classes parallèles ou pendant plusieurs années successives), contrairement à ce qui se passe pour un test standardisé, où un Q.C.M. est toujours économique. Le constructeur d'un test de classe aura donc quelquefois avantage à éviter cette forme de questionnaires, si des items à réponse ouverte sont possibles. Dans de nombreux cas cependant, elle est la seule possible, ou la seule valide, ou même la plus économique malgré tout, par exemple :

- les items de discrimination de sons ou d'intonations, où une réponse ouverte n'est guère concevable ;

- les items où la bonne réponse qu'on cherche à provoquer risque d'être remplacée par une autre réponse également acceptable, mais jugée moins intéressante par le constructeur du test (cf. 5.2.1.)

- les items de type vrai/faux, dont la préparation ne nécessite pas de recherche de distracteurs.

- les questionnaires dans lesquels les mêmes options peuvent être conservées pour toute une série d'items, ce qui les rend particulièrement économiques pour le constructeur. Cette formule est particulièrement intéressante pour les tests de classe, dont un sous-test entier peut être consacré à la vérification de l'assimilation d'une règle particulière qui vient de faire l'objet d'une unité de cours. On pourra l'appliquer facilement aux variantes distributionnelles d'un morphème grammatical (formes de l'article, des démonstratifs, des possessifs, des relatifs, etc.), aux oppositions de morphèmes en distribution complémentaire (opposition à/de/ø après les verbes "opérateurs", opposition avoir/être dans les formes verbales composées, opposition le/un/du/ø, opposition aussi/autant/si/tant, etc.)

- les items de compréhension lexicale, pour lesquels il serait peu valide d'exiger des sujets une réponse ouverte (qui serait presque forcément métalinguistique : recherche de synonymes, définitions, etc.).

.../...

5.2.7. Validité des Q.C.M.

On peut adresser aux Q.C.M. plusieurs sortes d'objections concernant leur validité.

a. Au niveau du choix des questions : aux difficultés de représentativité de l'échantillon que rencontre n'importe quelle épreuve, s'ajoute le risque que les questions soient choisies par le constructeur en fonction de leur commodité de rédaction (comment poser la question, quels distracteurs trouver) qu'en fonction de leur représentativité. Ce risque, qui existe aussi pour les questions à réponse ouverte mais brève, est effectivement plus grand encore pour les Q.C.M., qui exigent en outre des distracteurs efficaces, souvent difficiles à trouver. Le professeur doit en être conscient, et se souvenir que tout ne peut pas être testé par un Q.C.M.

b. Au niveau de la rédaction des questions : lorsqu'il s'agit par exemple de tester la compréhension d'un passage, les items, qui ne sont là que pour provoquer les réponses des élèves, ne doivent pas, eux, poser de problèmes de compréhension. Or il est impossible d'être certain à l'avance que tel sera le cas : le constructeur du test doit s'efforcer de poser ses questions de telle manière que tous les sujets les comprennent sans difficulté, mais c'est lui qui en est le juge et il peut se tromper. Une réponse erronée pourra être due, non à l'incompréhension du stimulus primaire (le passage), mais à celle du stimulus secondaire (l'item) ou des options proposées.

c. Au niveau de la passation, où toute une série de facteurs parasites peuvent biaiser les résultats : vitesse de lecture si le test est en temps limité (la mesure de cette vitesse pouvant être considérée comme valide s'il s'agit du stimulus primaire, non s'il s'agit du stimulus secondaire) ; capacité d'attention (il faut répondre au bon endroit, et les risques de se tromper de ligne ou de case sont multipliés si la feuille de réponses est séparée du livret de test) ; bref, la fameuse "aptitude à passer des tests" pourrait bien jouer un rôle plus important dans un Q.C.M. que dans toute autre forme de test. Ces risques ne doivent pas être exagérés, et diminuent à mesure que les élèves sont entraînés à cette technique.

d. La difficulté la plus épineuse à laquelle se heurte la validité des Q.C.M. est en définitive celle que pose la mesure de la compréhension proprement dite, avec toutes les implications sémantiques que ce terme comporte.

- Comment sait-on que quelqu'un a compris ? Il n'y a pas, en fait,

.../...

de critère de compréhension, dans la mesure où, si on peut prouver que quelqu'un n'a pas compris un énoncé donné, on ne peut jamais être certain qu'il l'a compris : ce qu'on peut établir dans le meilleur des cas, c'est que rien ne prouve l'incompréhension et que tout se passe comme si la compréhension était effective. C'est qu'en fait, même lorsqu'il s'agit de la langue maternelle, une interprétation erronée d'un énoncé se perpétue tant qu'elle fonctionne d'une manière satisfaisante, c'est-à-dire tant qu'elle n'est pas infirmée par une variation contextuelle ou situationnelle suffisante pour que la première interprétation se révèle comme erronée.

- La compréhension d'un énoncé englobe celle de ses présupposés. Il peut s'agir d'une implicitation discursive : certaines ambiguïtés de phrases isolées sont en effet levées au niveau du discours. Un énoncé comme "mon livre n'est pas rouge" peut être interprêté comme "ce livre rouge n'est pas à moi", ou "ce livre bleu m'appartient" ; "il est huit heures" peut signifier "allez-vous-en", ou "Pierre n'est pas encore rentré". Ou bien il peut s'agir de présupposition linguistique, la seule pour laquelle Oswald Ducrot[1] réserve le nom de présupposition : l'énoncé "Pierre se doute que Jacques va venir" présuppose l'information "Il est vrai que Jacques viendra", présentée comme allant tellement de soi qu'elle n'a pas besoin d'être explicitée par un acte particulier d'affirmation ; l'énoncé "Pierre s'imagine que Jacques va venir" laisse entendre de la même façon que Jacques ne viendra pas. S'il est relativement facile de tester la compréhension des présupposés linguistiques (toutes les occurrences des verbes "se douter" et "s'imaginer" comportent la même présupposition), il n'en est pas de même de l'implicitation discursive dont la compréhension constitue pour le constructeur de test (comme pour l'enseignant d'ailleurs) une difficulté considérable.

- Si l'on veut tester la compréhension d'un texte, il est impossible de la tester entièrement : les items devront-ils porter globalement sur l'ensemble du texte, ou sur ses différentes parties, ou encore sur chaque phrase prise isolément ? Si on pose des questions sur les éléments du texte, on laisse dans l'ombre leurs relations ; si on teste la compréhension de ces relations, on ne peut le faire que d'une manière approximative et très incomplète. Le test constitue comme un filtre qui,

.../...

---

(1) Oswald DUCROT : Dire et ne pas dire. Principes de sémantique linguistique. Hermann, Paris, 1972, p. 22.

sur l'ensemble du texte, ne peut être que grossier, et ne peut être fin que localement. Les tests de compréhension d'un texte ne sont généralement qu'un compromis entre les deux termes de cette alternative.

- Sauf peut-être dans les tout débuts de l'apprentissage, la compréhension d'une langue étrangère ne s'effectue pas d'une manière discrète, par tout ou rien : plus discontinue pour les débutants, dont on peut dire qu'ils comprennent ou qu'ils ne comprennent pas tel énoncé, elle devient de plus en plus continue à mesure que le niveau de l'apprentissage s'élève, que la compétence des sujets s'étend et que les énoncés qui leur sont soumis se diversifient et se rapprochent des énoncés authentiques. Or le système des Q.C.M. est par nécessité simplificateur : les items sont conçus de telle sorte que la réponse doive être considérée comme parfaitement acceptable ou complètement érronée. Et la multiplicité des questions ne compense pas les effets de cette simplification.

Malgré ces limites dans l'évaluation de la compréhension, le Q.C.M. reste néanmoins un instrument de mesure plus valide que d'autres procédés comme la paraphrase, qui pose le délicat problème des invariants (le problème du "dire autrement") qui implique qu'on dit autre chose, d'où la question de savoir quels éléments sémantiques doivent rester invariants quand on passe de l'énoncé primitif à sa paraphrase) et qui ramène d'ailleurs pratiquement au Q.C.M., puisqu'il est peu réaliste d'exiger des sujets qu'ils construisent eux-mêmes des paraphrases ; ou le résumé de texte qui se heurte d'une façon plus cruciale encore à la même difficulté des invariants (peut-on déterminer avec certitude quels éléments du texte et quelles relations doivent être conservées dans le résumé, et lesquels peuvent être éliminés sans inconvénient ?) et qui de plus contraint l'examinateur à renoncer à prétendre atteindre l'objectivité de notation[1].

Les tests de compréhension de la langue orale posent en outre le problème spécifique du rôle de la <u>mémoire</u> dans la compréhension auditive. Dans les premiers stades de l'apprentissage, seule la mémoire à court terme, la mémoire immédiate, entre en jeu : la mémoire à long terme, celle qui permet la rétention d'un énoncé long ou de plusieurs énoncés successifs, n'intervient qu'à un stade avancé. Il serait donc faux d'affirmer,que, dans tout test de compréhension auditive, le rôle de la mémoire doive être éliminée ou tout au moins minimisé : ceci n'est vrai que pour les débutants ou les élèves des niveaux intermédiaires, dont en effet la mémoire à long terme n'est pas encore développée normalement. Pour eux, un test constitué par un texte oral trop long, entendu une seule fois, risque d'être peu valide. A ce stade d'apprentissage, l'élève peut fort bien avoir

.../...

(1) L'essentiel de ce qui précède dans ce paragraphe d est inspiré d'un séminaire inédit de François BRESSON (Paris, B.E.L.C., mars 1973).

compris chacun des énoncés à mesure qu'il les a entendus, et être incapable de retenir ce qui précède le dernier énoncé entendu. Une réponse erronée à une question portant sur un énoncé entendu depuis trop longtemps n'est donc pas une preuve d'incompréhension de l'énoncé et le test manque son objectif.

Par contre, à un stade plus avancé, la mémoire à long terme devient un élément important de la compréhension auditive. Encore faut-il tenir compte de son véritable rôle qui, comme l'a établi F. Bresson, n'est pas de stocker l'information à la manière d'un ordinateur, selon l'illusion couramment admise, mais de l'organiser et de la reconstruire. C'est la méconnaissance de ce rôle de la mémoire qui a pu conduire à des erreurs méthodologiques comme celle qui consiste à fonder l'apprentissage d'une langue étrangère sur la mémorisation littérale (le "par coeur"), qui n'est pas transférable par elle-même. Pour qu'un test de compréhension d'un texte oral soit valide, il faut donc qu'il permette de vérifier si la mémoire a bien joué son rôle de transformation et si, comme c'est le cas dans la langue maternelle, c'est bien le contenu sémantique de l'énoncé qui a été retenu et non sa forme : dans le Q.C.M., la bonne réponse devra donc être conçue (tout comme, pour des raisons différentes, dans un test de compréhension d'un texte écrit : cf. ci-dessus 5.2.4.) de telle manière qu'elle ne reproduise pas les termes mêmes du texte.

Le problème du canal à utiliser pour présenter les options d'un Q.C.M. sur la compréhension d'un texte oral se pose en termes différents. Si les réponses sont proposées par écrit, bien que cette forme de test ne soit pas vraiment hybride à proprement parler (cf. 5.1. Remarque I), on peut considérer que cette présentation s'éloigne à l'excès, d'une situation authentique de compréhension. Mais si elles sont proposées oralement, le rôle de la mémoire devient ambigu : une liste relativement longue (4 ou 5) de réponses parallèles, qu'il faut retenir pour pouvoir y faire un choix, est-elle vraiment plus proche d'une situation naturelle de compréhension qu'une liste écrite ? Il faut alors trouver des palliatifs. comme la réduction du nombre des options (avec les inconvénients que cela comporte cf. 5.2.3. et 5.2.4.) ou la répétition de leur écoute (avec le risque de faire oublier le stimulus primaire qui, lui, n'aura pas été répété puisque c'est sa compréhension qu'on teste, et dont l'écoute par les sujets sera séparée du moment de leur réponse par l'audition de nombreux énoncés).

.../...

Sans doute faut-il, dans une succession de tests de classe, faire varier les procédés, afin que les variables parasites ne soient pas toujours les mêmes et que l'évaluation de la compréhension ne soit pas toujours liée à une même situation artificielle de testing. Sans doute aussi les solutions seront-elles différentes selon qu'il s'agit de la compréhension d'un texte oral fabriqué pour les besoins de la cause, ou d'un extrait authentique de langue parlée, qu'il est indispensable de proposer à des étudiants qui sont déjà parvenus à un certain niveau, afin de tester leur capacité de compréhension dans une situation proche d'une situation authentique de communication.

5.3.0. Les épreuves de production.

La technique des Q.C.M. n'est guère applicable qu'aux épreuves de "compréhension" et n'est pratiquement d'aucun secours pour les épreuves de production (cf. ci-dessus 5.2.1.). C'est dans cette catégorie d'épreuves qu'il est le plus difficile d'approcher de l'objectivité de cotation.

5.3.1. Les épreuves de production orale.

La production orale, qui met en jeu la skill d'élocution, est d'autant plus difficile à tester qu'elle est difficile à définir, car il s'agit d'un savoir-faire complexe. David P. Harris[1] y distingue plusieurs composantes :

1. la prononciation (éléments segmentaux, accent et intonation)
2. la grammaire
3. le vocabulaire
4. la facilité de parole (fluency)
5. sans doute la compréhension, dans la mesure où l'expression orale en est difficilement isolable.

C'est en fait avec la prononciation que, faute d'un consensus sur ce qui signifie la "bonne" prononciation d'une langue étrangère, le testing (tout comme l'enseignement) se trouve confronté aux plus grandes difficultés : peut-on se contenter de ne demander au sujet que d'être compréhensible ? ou bien faut-il exiger de lui une prononciation conforme à celle des Français natifs ? Dans le premier cas, par qui le sujet doit-il être compris ? Il n'est pas du tout certain que deux

.../...

(1) David P. HARRIS. Testing English as a Second Language. Mc Graw-Hill Book Company, London, 1969, chap. 8.

Français natifs différents trouveront également compréhensibles les énoncés d'un locuteur étranger ; encore moins s'il s'agit d'un Français natif et d'un professeur de la même nationalité que le sujet. Dans le second cas, comment définir une norme de la prononciation du français, et comment déterminer objectivement dans quelle mesure la prononciation du sujet s'en rapproche ?

Pour les tests standardisés, Harris préconise, afin d'atteindre un degré tolérable de fidélité dans la technique la plus courante d'évaluation de la production orale, l'entretien dirigé (scored interview), d'utiliser une échelle d'évaluation accompagnée d'instructions claires, précises et s'excluant mutuellement pour chaque degré, d'entraîner sérieusement et longuement les examinateurs à leur tâche, de mêler les jugements d'au moins deux examinateurs, et de respecter un certain nombre de conditions de procédure.

Voici l'échelle qu'il propose comme exemple pour l'évaluation de la prononciation :

5. Peu de traces d'accent étranger
4. Toujours intelligible, malgré l'existence d'un accent défini
3. Des problèmes de prononciation qui exigent une attention soutenue et conduisent parfois au malentendu
2. Très difficile à comprendre à cause des problème de prononciation ; on doit souvent lui demander de répéter
1. Problèmes de prononciation si graves que le discours est pratiquement inintelligible.

Des échelles semblables sont proposées pour les autres composantes, telle celle destinée à évaluer la facilité de parole :

5. Parle aussi couramment qu'un locuteur natif
4. La vitesse semble légèrement affectée par les problèmes linguistiques
3. La vitesse et l'aisance sont assez fortement affectées par les problèmes linguistiques
2. Habituellement hésitant ; souvent forcé au silence par ses lacunes linguistiques
1. Discours si haché et fragmentaire que la conversation est pratiquement impossible.

Pour les tests de classe, bien que l'entretien dirigé reste un moyen irremplaçable d'évaluation de la production orale parce qu'il constitue le procédé le plus naturel et le plus proche des situations authentiques

.../...

de communication, un certain nombre de réserves s'imposent :

- l'entretien dirigé est très peu économique : il nécessite une préparation sérieuse si l'on veut parvenir à la fois à mettre l'élève dans les conditions psychologiques optimales, à trouver des thèmes qui lui permettent de s'exprimer et de faire montre de ses capacités linguistiques, à varier les questions selon chaque sujet afin d'éviter les réponses préparées, à varier aussi la forme de la conversation pour éviter qu'elle se borne systématiquement à la formule "question du professeur/réponse de l'élève" et à donner à l'entretien l'apparence du naturel. De plus, le temps de la passation (nécessairement individuelle) doit être suffisamment long pour que le test soit valide : Harris l'estime à 10-15 minutes au minimum. Les horaires généralement restreints dont disposent les professeurs de langues excluent donc la possibilité d'entretiens dirigés fréquents : tout au plus un en début d'année scolaire et un en fin d'année pour évaluer les progrès accomplis.

- si les précautions à prendre pour harmoniser les jugements des différents examinateurs d'un test standardisé sont inutiles dans un test de classe, celui-ci court en revanche le risque d'une faible fidélité (effet de stéréotypie, cf. ci-dessus 1.1.).

- les échelles de Harris, qui sont faites pour couvrir tous les niveaux possibles, ne sont guère utilisables pour une classe normale, où il est (heureusement pour le professeur) rare que le niveau des élèves soit assez hétérogène pour que tous les degrés de l'échelle puissent être utilisés, et où (à moins de pratiquer un cours intensif) les élèves n'ont généralement pas le temps de s'élever d'un échelon même au bout d'une année scolaire entière. Ces échelles doivent donc être adaptées au niveau de la classe et, les degrés étant plus proches les uns des autres, il sera plus difficile d'éviter la subjectivité dans l'attribution des notes.

Il existe cependant d'autres techniques d'évaluation de la production orale, tels que les tests papier-crayon de prononciation (cf. 5.2.1.) qui ont l'avantage d'être économiques mais dont la validité est loin d'être prouvée, qui ne peuvent donc être utilisés seuls et qui ne testent de toute manière que la prononciation, et encore certains de ses éléments seulement.

Il y a aussi ce que Harris appelle les échantillons de discours hautement structurés, qui ne peuvent prétendre remplacer l'entretien dirigé, parce que tout à fait éloignés d'une situation authentique de

.../...

communication, mais qui ont sur lui le double avantage d'exiger des sujets une tâche plus uniforme parce que plus contraignante, et de permettre un étalonnage plus fin : la comparaison est ainsi rendue plus facile, non seulement entre les différents élèves, mais surtout entre des performances successives d'un même élève. Il en existe de plusieurs types :

- la répétition de phrases contenant une opposition phonétique dont on juge la réalisation, à l'exclusion des autres éléments. Inconvénient : on juge les capacités d'imitation d'un modèle plutôt que l'expression spontanée.

- la lecture de phrases écrites dans les mêmes conditions. Inconvénient : la lecture à voix haute constitue une capacité spécifique, surtout en matière d'intonation.

- la transformation de phrases selon une règle grammaticale précise (temps, nombre, etc.). Inconvénient : procédé très limité et très artificiel.

- la construction de phrases appropriées à des situations spécifiques. Exemple, emprunté à Harris : "Vous avez téléphoné à votre amie Marie, mais sa mère vous répond qu'elle n'est pas là. Demandez-lui de dire à Marie qu'elle vous appelle quand elle rentrera". Inconvénient : les consignes sont compliquées et risquent de poser des problèmes de compréhension ; les possibilités de ce procédé restent également très limitées.

- la réaction à des stimulus visuels : les sujets sont invités à examiner une série de dessins et à raconter ce qui se passe dans chaque scène représentée en images. Un bon exemple de cette technique est fourni par l'épreuve d'expression orale du Test C.G.M. 62[1], dont les "consignes de correction et de notation" recommandent à l'examinateur de tenir compte de trois éléments complémentaires :

"1. La quantité d'expression que l'étudiant fournit (l'épreuve étant de toutes façons arrêtée au bout de 5 minutes).

2. La correction grammaticale de ces expressions.

3. La qualité du style (emploi de structures complexes, utilisation de pronoms, mots de liaison, etc.)"

et fournissent une classification très élaborée des phrases pouvant être produites par les sujets et un barême de notation nuancée par un "indice d'expression" ("c'est le rapport du nombre d'idées exprimées

.../...

au temps écoulé") et un "indice de correction" ("il se calcule en divisant le nombre d'idées exprimées correctement par le nombre total d'idées énoncées").

Ce test offre les avantages de :

- donner les meilleures garanties possibles d'objectivité dans une épreuve d'expression orale.

- juger à la fois la correction grammaticale, la fluidité verbale et même la qualité stylistique, que D.P. Harris ne signale pas et dont il importe de tenir compte si l'on ne veut pas favoriser excessivement les sujets qui se contentent d'aligner des phrases très simples comportant peu de risques d'erreurs, mais dont l'accumulation est étrangère aux habitudes des locuteurs natifs.

- isoler nettement l'expression de la compréhension, puisque les sujets n'ont à comprendre que les consignes du test (qu'on peut donner dans leur langue maternelle) et qu'on ne tient aucun compte de la façon dont ils ont interprêté les images proposées, qui ne sont qu'un support non pertinent à l'expression.

On peut lui reprocher de ne pas tenir compte des variables lexicale et surtout phonétique, et d'être extrêmement peu économique. Mais c'est peut-être le seul type d'épreuve qui réunisse la plupart des avantages de l'entretien dirigé et des échantillons de discours structurés, en conciliant autant que possible l'expression quasi-spontanée en situation de communication et le maximum de garanties d'objectivité.

5.3.2. Les épreuves de production écrite.

Elles posent des problèmes différents de ceux que suscitent les épreuves de production orale, et en tout cas moins nombreux :

- la passation pouvant être collective, ces épreuves sont beaucoup plus économiques que les épreuves de production orale (autre facteur d'économie : la conservation des réponses des sujets ne nécessite pas d'appareillage coûteux d'enregistrement).

- l'orthographe est infiniment plus facile à tester que la prononciation, les problèmes de compréhensibilité et de norme ne se posant pratiquement pas.

.../...

(1) G. MIALARET, C. MALANDAIN : Test C.G.M. 62. Pour apprécier le niveau des connaissances linguistiques. Français. Réalisation du C.R.E.D.I.F. Paris, Didier, 1964, Epreuve n°2.

- la "facilité d'écriture", beaucoup plus inégale en langue maternelle que la facilité de parole, est une composante beaucoup moins pertinente quand il s'agit d'une langue seconde.

Si on entend par écriture les capacités spécifiques d'expression par ce canal de communication, et qu'on en exclue toute l'aire commune aux deux canaux, il faut remarquer que la référence aux natifs, indispensable dès qu'il s'agit de la langue orale, n'est plus pertinente : en effet, si dans les débuts de l'apprentissage d'une langue seconde, l'écriture n'est guère autre chose qu'un codage différent du matériau linguistique oral, elle devient par la suite un savoir-faire spécifique et aussi complexe que la parole, mais dont les composantes sont des habiletés que de nombreux individus n'atteignent jamais même dans leur langue maternelle.

Ces composantes peuvent être regroupées de plusieurs façons différentes par une analyse factorielle (1). David P. Harris (2) par exemple propose de distinguer :

1. Contenu : substence de l'écriture (les idées exprimées)
2. Forme : organisation du contenu.
3. Grammaire : emploi des formes et structures grammaticales.
4. Style : choix des structures et éléments lexicaux pour donner un ton particulier à l'écriture.
5. Mécanismes : usage des conventions graphiques (alphabet, orthographe, ponctuation, capitalisation).

Certaines de ces composantes sont faciles à tester objectivement : c'est le cas de la grammaire et surtout de l'orthographe (élément de l'écriture pour lequel la dictée traditionnelle, peu valide et très arbitraire dans sa cotation, est remplacée avantageusement par la dictée à trous : le texte entendu est reproduit sur la feuille de réponse, à l'exception des éléments graphiques, mots ou graphèmes, que l'on désire précisément tester).

Cependant, tout comme l'entretien dirigé pour la production orale, la composition reste irremplaçable quand on veut évaluer les capacités d'expression écrite d'un sujet. Ses inconvénients, qui sont comparables à ceux de l'entretien dirigé (faible fidélité en raison du manque d'uniformité des réponses et de la subjectivité de notation, possibilité

.../...

(1) cf. ci-dessus 1.1.1.

(2) David P. HARRIS, ouvrage cité, chap. 7.

donnée aux sujets d'éluder les difficultés, notation peu économique) sont longuement compensés par les avantages qu'elle présente :

- certains éléments du skill de l'écriture ne peuvent pas être mesurés autrement ("contenu" et "forme" de D.P. Harris).

- c'est le seul type d'épreuve qui puisse se comparer aux situations authentiques d'écriture (correspondance, rédaction d'article ou de rapport) et où l'élève puisse exprimer ses idées personnelles.

- elle est très économique à préparer, sinon à corriger.

La fidélité peut même être augmentée si on prend certaines précautions :

- demander aux sujets plusieurs échantillons courts sur des thèmes différents plutôt qu'un seul long.

- donner des tâches à la portée des sujets, en évitant celles qui exigent un haut degré de créativité ou des connaissances trop spécifiques (dissertations littéraires).

- fournir des consignes claires, précises et complètes.

- décider à l'avance (et en prévenir les élèves) du système de notation et du coefficient qu'on attribuera à chacune des composantes (système sans doute préférable pour un test de classe à celui qui consiste à porter un jugement global sur l'ensemble de la composition).

Voici un bon exemple d'épreuve d'expression écrite, tirée d'un test de contrôle élaboré par l'Association des Ecoles Populaires Supérieures d'Allemagne[1] :

"Cherchons famille allemande pour recevoir notre fils pendant vacances d'été. S'adresser à Monsieur Paysan, 109, av. St-Max, 31 Toulouse.

Ecrivez à M. Paysan en tenant compte de ce qui suit :

31. Dites comment vous avez appris que M. Paysan cherchait une famille.
32. Votre fils (16 ans) apprend le français.
33. Echange possible ?
34. Dates des vacances en Allemagne.
35. Proposez un séjour : dates
36. et logement.
37. Proposez une visite dans votre ville
38. et une excursion dans votre région.
39. Réponse rapide souhaitée.
40. Formule de politesse (Schlussformel)".

.../...

(1) Volkshochschul-Zertifikat französisch prüfungsaufgaben 1971. Ubungsausgabe A 1. Für den kursteilnehmer. Pädagogischen Arbeitsstelle des Deutschen Volkshochschul-Verbandes, Frankfurt, 1971.

5.4. Les composantes linguistiques.

Le troisième des critères de classement énumérés en 5.0., la composante linguistique testée, amène à distinguer de nouvelles classes d'items, si le but de l'épreuve est d'évaluer le savoir-faire spécifique des sujets dans la composante en question. On peut affiner la classification selon que ce critère se combine ou non avec les deux précédents.

C'est ainsi qu'il est difficile d'imaginer des tests de phonétique indépendamment du pôle de communication où on peut placer le sujet : il s'agira soit d'épreuves de production, soit d'épreuves de reconnaissance.

Par contre, on peut construire des sous-tests de grammaire dans lesquels ce critère est peu pertinent, et même dans lesquels la distinction entre les deux canaux de communication joue un rôle restreint (structures communes aux deux codes). Il en est de même pour les sous-tests de vocabulaire.

On trouvera, dans la littérature consacrée au testing des langues, des listes de types d'items de phonétique, de grammaire et de vocabulaire, accompagnées d'exemples nombreux(1).

5.5. Les niveaux de comportement.

Une dernière classification des items est possible ; c'est la classification taxonomique(2) qu'on peut effectuer parallèlement à la taxonomie des niveaux de performance qui constituent les objectifs pédagogiques.

.../...

(1) Robert LADO, ouvrage cité.
- Language Testing Symposium, ouvrage cité.
- David P. HARRIS, ouvrage cité.
- Rebecca M. VALETTE, ouvrage cité.
- Frédéric FRANCOIS et Emmanuel COMPANYS, ouvrage cité.

(2) Une taxonomie est une classification hiérarchisée, qui ordonne les éléments des plus simples aux plus complexes.

La taxonomie des objectifs pédagogiques établie par Benjamin S. Bloom et ses collaborateurs[1], principalement pour l'enseignement des sciences physiques et sociales, de l'histoire et de la littérature, a inspiré à J.B. Carrol[2] une ébauche de taxonomie des items des tests de langue, et surtout à Rebecca M. Valette et Renée S. Disick[3] une taxonomie des objectifs de performance dans l'enseignement d'une langue seconde, qui est directement applicable au testing.

R. Valette et R. Disick distinguent, parmi les comportements des domaines cognitif et psycho-moteur, qu'un professeur de langue peut tenter de faire acquérir à ses élèves ou à ses étudiants, 5 niveaux, allant du plus simple au plus complexe, les comportements de chacun de ces 5 niveaux étant subdivisés en deux composantes, le comportement interne (du sujet en tant que récepteur) et le comportement externe (de production). Cette taxonomie tient donc compte du second des critères de classification énumérés en 5.0. Elle englobe par contre dans les mêmes catégories les comportements qui ne se différencient que par les autres critères (canal de communication et composante linguistique), qui sont proprement linguistiques, donc fondés sur la matière enseignée et non sur le comportement de l'élève. Le tableau ci-dessous, qui résume les différents niveaux de la taxonomie des objectifs de performance dans les domaines cognitif et psycho-moteur, est une traduction de celui que proposent R. Valette et R. Disick.[4]

.../...

---

(1) Benjamin S. BLOOM et al., Taxonomy of Educational Objectives : The Classification of Educational Goals. Handbook I : Cognitive Domain, David McKay, New York, 1956.
Traduction française de Marcel LAVALLEE : Benjamin S. Bloom et ses collaborateurs. Taxonomie des objectifs pédagogiques. Tome 1 : Domaine cognitif, Education Nouvelle, Montréal, 1969.

(2) J.B. CARROL :"The Psychology of Language Testing" in Language Testing Symposium, ouvrage cité : §2 "The Taxonomy of Language Test Tasks".

(3) Rebecca M. Valette, Renée S. DISICK : Modern Language Performance Objectives and Individualization. A Handbook, Harcourt Brace Jovanovich, New York, 1972.

(4) Ouvrage cité, p. 41.

Niveau	Comportement interne	Comportement externe
1. AUTOMATISMES Les performances de l'étudiant sont produites par la mémoire mécanique plutôt que par la compréhension.	PERCEPTION L'étudiant perçoit les différences entre 2 (ou plus) sons, lettres, gestes, et fait entre eux des distinctions.	REPRODUCTION L'étudiant imite le discours, l'écriture, des gestes, des chansons et des proverbes de la langue étrangère.
2. CONNAISSANCE L'étudiant manifeste sa connaissance de faits, de règles et de données en relation avec l'apprentissage de la langue étrangère.	RECONNAISSANCE L'étudiant montre qu'il reconnaît les faits qu'il a appris en répondant à des Q.C.M. ou à des items du type vrai/faux.	RAPPEL L'étudiant manifeste qu'il se rappelle l'information enseignée en répondant à des questions à réponse construite.
3. TRANSFERT L'étudiant utilise ses connaissances dans des situations nouvelles.	RECEPTION L'étudiant comprend des passages oraux ou écrits recombinés ou des citations jamais rencontrées auparavant.	APPLICATION L'étudiant parle ou écrit dans une situation contrôlée, ou participe à des simulations culturelles.
4. COMMUNICATION L'étudiant utilise la langue et la culture étrangères comme des véhicules naturels pour la communication.	COMPREHENSION L'étudiant comprend un message en langue étrangère ou un signal culturel contenant un matériau inconnu dans une situation inconnue.	EXPRESSION L'étudiant utilise la langue étrangère pour exprimer oralement ou par écrit ses idées personnelles. Il utilise les signes gestuels comme partie intégrante de son expression.
5. CRITIQUE L'étudiant analyse ou évalue la langue étrangère, ou effectue à son sujet des recherches personnelles.	ANALYSE L'étudiant décompose la langue ou un passage littéraire en ses éléments stylistiques, thématiques, etc. essentiels. EVALUATION L'étudiant évalue et juge la pertinence et l'efficacité d'un échantillon de discours ou d'un passage littéraire.	SYNTHESE L'étudiant effectue des recherches originales, ou une étude personnelle, ou établit un plan en vue d'un tel projet.

Taxonomie des objectifs de performance dans les domaines cognitifs et psychomoteur (d'après Rebecca Valette et Renée Disick).

.../...

Outre cette taxonomie des comportements des domaines cognitif et psycho-moteur, qui sont centrés sur la matière enseignée, et toujours inspirées par les travaux de Benjamin S. Bloom et de ses collaborateurs, en l'occurrence la Taxonomie des Objectifs Pédagogiques, domaine affectif[1], Rebecca M. Valette et Renée S. Disick ont également établi une taxonomie des objectifs de performance dans le domaine affectif, qui se rapportent aux attitudes et aux sentiments des étudiants. Elles ont également réparti ces comportements affectifs en 5 niveaux, dont chacun est subdivisé en deux sous-catégories, ainsi qu'il apparaît dans le tableau ci-après.[2]

On trouvera dans le même ouvrage des exemples d'items (et de questionnaires d'attitudes) classés selon ces taxonomies.

.../...

(1) David R. KRATWOHL, Benjamin S. BLOOM and Bertram MASIA : Taxonomy of Educational Objectives, Handbook II : Affective Domain, David McKay, New-York, 1964.
Traduction française de Marcel LAVALLEE : Taxonomie des objectifs pédagogiques, tome 2. Domaine affectif, Education Nouvelle, Montréal, 1969.

(2) Rebecca M. VALETTE et Renée S. DISICK, ouvrage cité, p. 48.

Niveau		
1. RECEPTIVITE L'étudiant est ouvert à l'apprentissage d'une langue et d'une culture étrangères.	CONSCIENCE L'étudiant est conscient de l'existence de langues et de cultures autres que les siennes, et du fait qu'il existe entre elles des différences.	ATTENTION L'étudiant prête attention aux informations sur la langue et la culture étrangères, à la fois dans et hors de la classe. Il veille à la préparation soigneuse de ses devoirs.
2. REPONSE L'étudiant réagit (répond) positivement à l'apprentissage de la langue et de la culture étrangères.	TOLERANCE L'étudiant est tolérant à l'égard des différences dans l'expression en langue étrangère et les modes de vie étrangers. Il ne rejette ni ne raille les usages étrangers.	INTERET ET SATISFACTION L'étudiant s'intéresse aux activités se rapportant à l'étude de la langue étrangère, aime les activités qui lui sont proposées et est satisfait d'y participer.
3. APPRECIATION L'étudiant accorde de son plein gré de la valeur aux expériences linguistiques et culturelles.	VALORISATION L'étudiant considère l'étude de la langue et de la culture étrangères comme une valeur et comme quelque chose d'important.	ENGAGEMENT L'étudiant participe volontairement de temps en temps aux activités destinées à améliorer ses capacités linguistiques ou accroître sa connaissance de la langue et de la culture étrangères.
4. INTERIORISATION L'étudiant forme ses propres idées et valeurs basées sur ses expériences de l'apprentissage de la langue étrangère.	CONCEPTUALISATION L'étudiant développe un système personnel de valeurs se rapportant à l'étude de la langue étrangère.	ADHESION L'étudiant consacre la majeure partie de son temps et de son énergie à essayer d'apprendre davantage.
5. CARACTERISATION La langue et la culture étrangères sont devenues partie intégrante de la vie de l'étudiant, au point qu'il est caractérisé par ses activités en ce domaines.	INTEGRATION L'étudiant intégre les valeurs de la langue étrangère dans son système de valeurs personnel.	PROSELYTISME (leadership) L'étudiant joue un rôle essentiel dans la promotion de l'enseignement de la langue étrangère.

Taxonomie des objectifs de performance dans le domaine affectif.

(d'après R. VALETTE et R. DISICK)

.../...

6.0. Evaluation des résultats d'un test.

Les tests standardisés, qui sont des tests se référant à une norme[1] et pour lesquels c'est le classement des sujets qui est en définitive le plus important, nécessitent une pré-expérimentation minutieuse fondée sur un traitement statistique important.

Il n'en est pas de même des tests de classe, qui sont avant tout des tests se référant à un critère[1] et où il est plus important de comparer le résultat obtenu par tel élève au résultat considéré comme le critère d'atteinte de l'objectif, qu'au résultat obtenu par les autres sujets. Les tests de progrès étant diagnostiques au sens étroit du terme, et faits pour une évaluation qualitative plus que quantitative, il n'est pas nécessaire, et il est même dangereux, qu'ils s'entourent d'un appareil statistique compliqué.

Le professeur qui administre à ses élèves des tests ainsi conçus, non seulement peut faire l'économie des calculs complexes qui transforment le score brut obtenu à un test standardisé, en note normalisée, note-centile[2], note standard[3] ou classe d'échelle normalisée[3], mais encore a à se garder de l'illusion que les scores de ses élèves doivent se distribuer selon la loi normale des variables aléatoires, la fameuse courbe en cloche ou courbe de Gauss. Une telle distribution n'a de sens que si deux conditions sont remplies : d'une part que le nombre de sujets soit assez élevé pour que la variable score puisse être assimilée à une variable aléatoire, d'autre part que l'objectif du test soit une sélection. La plupart des tests standardisés remplissent ces conditions et dans le cas

.../...

---

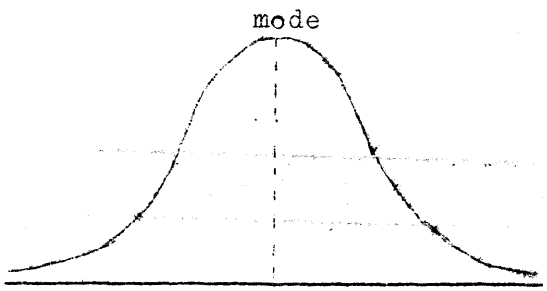
(1) cf. ci-dessus 2.3.

(2) Le centilage étant le système d'étalonnage dans lequel on divise une distribution normale (= de type gaussien) en cent classes d'effectif égal, la note centile indique, par rapport à une population ayant subi le même test, le pourcentage de sujets qui ont obtenu une note plus basse que le sujet considéré. La transformation en note-centile équivaut donc à donner à un sujet son classement dans un groupe d'étalonnage qui comprendrait 100 sujets. Ce système, qui donne un classement grossier et tendant à l'arbitraire dans les valeurs moyennes, tend à être abandonné.

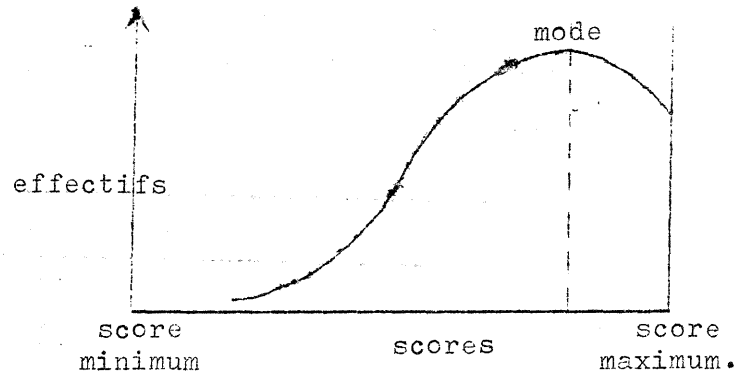
(3) cf. ci-dessous 6.2.

des tests de contrôle, la recherche de la courbe en cloche, avec des sujets ayant en principe suivi le même programme, signifie que ce qu'on mesure c'est la vitesse d'acquisition des élèves, donc en définitive leurs aptitudes.

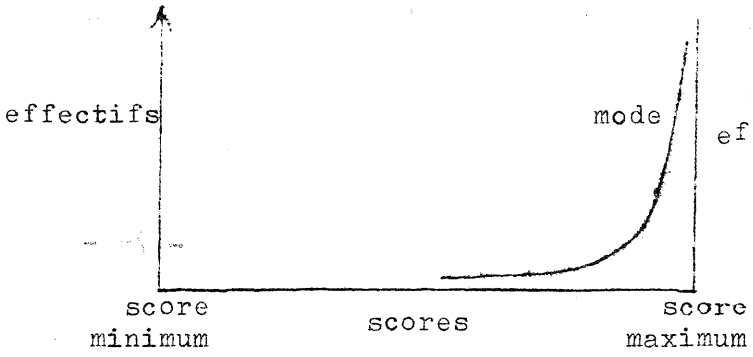
Tel n'est pas le cas des tests de classe, qui sont administrés à un petit nombre d'élèves et qui, sauf exceptions rares, sont les instruments d'une évaluation formative, visant à instruire et non à sélectionner. Ce que recherche le professeur, dans un enseignement individualisé, ce n'est pas de savoir quels sont parmi ses élèves les plus doués et les moins doués, ni de permettre aux premiers de faire la preuve de leur supériorité sur les seconds, mais de faire en sorte que tous ses élèves atteignent un niveau minimum qu'il a fixé comme objectif de performance. En administrant ces tests, il s'efforcera donc au contraire de faire en sorte que, une fois définis les objectifs de performance et les tests construits en fonction de ces objectifs, il obtienne, dans un graphique où sont portés en abscisses les scores possibles et en ordonnées l'effectif de chacun de ces scores (figures ci-dessous), une courbe dissymétrique dont le mode sera déplacé vers la droite, ou même une courbe en J, idéal de l'évaluation formative, mais que l'évaluation sommative s'efforce soigneusement d'éviter[1].



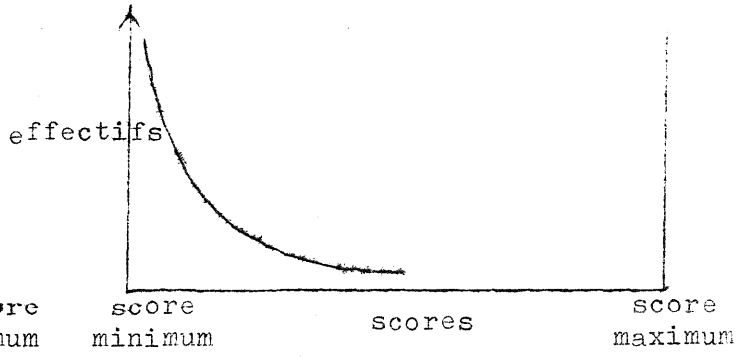
Courbe de Gauss



Courbe dissymétrique



Courbe en J



Courbe en i.

.../...

(1) Voir Gilbert de LANDSHEERE, Evaluation continue et examens. Précis de docimomologie. Labor, Bruxelles, 1971/ Nathan, Paris, 1972, 93 et suivantes. Sur la théorie de l'évaluation formative et la pédagogie de la courbe en J, ibid., p. 186 et suivantes.

(Courbe en J)

- Test très facile pour la population considérée : le mode est la valeur maximum de la variable.
- ou : situation avant l'apprentissage (pré-test)

(Courbe en i)

- test beaucoup trop difficile : le mode est la valeur minimum de la variable.
- ou : situation à la fin de l'apprentissage (post-test)

Le professeur qui administre un test de classe aura cependant quelquefois avantage à calculer :

- les paramètres de position et de dispersion de la variable score.
- le coefficient de corrélation entre deux tests ou entre deux passations du même test.
- le degré de validité relative de chacun des items d'un test par rapport au test pris dans son ensemble.

6.1. Calcul des paramètres de position (moyenne).

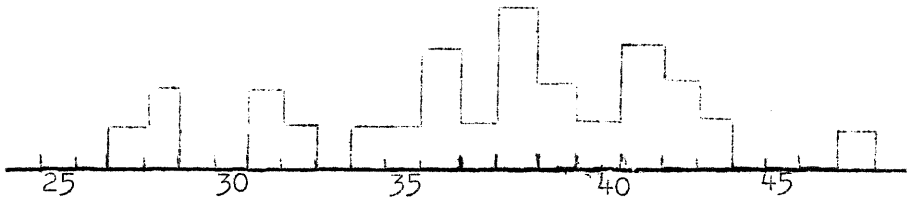
Prenons comme exemple les résultats obtenus à un test de progrès de 50 items administrés en janvier 1973 à un groupe de 26 étudiants de l'I.P.F.E. Ces résultats sont résumés dans le tableau ci-dessous, où la première colonne indique les scores, la deuxième l'effectif (= le nombre) des étudiants ayant obtenu le score considéré, et la troisième colonne les effectifs cumulés (chaque ligne totalisant les effectifs des lignes précédentes).

Scores	Effectifs	Effectifs cumulés
50	0	0
49	0	0
48	0	0
47	1	1
46	0	1
45	0	1
44	0	1
43	1	2
42	2	4
41	3	7
40	1	8
39	2	10
38	4	14
37	1	15
36	3	18
35	1	19
34	1	20
33	0	20
32	1	21
31	2	23
30	0	23
29	0	23
28	2	25
27	1	26
26	0	26

Test I.P.F.E. Tableau 1

.../...

Ces chiffres peuvent être représentés par l'histogramme ci-dessous :



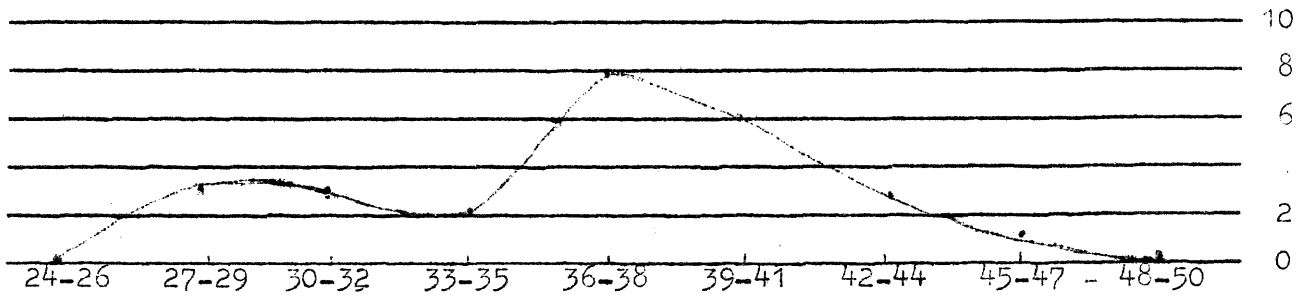
Test I.P.F.E. Histogramme 2

Pour corriger ("lisser") les irrégularités de ce profil, on peut diminuer le nombre de classes en regroupant les scores, par exemple 3 par 3, ce qui produit les effectifs suivants :

Scores	Effectifs	Effectifs cumulés
48 - 50	0	0
45 - 47	1	1
42 - 44	3	4
39 - 41	6	10
36 - 38	8	18
33 - 35	2	20
30 - 32	3	23
27 - 29	3	26
24 - 26	0	26

Test I.P.F.E. Tableau 3

On obtient de la sorte une courbe qui se rapproche davantage d'une courbe gaussienne, mais qui est pourtant bien différente de la courbe en cloche type :



Test I.P.F.E. - Graphique 4

.../...

Les paramètres de position (ou tendances centrales) de la variable x (dont les valeurs sont ici les scores obtenus) sont les suivants[1] :

- la médiane, qui est la valeur centrale de la série ordonnée des scores; dans notre exemple, puisqu'il y a 26 valeurs, c'est celle qui occupe les 13^{e} et 14^{e} rangs, soit 38.

- le mode, qui est la valeur dont la fréquence est le plus élevée ; en l'occurence, c'est la valeur 38 si on se réfère à l'histogramme 2, le groupe de valeurs 36-38 si on se réfère au graphique 4[2].

- la moyenne ($\bar{x}$), qui est le rapport entre la somme des scores x et leur nombre n :

$$\bar{x} = \frac{\Sigma x}{n}$$

et qui sera dans notre exemple :

$$\bar{x} = \frac{958}{26} = 36,85$$

La méthode la plus simple et la plus rapide pour le calcul d'une moyenne, bien que la formule qui l'exprime soit apparemment plus compliquée, est la suivante :

$$\bar{x} = \frac{\Sigma n_k x_k}{n}$$

où x_k représente chacune des valeurs différentes de x (chaque score) et où n_k représente l'effectif correpondant à chaque x_k[4].

.../...

---

(1) Voir Gaston MIALARET et Daniel PHAM. Statistique à l'usage des éducateurs, P.U.F., Paris, 1967, pages 53-57 ; Charles MULLER, ouvrage cité, p. 50.

(2) Le mode n'a de signification que si le nombre de classes n'est pas trop élevé par rapport au nombre des valeurs différentes prises par la variable. Il peut arriver qu'il y ait deux modes (la courbe a alors deux sommets et est dite bimodale). C'est généralement le signe que la population est hétérogène par rapport au caractère considéré.

(3) Le signe Σ sert à marquer la somme : Σx se lit "somme des valeurs prises par x".

(4) L'indice k signifie que les termes de chacun des produits à additionner sont le score et l'effectif situés sur la même ligne du tableau : $x_1 = 47$; $n_1 = 1$; $x_2 = 46$, $n_2 = 0$;.... Donc, $\Sigma n_k x_k$ signifie : $n_1 x_1 + n_2 x_2 + n_3 x_3 + \ldots + n_i x_i$, si le tableau comporte i lignes (ou i valeurs différentes du score).

Scores x_k	Effectifs n_k	$n_k \times k$
47	1	47
46	0	0
45	0	0
44	0	0
43	1	43
42	2	84
41	3	123
40	1	40
39	2	78
38	4	152
37	1	37
36	3	108
35	1	35
34	1	34
33	0	0
32	1	32
31	2	62
30	0	0
29	0	0
28	2	56
27	1	27
	$\Sigma n_k = n = 26$	$\Sigma n_k \times k = 958$

Test I.P.F.E. Tableau 5

Dans une distribution gaussienne, les trois paramètres de position prennent la même valeur. Des différences entre ces paramètres sont l'indice d'une dissymétrie dans la courbe, ce qui est le cas de notre exemple.

Les paramètres de position donnent une idée de la partie du champ de la variable dans laquelle ses valeurs ont tendance à se regrouper. Bien entendu, ils ne fournissent qu'une indication globale sur l'ensemble des scores. La comparaison entre leurs différentes valeurs, dans le cas de deux tests parallèles ou d'un test administré une seconde fois n'apporte de renseignements que sur la progression du groupe de sujets pris dans son ensemble.

6.2. Calcul des paramètres de dispersion (écart type).

Même si on s'en tient à une information globale sur l'ensemble des résultats obtenus, le calcul de la moyenne est insuffisant. La moyenne (ni les autres paramètres de position) ne donne aucune indication sur la façon dont les scores se répartissent : une moyenne pourra être identique avec une dispersion très réduite, c'est-à-dire si les scores sont très groupés autour de la moyenne (cas limite : tous les scores sont égaux) et

.../...

avec une grande dispersion, c'est-à-dire si les scores se répartissent sur un large éventail. Les paramètres de dispersion complètent donc sur l'ensemble des résultats les informations des paramètres de position. Ce sont[1] :

- la variance (Var ou σ^2), qui est la moyenne des carrés des déviations par rapport à la moyenne :

$$\sigma^2 = \frac{\Sigma n_k (x_k - \bar{x})^2}{n}$$

où $\bar{x}$ est la moyenne,
x_k chacune des valeurs de la variable,
$x_k - \bar{x}$ la déviation de chaque x_k par rapport à la moyenne[2],
n le nombre total de scores,
n_k l'effectif correspondant à chaque x_k ;

- et surtout l'écart type (σ), qui est la racine carrée de la variance, et qui est la moyenne des déviations par rapport à la moyenne des scores :

$$\sigma = \sqrt{\frac{\Sigma n_k (x_k - \bar{x})^2}{n}}$$

La moyenne $\bar{x}$ étant généralement un nombre fractionnaire (dans notre exemple 36,85), le calcul des déviations et de leurs carrés risque, par cette méthode, d'être long et difficile. Il vaut mieux utiliser comme origine des déviations un nombre entier, par exemple le nombre entier le plus proche de la moyenne vraie (dans notre exemple 37), qu'on appellera moyenne auxiliaire, et qui permettra de calculer une variance provisoire σ'^2, selon la même formule que celle citée ci-dessus pour σ^2, en remplaçant seulement $\bar{x}$ par $\bar{x}'$:

$$\sigma'^2 = \frac{\Sigma n_k (x_k - \bar{x}')^2}{n}$$

Pour obtenir la variance vraie σ^2 à partir de σ'^2, il suffira d'en soustraire le carré de l'écart entre la moyenne auxiliaire et la moyenne vraie :

$$\sigma^2 = \sigma'^2 - (\bar{x}' - x)^2$$

.../...

---

(1) Voir Charles MULLER, ouvrage cité, p. 52 et suiv. ; G. MIALARET et D. PHAM. ouvrage cité p. 57 et suiv. ; G. de LANDSHEERE, ouvrage cité, p. 96 et suiv.

(2) L'élévation au carré a pour effets d'éliminer les signes négatifs et de donner plus d'importance aux valeurs extrêmes de la déviation.

En pratique, cette correction étant faible (elle ne peut être supérieure à 0,25, carré de 0,5), on peut la négliger sans inconvénient et conserver la variance obtenue à partir de la moyenne auxiliaire.

Dans notre exemple, la façon de calculer la variance et l'écart type est exprimée dans le tableau ci-dessous :

Scores	Effectifs	Déviation par rapport à la moyenne auxiliaire.	Carré des déviations	Produit du carré des déviations par l'effectif
x_K	n_K	$x_K - \bar{x}'$	$(x_K - \bar{x}')^2$	$n_K (x_K - \bar{x}')^2$
47	1	10	100	100
46	0	9	81	0
45	0	8	64	0
44	0	7	49	0
43	1	6	36	36
42	2	5	25	50
41	3	4	16	48
40	1	3	9	9
39	2	2	4	8
38	4	1	1	4
37	1	0	0	0
36	3	-1	1	3
35	1	-2	4	4
34	1	-3	9	9
33	0	-4	16	0
32	1	-5	25	25
31	2	-6	36	72
30	0	-7	49	0
29	0	-8	64	0
28	2	-9	81	162
27	1	-10	100	100
	$\Sigma n_K = n = 26$			$\Sigma n_K (x_K - \bar{x}')^2 = 630$

Variance provisoire : $\sigma'^2 = \frac{\Sigma n_K (x_K - \bar{x}')^2}{n} = \frac{630}{26} = 24,27$

Variance vraie : $\sigma^2 = \sigma'^2 - (\bar{x}' - \bar{x})^2 = 24,27 - (0,15)^2 = 24,25$

Ecart type : $\sigma = \sqrt{24,25} = 4,92$, soit en arrondissant : 5

Test I.P.F.E. Tableau 6

Il existe une méthode simple donnant une approximation satisfaisante de l'écart type, sans passer par l'intermédiaire de la variance, et qui consiste à additionner les scores du sixième supérieur, les scores du sixième inférieur et à diviser la différence entre ces deux sommes par $\frac{n}{2}$ (1)

.../...

(1) D'après Paul B. DIEDERICH. Short-cut Statistics for Teacher-made Tests, Evaluation and Advisory Service Series, n° 5, 2e édition, Educational Testing Service, Princeton, 1964, p. 23.

Dans l'ensemble du test I.P.F.E., $\frac{n}{6} = 4\frac{1}{3}$; ce calcul sera donc basé sur les $4\frac{1}{3}$ meilleurs scores (= les 4 meilleurs + $\frac{1}{3}$ du cinquième) et sur les $4\frac{1}{3}$ plus bas :

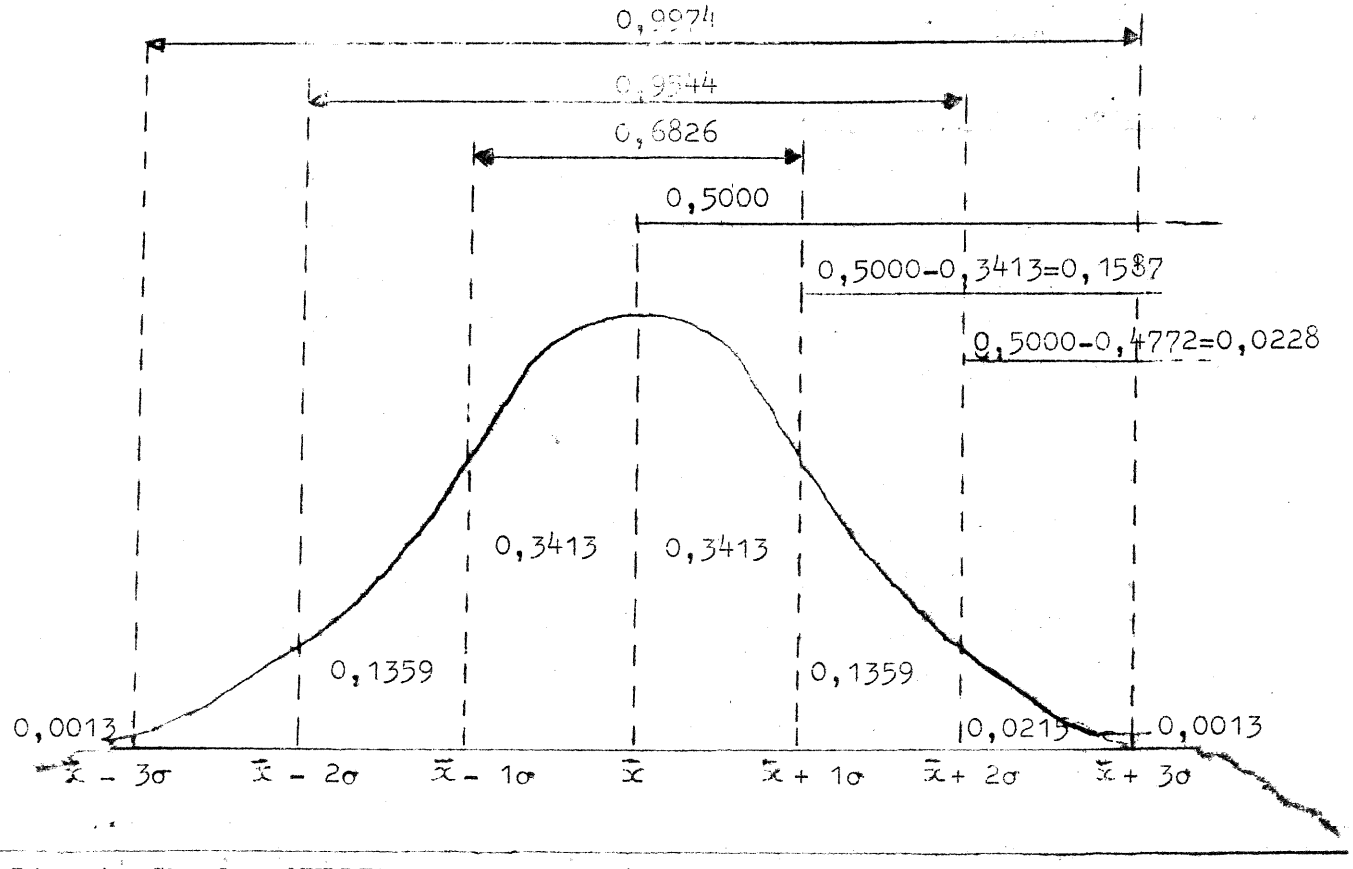
$$\sigma = \frac{(47 + 43 + 42 + 42 + 42 + 13,6) - (27 + 28 + 28 + 31 + 10,3)}{26 : 2}$$

$$= \frac{187,6 - 124,3}{13} = \frac{63,3}{13} = 4,87$$

ce qui constitue en effet une excellente approximation de l'écart type exact 4,92.

L'écart type constitue le complément indispensable de la moyenne quand il s'agit de comparer le résultat obtenu par un sujet aux résultats globaux du groupe auquel il appartient. Sans l'avoir calculé, on ne peut interpréter l'écart d'un score donné par rapport à la moyenne, car c'est l'écart type qui sert d'unité à cet écart.

En effet, la statistique de Laplace-Gauss, dite loi normale, à laquelle obéissent les variables aléatoires et à laquelle peut se ramener la distribution des scores obtenus à un test administré à un très grand nombre de sujets, établit de la façon suivante[1], en fonction de l'écart type, la probabilité pour que la variable x (le score) prenne une valeur comprise entre la moyenne $\bar{x}$ et une valeur quelconque :



(1) D'après Charles MULLER, ouvrage cité, p. 59.

Un nombre figurant sur le graphique ci-dessus, par exemple 0,3413, s'interprête de la façon suivante :

- la probabilité pour une variable aléatoire de prendre une valeur comprise entre la moyenne $\bar{x}$ et la valeur $\bar{x} + 1\sigma$ (ou $\bar{x} - 1\sigma$) est de 0,3413 (la probabilité maximale ou certitude, est égale à 1)

- le pourcentage de sujets obtenant à un test une note comprise entre $\bar{x}$ et $\bar{x} + 1\sigma$ (ou $\bar{x} - 1\sigma$) sera, pourvu que le nombre de sujets soit grand, voisin de 34,13 % (si ce n'est pas le cas, c'est que la variable n'obéit pas à la loi normale).

L'écart type est un élément du calcul de l'écart réduit, qui est la base du système d'étalonnage dit des notes standard ou notes Z dont la formule est :

$$z = \frac{x_k - \bar{x}}{\sigma} \quad (1)$$

C'est également par rapport à l'écart type que se construisent les échelles normalisées dont les plus utilisées sont :

- l'échelle à 5 classes, où les 3 classes centrales sont de l'étendue d'un écart type (les 2 classes extrêmes étant illimitées).

- l'échelle à 9 classes, où les 7 classes centrales sont de l'étendue d'un demi-écart type.

6.3. Notion de corrélation.

Il peut quelquefois être utile au professeur constructeur de tests de classe de vérifier le degré de corrélation qui existe entre deux tests, bien que cette notion soit plutôt utile à l'évaluation sommative (pour un test de classement par exemple, une haute corrélation entre un test long et un test court autorisera à utiliser plutôt le second pour des raisons d'économie).

On dira que la corrélation entre deux tests est élevée si la plupart des sujets obtiennent un score (ou un rang) comparable dans chacune des deux épreuves. La corrélation sera parfaite si chaque sujet obtient le même rang (ou le même écart réduit par rapport à la moyenne) à chacun des deux tests. La corrélation pouvant se calculer, et s'exprimer par un coefficient ou indice variant de +1 à -1, une corrélation parfaite se traduira par un coef-

.../...

(1) Voir Ch. MULLER, ouvrage cité p. 61 et suiv. ; G. MIALARET et D. PHAM, ouvrage cité p. 74 et suiv. ; G. de LANDSHEERE, ouvrage cité, p. 134.

ficient de corrélation de +1. La valeur o du coefficient indiquera une corrélation nulle, les deux variables étant totalement indépendante l'une de l'autre. Un coefficient de -1 marquera une corrélation inverse parfaite.

Les moyens de calculer ce coefficient sont indiqués dans tous les ouvrages de statistique[1]. La corrélation des rangs s'exprime par l'indice de Spearman ρ (rhô), qui se calcule à partir du rang à occuper par chaque sujet dans le premier test et du rang b qu'il occupe dans le second, de la différence d = a - b établie pour chaque sujet, élevée au carré et totalisée, et du nombre n de sujets. La formule en est :

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Le coefficient de Bravais-Pearson indique la corrélation, non plus des rangs, mais des déviations des scores par rapport à la moyenne de chaque épreuve. Si on désigne par x_k le score obtenu par chaque sujet au premier test, dont la moyenne est $\bar{x}$ et l'écart type σ_x, et par y_k le score obtenu au second test, dont la moyenne est $\bar{y}$ et l'écart type σ_y, le coefficient de Bravais-Pearson, représenté par r, sera :

$$r = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{n\sigma_x\sigma_y}$$

Mais il s'agit là de calculs longs et complexes, rarement utiles pour un test de classe. Paul B. Diederich[2] indique un moyen rapide et simple de calculer un indice approximatif de corrélation : il consiste à compter le pourcentage de sujets qui, après classement des résultats obtenus à chaque test, figurent dans la moitié supérieure du groupe pour chacune des deux épreuves. Il est facile de voir que si le classement est exactement le même dans les deux cas, ce pourcentage atteindra son maximum, c'est-à-dire 50% et l'indice sera égal à 1. Si ce pourcentage se rapprochait de 0%, il s'agirait d'une très forte corrélation inverse, la corrélation nulle correspondant à un pourcentage de 25%. Voici le tableau de correspondance entre les pourcentages possibles et l'indice de corrélation r, que propose Paul B. Diederich :

.../...

(1) Voir en particulier G. MIALARET et D. PHAM, ouvrage cité, pages 141 et suiv. ; Ch. MULLER, ouvrage cité, pages 111 et suiv.

(2) Paul B. DIEDERICH, ouvrage cité, p. 34.

%	r	%	r	%	r	%	r	%	r
45	0,95	37	0,69	29	0,25	21	-0,25	13	-0,69
44	0,93	36	0,65	28	0,19	20	-0,31	12	-0,73
43	0,91	35	0,60	27	0,13	19	-0,37	11	-0,77
42	0,88	34	0,55	26	0,07	18	-0,43	10	-0,81
41	0,85	33	0,49	25	0	17	-0,49	9	-0,85
40	0,81	32	0,43	24	-0,07	16	-0,55	8	-0,88
39	0,77	31	0,37	23	-0,13	15	-0,60	7	-0,91
38	0,73	30	0,31	22	-0,19	14	-0,65	6	-0,93

Au lieu de se calculer, la corrélation peut se visualiser par un diagramme (cf. ci-dessous), où chaque point représente un sujet, l'abscisse correspondant au score x qu'il a obtenu au premier test, l'ordonnée au score y qu'il a obtenu au second. Après avoir porté chaque couple de scores (x_k, y_k) sur le diagramme, on obtient un nuage de points, dont l'aspect renseigne immédiatement sur le degré de corrélation entre les deux épreuves : si le nuage de points est nettement groupé autour de la diagonale sud-ouest/nord-est, la corrélation est élevée. Un nuage orienté selon la diagonale nord-ouest/sud-est indiquerait une forte corrélation inverse. Une corrélation faible se traduit par une dispersion des points sur toute la surface du diagramme :

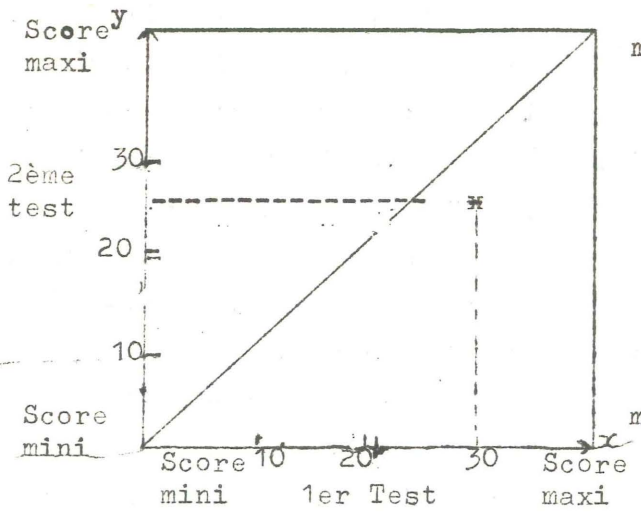
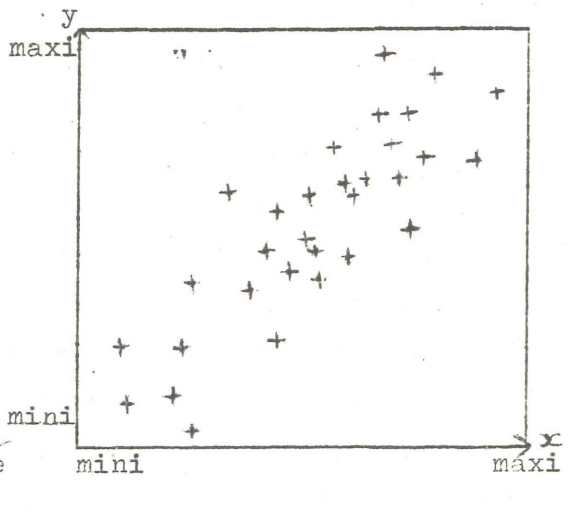


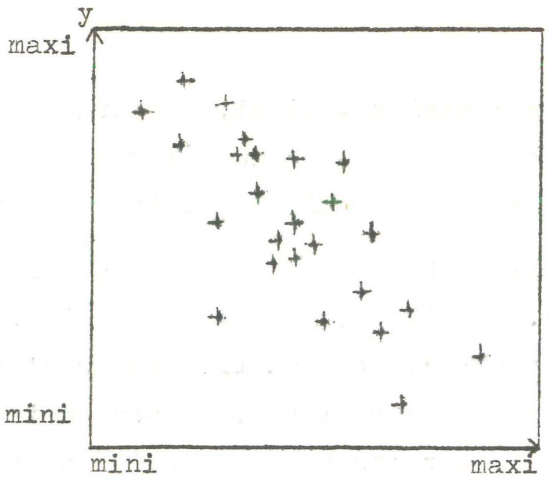
diagramme de corrélation



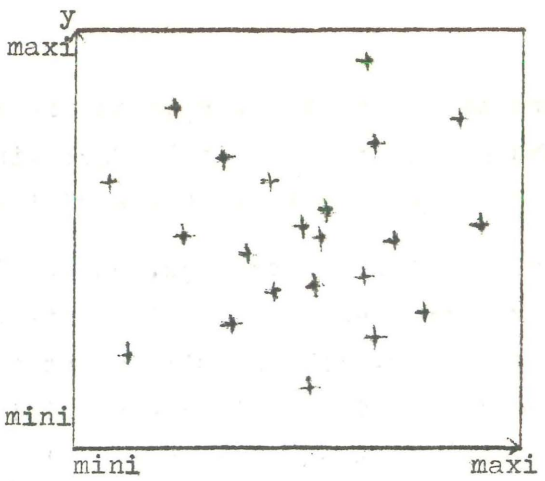
forte corrélation

Exemple : $x = 30, y = 25$

.../...



corrélation inverse



Faible corrélation

Le calcul des coefficients de corrélation peut servir en particulier à la mesure de la fidélité d'un test : test-retest avec le même test, ou avec les deux moitiés (méthode pairs-impairs) d'un même test[1].

Dans ce dernier cas, chaque demi-test étant moins fidèle (parce que plus court) que le test entier, l'indice de corrélation sous-estime la fidélité réelle du test entier. On corrige habituellement cette erreur par la formule suivante, qui dégage de l'indice de corrélation r calculé par la méthode pairs-impairs l'indice de fidélité du test, dont la valeur maximale est également 1 :

$$\text{Indice de fidélité} = \frac{2\,r}{1 + r}$$

Cette corrélation interne d'un test pourra cependant être évaluée d'une manière moins quantitative et plus détaillée par l'analyse des items, qui est en quelque sorte une estimation de la corrélation entre chaque item et l'ensemble du test.

6.4. L'analyse des items.

Indispensable dans l'élaboration d'un test standardisé, l'analyse des items, effectuée au stade de la pré-expérimentation, a pour but essentiel d'évaluer chacun des items composant le test du point de vue de son pouvoir discriminatif, comparé au pouvoir discriminatif du test (ou du sous-test) entier : compte-tenu de l'information fournie par l'ensemble du test sur le niveau respectif des sujets, il s'agit de calculer dans quelle mesure

.../...

(1) : cf. ci-dessus 1.3.

l'information apportée par chaque item y contribue. Ainsi, les items qui contribuent insuffisamment à discriminer les meilleurs sujets des moins bons sont jugés inutiles et sont éliminés de la version définitive du test.

Dans un test de progrès, où le classement des élèves les uns par rapport aux autres est de peu d'importance, ce rôle de l'analyse des items sera secondaire. Elle servira avant tout à permettre au professeur d'établir un diagnostic précis des acquisitions de ses élèves, et accessoirement de vérifier si chacun des items qu'il a construits a une validité comparable à celle du test entier.

Pour procéder à l'analyse des items(1), on divise les copies, classées par ordre de mérite(2), en trois parties égales ; les meilleurs scores constitueront le tiers supérieur (I), les plus bas le tiers inférieur (III) et les scores intermédiaires le tiers moyen (II). Si le nombre total (N) n'est pas divisible par 3, il suffit que les deux tiers extrêmes soient exactement égaux.

Exemple :

N = 37
I = 12 (les 12 meilleurs scores)
II = 13 (les 13 scores intermédiaires)
III = 12 (les 12 scores les plus bas)

Il peut arriver que des scores identiques soient partagés entre deux tiers contigus. Cela n'a pas grande importance et la répartition peut être faite au hasard.

Pour chaque item, on fait le décompte des réponses correctes fournies par chaque tiers et on calcule la différence entre le nombre de réponses

.../...

(1) D'après Elisabeth INGRAM "Item Analysis" in Language Teaching Symposium. ouvrage cité.
Voir aussi Rebecca M. VALETTE, ouvrage cité, pages 37-43. David P. HARRIS, ouvrage cité, chap. 9 ; Gilbert de LANDSHEERE, ouvrage cité, p. 75

(2) Cet ordre est généralement établi en fonction des résultats obtenus au test entier. Cependant, lorsque le test est composé de sous-tests faisant intervenir des skills différents et risquant d'avoir entre eux une faible corrélation, on a intérêt à établir cet ordre en fonction du sous-test spécifique.

correctes fournies par le tiers supérieur et le nombre de réponses correctes fournies par le tiers inférieur ; cette différence, divisée par l'effectif (égal) de ces tiers (8 dans notre exemple), donne l'indice de discrimination de l'item, Le total des réponses correctes fournies par l'ensemble des sujets (les trois tiers réunis), permet, divisé par N, d'établir l'indice de difficulté de l'item.

L'indice de discrimination peut donc varier de 1 (quand le tiers supérieur tout entier a fourni la réponse correcte et que le tiers inférieur tout entier a donné une mauvaise réponse : c'est la discrimination maximale) à théoriquement -1 (quand c'est l'inverse : discrimination négative absolue) ; un indice égal à 0 (le nombre de réponses correctes est le même dans le tiers supérieur et le tiers inférieur) indique une discrimination nulle : un tel item n'a joué aucun rôle dans le classement des sujets.

L'indice de difficulté peut varier de 1 (100 % des sujets ont fourni la bonne réponse : difficulté minimale) à 0 (personne n'a fourni la bonne réponse : difficulté maximale).

Exemple : N = 37

N° de l'item	I	II	III	Total I + II + III	Indice de difficulté: $\frac{I + II + III}{N}$	Différence I - III	Indice de discrimination: $\frac{I - III}{N/3}$
1	12	13	12	37	1,00	0	0
2	12	12	10	34	0,92	2	0,17
3	8	6	6	20	0,54	2	0,17
4	12	6	2	20	0,54	10	0,83
5	6	6	8	20	0,54	-2	-0,17
6	11	9	6	26	0,70	5	0,42
7	12	12	8	32	0,86	4	0,33
8	10	4	2	16	0,45	8	0,66
9	8	2	3	13	0,35	5	0,42
etc							

Cet exemple (imaginaire) montre la différence des informations apportées par l'analyse des items selon qu'il s'agit d'un test de classe ou de la pré-expérimentation d'un test de classement, Pour un test de classement, on éliminerait de la version définitive :

- les items n° 1 et 2, trop faciles et par conséquent trop peu discriminatifs (encore qu'il soit admis dans ce genre de tests que les premiers items soient faciles afin de mettre les sujets en confiance).
- l'item n° 9, trop difficile bien que relativement discriminatif.

.../...

- l'item n° 3, de difficulté moyenne, mais égale pour tous les sujets donc à indice de discrimination faible.

- l'item n° 5, mieux réussi par les sujets les plus faibles que par les meilleurs (indice de discrimination négatif), ce qui prouve son absence de validité, puisqu'il semble mesurer une capacité différente de celle qui est mesurée par le test entier (cette décision présupposant d'ailleurs que l'on considére comme valide le test dans son ensemble).

Seuls seraient conservés les items n° 4, 6 et 8, de difficulté moyenne et à indice de discrimination suffisant, et peut-être l'item n° 7, dont le pouvoir discriminatif est à la limite de l'acceptabilité.

Dans un test de classe, seuls l'item n° 5 et peut-être l'item n° 3 peuvent être considérés comme mauvais (la question peut avoir été mal posée ou comporter un piège trop subtil dans lequel seront tombés surtout les meilleurs sujets). Tous les autres items apportent au professeur des informations utiles sur le niveau de sa classe : il verra que les items 1, 2 et 7 ne font pas difficulté pour l'ensemble de ses élèves, et il saura qu'il a à reprendre, pour tout ou partie de sa classe, ce qui constituait le contenu des autres items.

En fait, c'est donc surtout l'indice de difficulté des items qu'il doit calculer à l'occasion de chaque test. Pour ce calcul, l'analyse complète des items n'est pas nécessaire : il suffit de faire le décompte du nombre total des réponses correctes de chaque item, sans avoir besoin de distinguer les différents tiers. L'analyse complète pourra être effectuée de temps en temps, afin de déceler les items peu valides, mais de toute manière, avec un nombre de sujets aussi réduit, elle ne peut donner que des indications grossières et approximatives.

L'analyse des items peut même (elle doit l'être dans le préexpérimentation d'un test standardisé) être raffinée dans le cas des Q.C.M. par une analyse des distracteurs, qui donne des informations sur l'efficacité et la valeur des distracteurs et permet d'éliminer pour la version définitive du test ceux qui sont inutiles parce que totalement inefficaces, ou non valides parce que trop efficaces chez les meilleurs sujets. Dans le cas d'un test de progrès, le professeur aura tout autant intérêt à relever les distracteurs qui se sont révélés efficaces dans un Q.C.M. construit par lui, qu'à relever les fautes commises dans une épreuve traditionnelle ou ou dans un questionnaire à réponse construite. C'est ce genre d'analyse qui lui permet de mieux ajuster son enseignement et de mieux mesurer, à travers des tests successifs, les progrès accomplis par ses élèves et ceux qu'il leur reste encore à accomplir. C'est bien là en définitive le but des tests de progrès dans la classe.

BUREAU
POUR
L'ENSEIGNEMENT
DE LA
LANGUE
ET DE LA
CIVILISATION
FRANCAISES
A L'ETRANGER

LES TESTS DE LANGUE

BIBLIOGRAPHIE SELECTIVE.

A. Testing des langues étrangères.

Ouvrages fondamentaux.

HARRIS (David P.) : Testing English as a Second Language.
New York, Mc Graw-Hill, 1969.

LADO (Robert) : Language Testing. The Construction and Use of Foreign Language Tests.
London, Longmans, 1961. 389p.

LANGUAGE Testing Symposium. A Psycholinguistic Approach. Ed. by Alan DAVIES.
London, Oxford University Press, 1968. 214p.

VALETTE (Rebecca M.) : Modern Language Testing. A Handbook.
New York, Harcourt, Brace and World, 1967. 200p.

Ouvrages et articles divers.

ACTES du deuxième congrès de la Fédération Internationale des Professeurs de français tenu au Domaine Universitaire de Saint-Martin-d'Hères, Université des Langues et Lettres, Grenoble, du 23 au 29 juillet 1972.
F.I.P.F. Bulletin 6-7, 2ème semestre 1972-1er semestre 1973.

BEARDSMORE (H. Baetens) : "Testing Oral Fluency".
in : Rapport d'activités de l'Institut de Phonétique 1971-1972, n° 6.
Université libre de Bruxelles - Septembre 1972.

BECK (Antoine) : "Les tests dans l'enseignement des langues vivantes"
Cahiers d'Allemand, Revue de linguistique et de pédagogie, n° 1.
Paris, Didier, Septembre 1970.

BECK (Antoine) : "Les tests de niveau à l'entrée en classe de seconde"
Cahiers d'Allemand, n° 4.
Paris, Didier, janvier 1973.

CENTRE INTERNATIONAL D'ETUDES PEDAGOGIQUES - SEVRES.
Evaluation des connaissances et apprentissage du français.
Sèvres : C.I.E.P., 1973 - 238p., 30cm (Dossier n° 48).

FOREIGN Language Testing. Specialised Bibliography 1
London, Centre for Information ou Language Teaching, 1971, 16p.

FRANCOIS (Frédéric) et COMPANYS (Emmanuel) : Les tests de langue.
Paris, B.E.L.C., 1969. 66p.

MOST (R.L.) : "Testing Aural Comprehension"
in : Rapport d'activités de l'Institut de Phonétique 1971-1972, n° 6
Université libre de Bruxelles, septembre 1972.

.../...

PALMER (Adrian S.) : "Testing Communication".
IRAL, X/1, February 1972. pp. 35-46.

PIMSLEUR (Paul) : "Testing Foreign Language Learning" in VALDMAN (Albert) (Ed. by). Trends in Language Teaching.
New York, McGraw-Hill, 1966. pp. 175-214.

SAVARD (Jean-Guy) : Bibliographie analytique de tests de langue/Analytical Bibliography of Language Tests.
Preface de W.F. MACKEY.
Québec, Presses de l'Université Laval, 1969. 372p.

TRUCHOT (Claude) : "Les tests de langue : réévaluation critique".
Les Langues Modernes, 65ème année, n°2, mars-avril 1971. pp. 103-116

UPSHUR (John A.) and FATA (Julia), ed. : Problems in Foreign Language Testing : proceedings of a conference held at the University of Michigan, September 1967. 183p.
Language Learning, spécial issue n° 3, August 1968.

VALETTE (Rebecca M.) : Directions in Foreign Language Testing.
New York, Modern Language Association, 1970.

B. Quelques tests de français langue étrangère.

CAPELLE (Janine et Guy) : Tests "La France en Direct".
Paris, Hachette, 1969.
(Pour niveau débutants : tests de la méthode "La France en Direct").

CENTRE DE RECHERCHE ET D'ETUDE POUR LA DIFFUSION DU FRANCAIS (C.R.E.D.I.F.) : Test T. 69, pour apprécier le niveau des connaissances en langue française, 2ème degré.
Paris, C.R.E.D.I.F., Didier, 1973.
- Livret de l'examinateur, 54p., 20cm.
- Feuilles de test.

CENTRE DE RECHERCHE ET D'ETUDE POUR LA DIFFUSION DU FRANCAIS (C.R.E.D.I.F.) : Test "Bonjour Line".
Paris, Didier, 1963.
(Pour niveau débutants : tests de la méthode "Bonjour Line".

CENTRE DE RECHERCHE ET D'ETUDE POUR LA DIFFUSION DU FRANCAIS (C.R.E.D.I.F.)): Voix et Images de France. Epreuves de contrôle.
Paris, Didier, 1967.
(Pour niveaux débutants : tests de la méthode "V.I.F.").

MIALARET (G.) et MALANDAIN (C.) : Test C.G.M. 62. Pour apprécier le niveau des connaissances linguistique : Français.
Réalisé par le C.R.E.D.I.F.
Paris, Didier, 1964.

MODERN LANGUAGE ASSOCIATION OF AMERICA M.L.A. Cooperative Foreign Language Tests. Forms LA and MA.
Princeton, Educational Testing Service, 1965.
(Pour niveau intermédiaire).

MODERN LANGUAGE ASSOCIATION OF AMERICA M.L.A. Foreign Language Proficiency Tests for Teachers and Advanced Students. Forms HA and HB.
Princeton, Educational Testing Service, 1968.
(Pour niveau avancé).

PIMSLEUR (Paul) : Modern Foreign Language Proficiency Tests. French. Forms A and C.
New York, Harcourt, Brace and World, 1967.
(Pour niveaux débutant et intermédiaire).

.../...

VALETTE (Rebecca M.) : Objective Tests. French I and French II.
Boston, Ginn and Company, 1967.
(Pour niveau secondaire : tests de la méthode "French I and II" de O'Brien, Lafrance, et Brachfeld).

C. Evaluation et docimologie.

BLOOM (Benjamin S.) ; HASTINGS (Thomas J.) ; MADAUS (George F.) : Handbook on formative and summative evaluation of student learning.
New York, London, McGraw,Hill, 1971.

BONBOIR (ANNA) : La docimologie.
Paris, Presses Universitaires de France, 1972 (Coll. Sup. L'éducateur, 196p.).

DE LANDSHEERE (Gilbert) : Evaluation continue et examens. Précis de docimologie.
Bruxelles, Labor, Paris, Nathan, 1972 (Coll. Education 2000). 230p.

PIERON (Henri) : Examens et docimologie.
Paris, Presses Universitaires de France, 1969 (Coll. Sup. le psychologue). 195p.

D. La mesure par les tests.

BONBOIR (Anna) : La méthode des tests en pédagogie.
Paris, Presses Universitaires de France, 1972 (Coll. Sup. L'éducateur). 144p.

DE LANDSHEERE (Gilbert) : Le test de closure, Mesure de la lisibilité, et de la compréhension. Bruxelles, Labor, Paris, Nathan, 1973.
128p., graph.

DE LANDSHEERE (Gilbert) : Les tests de connaissances.
Bruxelles, Editest, 1965.

PICHOT (Pierre) : Les tests mentaux.
Paris, Presses Universitaires de France, 1968 (Que sais-je ? n° 626)
128p.

E. Objectifs pédagogiques.

BLOOM (Benjamin S.) et al. Taxonomy of Educational Objectives :
The Classification of Educational Goals, Handbook I : Cognitive Domain.
New York, David Mc Kay, 1956.
Trad. française de Marcel Lavallée : Taxonomie des objectifs pédagogiques. Tome 1 : Domaine cognitif.
Montréal, Education Nouvelle, 1969, 233p.

KRATWOHL (David R.) ; BLOOM (Benjamin S.) and MASIA (Bertram) : Taxonomy of Educational Objectives, Handbook II : Affective Domain.
New York, David Mc Kay, 1964.
Trad. française de Marcel Lavallée : Taxonomie des objectifs pédagogiques. Tome 2 : Domaine affectif.
Montréal, Education Nouvelle, 1969. 233p.

MAGER (Robert F.) : Preparing Instructional Objectives.
Palo Alto (Calif.), Fearon Publishers, 1962, 62p.
Trad. française de George Decote : Comment définir des objectifs pédagogiques.
Paris, Gauthier-Villars, 1971.

VALETTE (Rebecca M.) ; DISICK (Renée S.) : Modern Language Performance Objectives and Individualization. A Handbook.
New York, Harcourt Brace Jovanovich, 1972, 261p.

.../...

F. Statistiques.

DIEDERICH (Paul B.) : Short-cut statistics for teacher-made tests. Evaluation and Advisory Service Series, n° 5, 2nd ed. Princeton, Educational Testing Service, 1964.

MIALARET (Gaston) et PHAM (Daniel) : Statistique à l'usage des éducateurs. Paris, Presses Universitaires de France, 1967. 266p.

MULLER (Charles) : Initiation à la statistique linguistique. Paris, Larousse, 1968, 248p.

Notions élémentaires de statistiques à l'usage des professeurs de langue. Commission des Tests, Institut de Linguistique Appliquée et C.R.D.P., Strasbourg.

LEXIQUE DU TESTING

ACQUISITION (test d')

Dans certaines terminologies, synonyme de "test de CONTROLE"

ADMINISTRATION (d'un test)

Le fait d'administrer, de faire passer un test à des sujets (point de vue du constructeur ou de l'examinateur.)

ANALYSE DES ITEMS

Opération destinée à comparer l'information apportée à chaque item à l'information fournie par le test dans son ensemble, et à évaluer la validité de chaque item, cf. 6.4.

APPRECIATIVE (Notation)

Notation non chiffrée, exprimée par une appréciation subjective (exemples : excellent, bien,, passable, etc.)

APPRENTISSAGE (Effet d')

Phénomène qui fait qu'au cas de deux passations rapprochées d'un même test par les mêmes sujets, les performances de la seconde passation sont légèrement supérieures en moyenne à celles de la première.

APTITUDE (Test d')

Test de pronostic destiné à évaluer les probabilités de succès dans un type déterminé d'apprentissage ultérieur.

CENTILAGE

Système d'étalonnage consistant à diviser une population en 100 classes d'effectif égal, et à attribuer à chaque sujet une note-centile, exprimant le pourcentage de sujets se classant après le sujet considéré.
Exemple : une note-centile de 85 signifie que le score correspondant est supérieur à 85 % des scores.

CENTILE

Rang occupé sur une échelle d'évaluation à 100 degrés.
Syn. : NOTE-CENTILE

CHOIX

Synonyme d'OPTION

CHOIX MULTIPLES

Voir : Q.C.M.

CLASSE

1. Groupe d'individus rangés sur le même degré d'une échelle d'évaluation.
2. (Test de) Par opposition à un test STANDARDISE, test construit par un professeur pour ses élèves.
   Angl. : Classroom test.

CLASSEMENT (Test de)

Type de test de niveau destiné à répartir les sujets en plusieurs classes de niveau pour une répartition déterminée.
Syn. : test d'ORIENTATION
Angl. : "Placement test".

.../...

COMPLETION

Tâche exigée des sujets dans certains items et consistant à complèter un énoncé inachevé ou à trous.

COMPREHENSION (Epreuve de)

1. Au sens large, par opposition à "épreuve de PRODUCTION", épreuve dans laquelle le sujet est placé en position de récepteur.
2. Au sens restreint, par opposition aux épreuves qui ne mettent en jeu que la perception (exemple : test de DISCRIMINATION phonétique), épreuve destinée à vérifier si le sujet est capable d'appréhender la signification d'énoncés oraux ou écrits.

CONSIGNES (d'un test)

1. Instructions données aux sujets, avant le test ou le sous-test, pour leur préciser la nature de la tâche demandée et éventuellement leur fournir des indications sur la présentation de leurs réponses.
   Syn. : DIRECTIVES
2. Indications réservées à l'examinateur sur les modalités d'administration et de correction du test.
   Syn. : INSTRUCTIONS

CONSTRUCTION (d'un test)

Elaboration, fabrication d'un test. L'auteur d'un test s'appelle le CONSTRUCTEUR du test.

CONTAMINATION

1. En docimologie, influence exercée sur le jugement de l'examinateur par des facteurs extérieurs à ce qui fait l'objet de l'évaluation.
2. Influence négative exercée sur l'apprentissage par des erreurs qui n'ont pas été produites par l'élève lui-même.

CONTROLE (Test de)

Test de diagnotic mesurant l'acquisition d'un programme déterminé à l'issue d'un cycle d'enseignement. Cf. 2.2.
Angl. : Achievement test, attainment test.

CORRELATION

En statistique, degré de similitude entre les valeurs correspondantes des caractères de deux (ou plusieurs) variables, par exemple entre les résultats obtenus à deux tests par le même groupe de sujets.
La corrélation entre deux tests sera forte, ou élevée, si la plupart des sujets obtiennent aux deux épreuves des résultats similaires. Sinon, la corrélation sera faible, ou basse. Elle se mesure par un INDICE DE CORRELATION.

CRITERE

1. (de classification). Caractère dont la présence ou l'absence permet d'effectuer une distinction entre deux éléments lors d'une classification.
2. Tâche ou ensemble de tâches, qu'un sujet doit avoir accomplies pour que sa performance soit considérée comme acceptable.
   TEST SE REFERANT A UN CRITERE : Par opposition aux tests se référant à une NORME, test destiné à mesurer, en fonction d'un critère déterminé à l'avance, si les objectifs de performances ont été atteints.
   Angl. : criterion-referenced test.

.../...

DIAGNOSTIC (Test de)

1. Au sens large, par opposition aux tests de PRONOSTIC, test destiné à mesurer des acquisitions.
2. Au sens restreint, synonyme de "test de PROGRES".
   Syn. : TEST DIAGNOSTIQUE.

DIRECTIVES

Synonyme de CONSIGNES 1.

DISCRET (Item)

Par opposition à "item sur PASSAGE", item contenant en lui-même tous les éléments de la question.

DISCRIMINATION

1. Distinction, effectuée par un test, entre les sujets capables de réussir certaines tâches et ceux qui ne le sont pas. La capacité d'un test d'effectuer cette distinction, s'appelle son POUVOIR DISCRIMINATIF.
2. Tâche exigée des sujets dans certaines épreuves des tests de langue et consistant à reconnaître si deux (ou plusieurs) énoncés sont différents ou identiques.

DISTRACTEUR

Dans un item à choix multiples, option autre que la réponse correcte.

DOCIMASTIQUE

"Technique des examens" (G. de LANDSHEERE, ouvrage cité, p. 13)

DOCIMOLOGIE

"Science qui a pour objet l'étude systématique des examens, en particulier des systèmes de notation, et du comportement des examinateurs et des examinés" (G. de LANDSHEERE, ouvrage cité, p. 13).

DOXOLOGIE

"Etude systématique du rôle que l'évaluation joue dans l'éducation scolaire" (G. de LANDSHEERE, ouvrage cité, p. 13).

ECART REDUIT

Quotient z de la déviation d'une valeur x de la variable par rapport à la moyenne $\bar{x}$, divisée par l'écart type σ :

$$z = \frac{x - \bar{x}}{\sigma}$$

ECART TYPE

Paramètre de dispersion le plus couramment utilisé, qui exprime la moyenne des déviations (écarts) par rapport à la moyenne arithmétique. C'est la racine carrée de la variance.

$$\sigma = \sqrt{\frac{\Sigma n_K (x_K - \bar{x})^2}{n}}$$

Syn. : ECART QUADRATIQUE MOYEN, ECART ETALON
Angl. : standard déviation.

.../...

ECHELLE D'EVALUATION

Série continue de degrés d'évaluation, servant à ordonner des performances individuelles, par rapport à une qualité déterminée, en un certain nombre de classes.
Syn. : ECHELLE DE NOTATION

ECHELLE NORMALISEE

Système d'étalonnage dans lequel les classes (généralement 5 ou 9) sont déterminées par les déviations par rapport à la moyenne, calculées en écarts types.

ECONOMIE

Qualité d'un test qui fait qu'il peut être considéré comme peu coûteux en argent, et surtout en temps et en énergie. On peut distinguer l'économie de construction, de passation, de correction d'un test.
Syn. : ECONOMIE D'EMPLOI

EFFICIENCE (Test d')

Dans certaines terminologies, désigne ce qui est appelé ici test de NIVEAU.

EPREUVE

1) Terme générique désignant toute tâche ou ensemble de tâches dont l'accomplissement est soumis à une évaluation.
Angl. : test.
2) Synonyme de SOUS-TEST.

EXAMINATEUR

Personne chargée de la correction d'une épreuve. Dans les épreuves à passation individuelle (exemple : entretien dirigé), l'examinateur est en outre chargé de poser les questions.

EXERCICE-TEST

Test de classe conçu comme un exercice autant que comme une évaluation.

EXPRESSION (Epreuve d')

Parmi les épreuves de production des tests de langue, apellation habituellement réservée aux épreuves qui exigent du sujet qu'il produise des énoncés comme dans une situation authentique de communication.

FEUILLE DE REPONSES

Feuillet sur lequel les sujets doivent inscrire leur réponse, et qui peut être distinct du livret de test contenant les consignes et les questions. Pour la correction automatique, on utilise des cartes de réponse.
Angl. : answer sheet.

FIDELITE

Qualité d'une épreuve qui fait que les résultats obtenus restent stables et constants indépendamment des conditions de passation et de correction.
Angl. : reliability.

.../...

FORMATIVE (Evaluation)

Par opposition à SOMMATIVE, évaluation qui consiste à comparer les résultats obtenus par un sujet donné aux objectifs de performances visés, et qui s'intègre par conséquent au processus d'apprentissage.

GRILLE DE REPONSES

Document destiné à l'examinateur et où apparaissent les réponses considérées comme correctes.
Angl. : answer grid.

HALO (Effet de)

Facteur de contamination dans la notation, qui fait varier affectivement le jugement d'un examinateur en fonction de préjugés favorables ou défavorables dus à des caractères extérieurs à ce qu'il s'agit de mesurer.

HYBRIDE (Test)

Par opposition à un test PUR :

1. Test de langue faisant intervenir deux skills différents du comportement linguistique.
2. Test de langue dont le stimulus primaire et le stimulus secondaire empruntent un canal de communication différent.

i (Courbe en)

Courbe représentant une série de valeurs dont la distribution est telle que le mode est proche ou se confond avec la valeur minimale.
Cf. 6.0.

INDICE DE CORRELATION (de deux tests)

Nombre fractionnaire, situé entre + 1 et - 1, et indiquant le degré de corrélation entre deux variables, par exemple entre deux séries de résultats obtenus par le même groupe de sujets à deux tests. L'indice 1 marque une corrélation parfaite, - 1 une corrélation inverse absolue, 0 une corrélation nulle.

L'indice de SPEARMAN ρ mesure la corrélation des rangs :

$$\rho = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

L'indice de BRAVAIS-PEARSON r mesure la corrélation des déviations par rapport à la moyenne :

$$r = \frac{\sum (x_k - \bar{x})(y_k - \bar{y})}{n\sigma_x\sigma_y}$$

cf. 6.3.
Syn. : COEFFICIENT DE CORRELATION.

INDICE DE DIFFICULTE (d'un item)

Nombre fractionnaire, situé entre les limites 0 et 1, et indiquant la proportion des sujets ayant fourni une réponse correcte à l'item considéré :

$$\text{Indice de difficulté} = \frac{\text{Nombre de réponses correctes}}{\text{Nombre de sujets}}$$

L'indice 1 signifie que 100% des sujets ont fourni une réponse correcte et que la difficulté de l'item est minimale pour la population considérée. L'indice 0 exprime la difficulté maximale (aucun sujet n'a fourni la réponse correcte.) Cf. 6.4.

.../...

INDICE DE DISCRIMINATION (d'un item)

Nombre fractionnaire, situé entre les limites +1 et -1, et indiquant le pouvoir discriminatif d'un item. On l'obtient en comparant le nombre de réponses correctes fournies par le groupe des sujets ayant obtenu les meilleurs scores au test entier (le tiers, ou le quart, ou 27% des sujets, selon la technique adoptée d'analyse des items) et le nombre de réponses correctes fournies au même item par le groupe, d'effectif égal, des sujets ayant obtenu les scores les plus bas. Par exemple, si l'effectif de ces deux groupes est le tiers :

Indice de discrimination =

$$\frac{\text{Nombre de réponses correctes du 1/3 supérieur - Nbre de réponses correc du 1/3. inférieur}}{\text{1/3 du nombre total de sujets.}}$$

cf. 6.4.

INDICE DE FIDELITE (d'un test)

Nombre fractionnaire, situé entre 0 et 1 et indiquant le degré de corrélation interne du test. Il se calcule à partir de l'indice de corrélation entre les deux moitiés du test (méthode pairs/impairs).

$$\text{Indice de fidélité} = \frac{2r}{1 + r}$$

INSTRUCTIONS

Synonyme de CONSIGNES 2.

ITEM

Mot latin = "en outre, de plus"
Chacun des points numérotés d'un test, sollicitant du sujet une réponse particulière et donnant lieu à une notation séparée.
Syn. : QUESTION

J (Courbe en)

Courbe représentant une série de valeurs dont la distribution est telle que le mode est proche ou se confond avec la valeur maximale.
Cf. 6.0.

MEDIANE

L'une des tendances centrales (ou paramètres de position) d'une variable : c'est la valeur qui se situe exactement au milieu de la série ordonnée.
Syn. : MEDIAN

MODE

L'une des tendances centrales d'une variable : c'est la valeur la plus fréquente de la variable (le sommet de la courbe).
Une courbe à deux modes est dite "bimodale".

MOYENNE

1. L'une des tendances centrales d'une variable, obtenue en divisant la somme de toutes ses valeurs par son effectif (pour un test, en divisant la somme des scores par le nombre de sujets) :

$$\bar{x} = \frac{\Sigma x}{n}$$

Syn. : MOYENNE ARITHMETIQUE

2. Note correspondant à la moitié de la note maximum (exemple : 10 sur 20)

.../...

NIVEAU (Test de)

1. Test de pronostic évaluant les acquisitions en fonction de certaines prévisions (tests d'orientation, de sélection). Cf. 2.1.
   Angl. : Proficiency Test.
2. Dans certaines terminologies, synonyme de test de CONTROLE.

NORMAL

Conforme à la norme statistique.
- LOI NORMALE, ou loi de LAPLACE-GAUSS : loi de distribution statistique des valeurs d'une variable aléatoire.
- DISTRIBUTION NORMALE, ou GAUSSIENNE : distribution conforme à la loi normale.
- COURBE NORMALE, ou COURBE EN CLOCHE, ou COURBE DE GAUSS : courbe figurant une distribution normale.

NORMALISATION

Opération d'étalonnage consistant à définir des normes statistiques permettant de situer les performances individuelles et à établir des notes dites NORMALISEES.

NORME

1. Distribution statistique conforme à la loi normale.
TEST SE REFERANT A UNE NORME : Par opposition aux tests se référant à un CRITERE, test destiné à situer une performance individuelle parmi celles d'une population.
Angl. : norm-referenced test.

2. En pédagogie et dans une perspective dite "normative", désigne ce qui doit être considéré comme correct par opposition à ce qui est jugé inacceptable.

NOTATION

Attribution de notes à des performances scolaires.

NOTE

Appréciation, chiffrée ou non, traduisant l'évaluation d'une performance. La note peut être soit un score brut, soit un score transformé (par exemple par une normalisation), soit une appréciation subjective.

NOTE STANDARD

Note normalisée obtenue en divisant par l'écart type σ la déviation d'un score brut x par rapport à la moyenne $\bar{x}$ (c'est l'écart réduit du score considéré) :

$$z = \frac{x - \bar{x}}{\sigma}$$

Syn. : NOTE Z

OBJECTIVITE

Qualité d'un test qui fait que les résultats obtenus par un sujet donné seront constants quel que soit l'examinateur (celui-ci pouvant éventuellement être remplacé par une machine). C'est l'un des facteurs de la fidélité.
Syn. : OBJECTIVITE DE NOTATION

OPTION

Chacune des réponses proposées dans un item à choix multiple (réponse correcte ou distracteur).
Syn. : CHOIX, REPONSE 2.

.../...

ORIENTATION (Test d')

Synonyme de test de CLASSEMENT.

PARALLELES (Tests)

Tests de forme et de contenu différents, mais de niveau et d'objectif équivalents.

PARAMETRE

En statistique, grandeur mesurable permettant de présenter de façon plus simple et plus abrégée les caractéristiques principales d'un ensemble statistique.

PARAMETRES DE DISPERSION (d'une variable)

Paramètres donnant une indication générale sur la façon dont les valeurs de la variable sont plus ou moins groupées autour de la moyenne arithmétique. Les plus couramment utilisés sont la variance et surtout l'écart type.

PARAMETRES DE POSITION (d'une variable)

Paramètres indiquant l'ordre de grandeur de l'ensemble des valeurs de la variable : ce sont les valeurs centrales, indiquant dans quelle partie du champ les valeurs de la variable ont tendance à s'accumuler. Les plus couramment utilisés sont la médiane, le mode et surtout la moyenne.
Syn. : TENDANCES CENTRALES.

PASSAGE (Item sur)

Par opposition à un item DISCRET, item d'un test de compréhension dont la réponse nécessite la lecture ou l'audition préalable d'un texte.

PASSATION (d'un test)

Le fait de passer, de subir un test (point de vue du sujet ; du point de vue du constructeur ou de l'examinateur, on parlera d'ADMINISTRATION).

PERFORMANCE

1. Comportement par lequel un sujet manifeste sa compétence dans le domaine qui fait l'objet de l'évaluation.
2. Résultat obtenu par un sujet à une épreuve donnée.

POPULATION

1. En termes de statistique, tout ensemble d'individus (= de personnes, ou d'objets, etc.) soumis à une analyse statistique.
2. Ensemble des sujets soumis à un même test.

POST-TEST

Par opposition à PRE-TEST, seconde passation, après l'apprentissage, d'un test qui a été admihistré une première fois avant l'apprentissage.

POUVOIR DISCRIMINATIF

Qualité d'un test ou d'un item qui fait qu'il contribue plus ou moins à distinguer les meilleurs sujets des plus faibles en ce qui concerne le domaine considéré. Le pouvoir discriminatif d'un item par rapport à celui du test entier se mesure par un indice de discrimination.

PRE-TEST

1. Version provisoire, destinée à la pré-expérimentation, d'un test standardisé.
2. Par opposition à un POST-TEST, première passation, avant l'apprentissage, d'un test destiné à être administré une seconde fois après l'apprentissage.

PRODUCTION (Epreuve de)

Par opposition à une épreuve de COMPREHENSION, épreuve dans laquelle le sujet est placé en position d'émetteur et doit construire lui-même sa réponse.

PROGRES (Test de)

Test de diagnostic destiné à mesurer les résultats obtenus par les élèves en comparaison avec les objectifs de l'unité de cours, et à jouer un rôle de feed-back sur le processus d'apprentissage. Cf. 2.2.

PRONOSTIC (Test de)

Test destiné à établir certaines prévisions à partir de données actuelles, par opposition aux tests de DIAGNOSTIC, cf. 2.1.
Syn. : TEST PRONOSTIQUE

PUR (Test)

Par opposition à un test HYBRIDE :

1. Test de langue ne faisant intervenir qu'un seul des quatre skills du comportement linguistique.
2. Test de langue dont le(s) stimulus( ) emprunte(nt) un seul canal de communication (oral ou écrit).

Q.C.M. = QUESTIONNAIRE A CHOIX MULTIPLES

1. Par opposition aux items o REPONSE OUVERTE, où le sujet doit construire lui-même sa réponse, questionnaire composé d'items comportant un nombre limité de réponses proposées par le constructeur du test, et parmi lesquelles le sujet doit choisir la réponse qu'il considère comme correcte.
   Syn. : QUESTIONNAIRE A REPONSES FERMEES
2. Au sens restreint et par opposition aux items de type VRAI/FAUX, questionnaire composé d'items comportant au moins 3 options.

QUALIFICATION (Test de)

Dans certaines terminologies, désigne ce qui est appelé ici test de NIVEAU.

QUESTION

1. Synonyme d'ITEM.
2. Synonyme de STIMULUS PRIMAIRE.
3. Synonyme de STIMULUS SECONDAIRE.

QUESTIONNAIRE

Ensemble d'items.
QUESTIONNAIRE A CHOIX MULTIPLES : Voir Q.C.M.

QUIZ

Test de classe très bref portant sur le contenu d'une petite unité de cours. Cf. 2.3.

RENDEMENT (Test de)

Dans certaines terminologies, catégorie de tests englobant les tests de niveau et les tests de contrôle.

REPONSE

1. Toute réaction du sujet à une question. Cette réponse peut être non-linguistique (cocher une case, exécuter un ordre). Si elle est linguistique, elle peut être très brève ou relativement longue (composition écrite, récit oral).
2. Synonyme d'OPTION.

REPONSE FERMEE (Item à)

Synonyme d'item à CHOIX MULTIPLES.

REPONSE OUVERTE (Item à)

Par opposition aux items à CHOIX MULTIPLES, item pour lequel le sujet doit construire (énoncer ou rédiger) lui-même sa réponse.
Syn. : Item à REPONSE LIBRE.

REPRESENTATIVITE

Qualité d'un test qui fait que l'échantillon que constitue son contenu est représentatif (est une réduction non déformante) du contenu de l'apprentissage qu'il est censé mesurer (test de diagnostic) ou des capacités exigées pour atteindre un certain niveau (test de niveau) cf. 3.2., 3.3.

SCORE

Résultat chiffré traduisant une évaluation effectuée selon un calcul objectif préétabli, avant toute éventuelle transformation.
Syn. : SCORE BRUT.

SELECTION (Test de)

Type de test de niveau destiné à répartir les sujets en deux classes en fonction d'un seuil de succès. Cf. 2.2.

SKILL

Mot anglais utilisé tel quel en français ou traduit par "habileté, savoir-faire", et désignant l'ensemble des dimensions intervenant dans l'accomplissement d'une tâche définie. On distingue traditionnellement dans le comportement linguistique 4 skills :

- compréhension orale (audition),
- production orale (élocution),
- compréhension écrite (lecture),
- production écrite (écriture).

SOMMATIVE (Evaluation)

Par opposition à FORMATIVE, évaluation dont le but est de comparer entre elles ou de classer les performances d'une population.

SOUS-TEST

Partie d'un test constituant un ensemble particulier d'items, généralement parce qu'elle fait intervenir un skill ou une composante linguistique particuliers, ou parce qu'elle comporte des consignes spécifiques.

.../...

STANDARDISATION

Uniformisation des conditions de passation et de correction d'un test, afin que tous les sujets soient placés dans la même situation au moment de l'administration, et que leurs performances soient évaluées de la même façon lors de la correction.

STANDARDISE (Test)

Par opposition à un test de CLASSE, test destiné à être administré à une population nombreuse.

STEM

Dans un Q.C.M., partie de l'item extérieure aux options.

STEREOTYPIE

Facteur de contamination dans la notation, qui instaure dans le jugement porté sur l'élève par l'examinateur une sorte d'immuabilité par rapport au jugement qu'il avait porté sur lui en fonction de performances antérieures.

STIMULUS

Tout ce qui, dans un test, est destiné à provoquer une réponse du sujet.

STIMULUS PRIMAIRE

Donnée première présentée par le constructeur et dont l'appréhension par le sujet est nécessaire pour répondre. Exemples : texte oral ou écrit, série d'images, sur lequels sont posées des questions - stimulus secondaire.

STIMULUS SECONDAIRE

Question posée à propos d'un stimulus primaire et sollicitant du sujet une réponse particulière.

SUJET

Individu soumis à un test.

TAXONOMIE

Classification hiérarchisée dans laquelle les éléments sont ordonnés du plus simple au plus complexe.
Syn. : TAXINOMIE.

TENDANCES CENTRALES (d'une variable)

Synonyme de PARAMETRES DE POSITION.

TEST-RETEST

Moyen de mesurer la fidélité d'un test, consistant à administrer aux mêmes sujets soit deux fois le même test, soit deux tests parallèles.
Cf. 1.3.

TROU

Blanc remplaçant dans un texte écrit une suite quelconque de graphèmes (mot, élément de mot, suite de mots) que le sujet est invité à reconstituer par complétion.

VALIDITE

Qualité d'une épreuve qui fait qu'elle mesure effectivement ce qu'elle est censée mesurer. Cf. 3.0.

VALIDITE CONCOURANTE

Validité établie par corrélation avec un autre test reconnu comme valide.

.../...

VALIDITE DE CONTENU

Qualité d'une épreuve qui fait que son contenu correspond à ce qu'elle est censée mesurer. Cf. 3.0.

VALIDITE PREDICTIVE

Qualité d'un test de pronostic qui fait qu'il permet de prédire avec le minimum d'erreur les chances de succès des sujets dans une activité déterminée.

VARIANCE

Paramètre de dispersion d'une variable. C'est la moyenne σ^2 des carrés des déviations par rapport à la moyenne :

$$\sigma^2 = \frac{\sum n_k (x_k - \bar{x})^2}{n}$$

C'est le carré de l'écart type.

VRAI/FAUX (Item de type)

Type d'items ne comportant que deux options (vrai ou faux, oui ou non).

Z (Note)

Synonyme de NOTE STANDARD

TABLE DES MATIERES

.../...

Pages