Cadre européen commun de référence pour les langues :
Apprendre, Enseigner, Evaluer

Evaluation de compétences en langues et conception de tests

Préparé sous la direction de

M. Milanovic (A.L.T.E.)

Division des Politiques Linguistiques
Strasbourg, octobre 2002

TABLE DES MATIÈRES

1.0 INTRODUCTION GÉNÉRALE

1.1 Le but de ce guide

Ce guide a été conçu pour aider tous ceux qui sont engagés dans la préparation de tests de langue et, notamment, s'ils veulent utiliser le « Cadre européen commun de référence pour les langues : Apprendre, Enseigner, Evaluer » du Conseil de l'Europe. On a voulu faire de ce guide un outil approprié, non seulement pour les élaborateurs d'examens en situation officielle comme celle de préparation d'épreuves nationales, mais aussi pour les professeurs qui préparent des tests pour leur classe. Trouver un équilibre qui réponde aux besoins de ces deux groupes a constitué un véritable défi; c'est pourquoi nous encourageons les lecteurs qui préparent du matériel d'évaluation à examiner et à prendre en compte les conseils prodigués dans ce guide, en fonction de leur propre but ainsi que du temps et des moyens dont ils disposent. On a plus mis l'accent sur les questions de *procédures* que sur celles de produit, persuadés que des principes clairs et des démarches bien planifiées débouchent sur des produits bien conçus et non l'inverse.

1.2 Une approche communicative de la langue

De tout temps, les techniques d'évaluation ont reflété la vision que l'on avait de la langue et de son usage à un moment donné. Ce que l'on teste, et le genre de tâche ou de type d'item choisis pour tester, révéleront l'influence de la pensée dominante sur ce qu'est la capacité langagière et ce que nous faisons précisément quand nous utilisons la langue dans son usage quotidien. L'évaluation de la langue comme outil de communication s'est développée à partir d'un déplacement de la théorie de l'enseignement/apprentissage des langues et de la méthodologie, d'une centration sur la structure vers une accentuation de l'importance du discours, *de la langue telle qu'elle est utilisée.*

Le Cadre de référence du Conseil de l'Europe est le développement normal des travaux antérieurs du Conseil. Il se fonde sur un certain nombre de recherches dont l'influence a été mondiale et qui ont été adoptées dans les métiers de l'enseignement des langues. On pense, bien évidemment, au Threshold Level (van Eck, 1975; van Eck et Trim, 1990), démonstration de l'approche communicative dont l'effet sur la pratique de classe et l'évaluation a été étendu et durable. La Préface de l'édition de 1980 du Threshold Level English recommande une approche fonctionnelle de l'enseignement de la langue afin de "transformer un enseignement de la langue d'une stérilité scolastique, dominé par la grammaire, en un médium vital pour un échange plus libre des personnes et des idées"; cette approche met l'accent sur la langue telle qu'elle est utilisée pour répondre aux besoins quotidiens d'un adulte vivant dans un pays étranger.

Le Threshold Level ou, en français, le Niveau seuil, n'est en aucune façon un cours, un programme ou une liste exhaustive d'éléments linguistiques qu'un apprenant devrait savoir à un certain niveau; c'est une déclaration d'*objectifs* ou encore une tentative "pour définir comment un apprenant devrait être capable d'utiliser une langue afin de se conduire de manière autonome dans un pays où cette langue est le moyen de communication de la vie quotidienne". Cela signifie que l'on doit donner aux apprenants non seulement les moyens de faire des choses comme acheter du lait ou faire réparer sa voiture mais aussi échanger des informations et des opinions avec autrui, parler de ce que l'on aime ou pas et raconter ses expériences. L'accent est nettement mis sur la langue comme outil social ou comme moyen de permettre aux gens d'interagir les uns avec les autres. Le point de départ est l'éventail des situations dans lesquelles se trouvent couramment des apprenants dans un pays étranger; le but est de les rendre capables d'utiliser la langue pour faire ce qu'il est convenable de faire dans ces situations.

1.3 Un modèle pour l'évaluation de la langue

Depuis la publication initiale du Threshold Level, un certain nombre de modèles de compétence communicative ont été proposés. Le plus connu est peut-être celui de Canale et Swain (1981) qui divise la compétence communicative en quatre composantes: grammaticale, sociolinguistique, discursive et stratégique. À la fin des années 80, Bachman (1990) a exposé sa première approche synthétique de la compétence langagière (Communicative Language Ability - CLA) inspirée, de manière évidente, des travaux de Canale et Swain. Il postulait que la compétence langagière recouvre le savoir linguistique ou compétence, combiné à la capacité de le mettre en œuvre de manière appropriée.

Pour l'évaluateur en langues, un modèle de compétence linguistique ou de compétence langagière est important parce qu'il fournit une base utile à la définition du champ de compétence à évaluer. Il est nécessaire d'avoir tout d'abord une idée claire de ce que l'on teste pour décider si un test est valide ou pas (c'est-à-dire s'il évalue vraiment ce qu'il prétend évaluer); cela permet également de concevoir des outils pour le rédacteur d'items ou l'élaborateur de tests tels que des listes de contrôle du contenu du test. Le but général de toute forme d'évaluation en langue est d'avoir un échantillon des compétences langagières des candidats qui permette d'obtenir une représentation réaliste de leur niveau de capacité à utiliser la langue en situation hors test.

Le Cadre de référence actuel contient également un modèle de compétence langagière. On peut le présenter essentiellement comme une définition de la compétence communicative: la **compétence communicative** (sociolinguistique, linguistique, pragmatique) est une forme de **compétence générale** qui conduit à des **activités langagières** (interaction, production, réception, médiation) mettant en œuvre des **tâches,** des **textes** (ou discours) et des **stratégies** dans quatre **domaines** principaux (public, professionnel, éducationnel, personnel) où se présentent des **situations** caractérisées par des **lieux**, des **organismes** (ou institutions) qui structurent l'interaction, des **acteurs** avec des rôles définis, des **objets** (animés et inanimés) qui constituent l'environnement, des **événements** qui y ont lieu et des **opérations** qui y sont exécutées (voir Chapitre 4 du Cadre de référence).

Le Cadre de référence offre aux concepteurs de tests de langue, et à tous ceux qui sont engagés dans la production d'examens, la possibilité d'aller d'un commun accord vers un système commun d'évaluation en langues que justifient les valeurs fondamentales du Conseil de l'Europe dans la conception qu'il a de la citoyenneté européenne, tout en préservant la culture de l'évaluation de chacun et leurs traditions dans ce domaine et en mettant en valeur tout ce qui, dans leur pratique, rejoint les pratiques professionnelles reconnues. Ce guide s'intéresse directement à la tâche immédiate qui attend les examinateurs, à savoir la création d'une gamme étendue de tests qui aient un lieu et une identité selon le Cadre de référence tout en se conformant aux normes européennes et internationales de l'évaluation.

1.4 Autres paramètres pour l'élaboration de tests de langue

Il est important de souligner qu'il n'y a pas obligatoirement de réponse juste dans l'absolu en termes d'évaluation en langue. Aucun mode d'évaluation n'est intrinsèquement meilleur ou pire qu'un autre. Le choix d'un mode d'évaluation dépend d'un certain nombre de facteurs et se fait à la lumière des réponses à un certain nombre de questions. Par exemple:

- le test est-il un test de compétence générale ou évalue-t-il surtout ce qui a été appris en cours ?
- quelle est sa durée ?
- quel est le niveau de performance attendu ?
- a-t-il pour but de classer les élèves ?
- les résultats serviront-ils pour faire un diagnostic ?

On traitera certaines de ces questions ultérieurement dans ce guide.

Si l'on peut dire du Cadre de référence qu'il fournit l'approche théorique nécessaire à la conception et à la production de tests, ce guide a pour but de proposer en résumé un état synthétique des pratiques de production des tests que tout concepteur doit identifier afin de produire un "bon" test au sens le plus large du terme. Le Cadre de référence se focalise sur des questions de contenu; ce guide se concentre plus sur les procédures mises en œuvre pour la conception et la production de tests en utilisant les données du Cadre de référence comme point de départ. Régulièrement dans ce guide nous renverrons le lecteur à certaines parties du Cadre pour plus de détails. En règle générale, les examinateurs trouveront que les Chapitres 3 (Niveaux communs de référence), 4 (L'utilisation de la langue et l'apprenant/utilisateur), 7 (Les tâches et leur rôle dans l'enseignement des langues) et 9 (Evaluation) sont particulièrement utiles.

En définitive, le but du producteur de test est d'assortir la méthode la plus appropriée à l'objectif déclaré d'un test donné. Pour ce faire, il faut essayer d'équilibrer les qualités essentielles d'un test, à savoir: fidélité, validité, faisabilité et impact. La façon d'allier ces qualités dépendra des raisons qu'il y a à produire tel ou tel test. Pour un concours officiel, par exemple, dont les résultats ont une importance décisive sur la vie des gens, fidélité et validité seront les critères les plus importants. En revanche, s'il s'agit de l'évaluation en classe, on pensera à la faisabilité et à l'impact. Ce qui compte, c'est que le rédacteur du test soit pleinement informé des variables que l'on peut manipuler lors de l'élaboration du test et prenne les décisions de manière claire et rationnelle.

Les utilisateurs du Guide qui sont engagés dans la conception de tests envisageront et expliciteront selon le cas:

- *jusqu'où les approches de l'évaluation en langue actuellement en vigueur dans leur système reflètent une façon particulière de considérer la langue et son usage*
- *jusqu'où ces approches sont centrées sur le savoir linguistique et/ou sur la performance communicative*
- *dans quelle mesure ces approches ont un lien avec un modèle explicite de capacité langagière*
- *jusqu'où la spécification du Threshold ou du Niveau seuil, ainsi que celle du Cadre de référence européen, offrent une approche théorique pour la conception et la production de tests*
- *quelle est l'importance relative, dans la conception des tests de langue, de facteurs tels que la culture pédagogique, l'impact social, la disponibilité des ressources, etc.*
- *ce que pourrait être un équilibre approprié des qualités essentielles des tests que sont la fidélité, la validité, la faisabilité et l'impact.*

2.0 PROCESSUS D'ÉLABORATION DE TESTS

Il est utile et important d'envisager le processus d'élaboration de tests comme cyclique et réitéré. Cela suppose que l'on réinjecte dans la démarche la connaissance et l'expérience acquises aux différentes étapes de la procédure pour la réévaluation continue d'un test donné et chacune de ses passations.

La Figure 1 est une tentative de schématisation du processus qui montre sur un schéma directeur les étapes à parcourir à partir de la perception initiale du besoin d'un nouveau test.

- **perception du besoin d'un nouveau test**
- **planification**
- **conception**
- **élaboration**
- **opérationalisation**
- **contrôle**

Ces étapes ne sont pas toutes toujours nécessaires; qu'elles soient toutes respectées ou pas, elles relèvent également d'une décision rationnelle fondée sur les exigences particulières de la situation d'élaboration du test.

2.1 La nature cyclique du processus d'élaboration des tests

La Figure 1 met en évidence la nature cyclique du processus d'élaboration

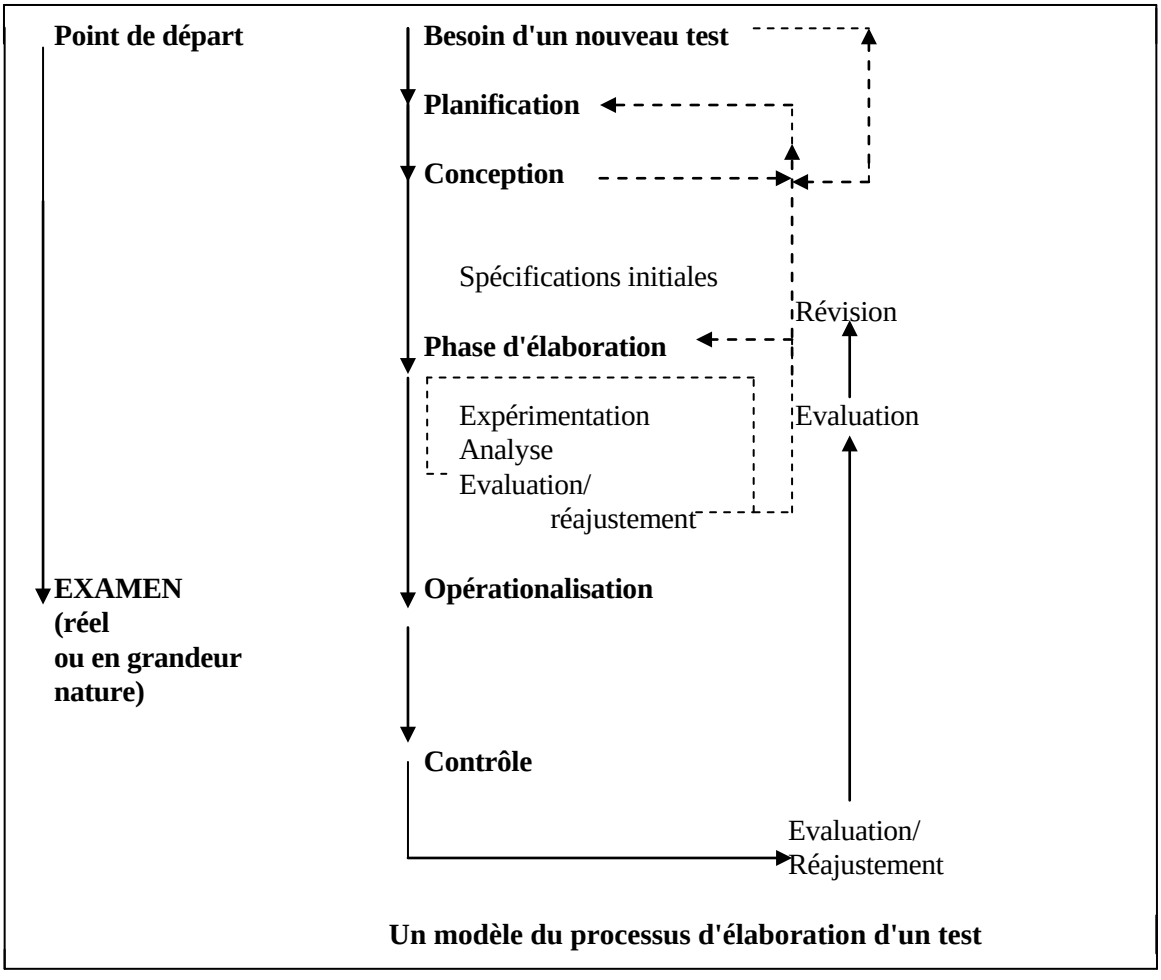

Figure 1 : Modèle de développement d'un test

Une fois reconnu le besoin d'un nouveau test, le modèle suppose une phase de **planification** durant laquelle on recueille les données sur les demandes précises des candidats. En situation de classe, cette étape peut se fonder sur la connaissance personnelle directe que l'on a des étudiants et du programme d'enseignement. Dans des contextes plus larges, on peut collecter l'information à l'aide de questionnaires, de consultation officielle et ainsi de suite. Quelle que soit la situation, on visera à se faire une image claire de ce que sont vraisemblablement les candidats potentiels et à savoir qui seront les utilisateurs des résultats.

La phase de planification est suivie d'une phase de **conception** durant laquelle on essaiera de définir les spécifications initiales d'un test convenable pour les candidats. Les spécifications décrivent et commentent la forme du test et tous les aspects de son contenu ainsi que toutes les variables et contraintes qui l'affectent. Les premières décisions peuvent être prises ici en ce qui concerne la longueur de chaque partie du test, les types d'items choisis et la gamme des sujets utilisables. C'est aussi à ce niveau qu'il faut rédiger des échantillons et les soumettre aux utilisateurs concernés pour réactions et commentaires. Même pour un test de classe, il est toujours bon de soumettre des échantillons à un(e) collègue car une réaction extérieure peut constituer un feed-back inestimable pour le processus d'élaboration.

Pendant la phase d'**élaboration**, on expérimentera ou prétestera l'échantillon. Cela signifie qu'on fera une simulation dans les conditions de l'examen avec des étudiants semblables aux candidats potentiels (en termes d'âge, de formation, etc.) et au niveau approprié. Cette phase peut entraîner l'analyse et l'interprétation des données fournies par les notes des candidats; on peut aussi recueillir des informations utiles à l'aide de questionnaires proposés aux candidats et à leurs enseignants, de comptes rendus ainsi que d'enregistrements audio et vidéo et d'observations. On peut alors prendre des décisions quant au niveau de difficulté du matériel et à son adéquation pour une utilisation dans l'examen définitif. L'expérimentation permet aussi de mettre en place un barème et une échelle de notation assez complets. Même une expérimentation à petite échelle de tests au niveau d'un établissement scolaire ou d'une classe et qui n'utiliserait qu'une poignée de candidats peut se révéler informative sur des points tels que la durée des épreuves individuelles, la clarté des consignes, l'espace laissé pour la réponse, etc. À ce niveau, on peut encore apporter des modifications radicales aux spécifications, aux types d'items utilisés ou à tout autre aspect du test qui pose problème.

Une fois achevées les phases de **planification, conception** et **élaboration**, les spécifications trouvent leur forme définitive, on rédige le contenu et l'on met le test en forme. On met alors en place la procédure de passation et de notation du test. C'est la phase d'**opérationalisation** durant laquelle le test est administré aux candidats (Les différentes étapes de cette phase sont présentées en détail dans la Figure 3; la procédure que nous décrivons trouve le mieux sa place dans les examens scolaires de fin d'année, de fin de cours ailleurs et pour ceux administrés sur une grande échelle).

Lorsque le test est opérationnel, la procédure d'élaboration entre dans sa phase de **contrôle** ou **révision** pendant laquelle il faut soigneusement vérifier les résultats de la passation de l'examen. Cela suppose que l'on ait un feed-back régulier de la part des candidats et des enseignants des écoles dans lesquelles l'examen est passé; il faut aussi analyser la performance des candidats. On utilise ces données pour évaluer l'efficacité du test et envisager sa révision, le cas échéant. On peut conduire une recherche sur divers aspects de la performance du candidat et de l'examinateur afin de voir quelles améliorations apporter au test ou à son administration. Il est probable que le test doive être révisé ultérieurement et toute révision majeure suppose que l'on revienne à la case départ du cycle d'élaboration.

Les utilisateurs du Guide qui sont engagés dans la conception de tests envisageront et expliciteront selon le cas:

- *si leur situation exige un test entièrement nouveau ou si des révisions appropriées peuvent être apportées à un test existant*
- *qui sont les candidats potentiels, quel est leur niveau et quelles sont leurs demandes*
- *comment les contraintes et variables locales affecteront le contenu et la forme du test*
- *comment seront réalisés une expérimentation ou un pré-test adéquats pendant la phase d'élaboration*
- *quelles seront les méthodes les plus appropriées pour le contrôle et l'évaluation à long terme de l'examen*
- *qui seront les utilisateurs des résultats de l'examen et comment seront interprétés les résultats.*

2.2 La définition des spécifications

Lorsqu'on a planifié les spécifications d'un test nouveau (ou révisé), l'objectif sous-jacent est toujours de produire un test qui

- soit **valide** (c'est-à-dire qui propose une démarche adéquate pour mesurer ce qu'il prétend mesurer);
- soit **fidèle** (c'est-à-dire que les résultats obtenus sont aussi exempts que possible d'erreurs de mesure);
- ait un **impact** (c'est-à-dire qui ait un effet positif sur les individus et la pratique de classe);
- soit **faisable** (c'est-à-dire que ses exigences en matière de ressources sur le concepteur et l'administrateur soient compatibles avec les ressources disponibles).

Il faut garder ces facteurs constamment à l'esprit et maintenir entre eux un équilibre acceptable.

La première étape de la planification suppose que l'on se mette en situation d'analyse. Cela signifie que l'on considère le besoin de test dans le contexte des influences diverses qui affecteront sa forme définitive; le but de l'analyse est d'identifier les principales **contraintes** et **variables** pertinentes pour le projet. Elles concernent tous les aspects de ce que le test doit faire afin d'atteindre son but, ainsi que les restrictions imposées au test par les circonstances dans lesquelles il sera utilisé.

2.2.1 Variables et contraintes

En gros, on distingue deux sortes de **variables** que l'on peut qualifier de **professionnelles** et **matérielles**.

Les variables professionnelles se rapportent à ce qu'il faut précisément évaluer et comprennent:

- les situations de communication réelles dans lesquelles les candidats auront besoin de la langue;
- le niveau de performance attendu dans ces situations;
- les événements propres à ces situations réelles qui doivent se retrouver dans le contexte de l'évaluation;
- les informations à donner aux utilisateurs du test avant et après.

Les variables matérielles sont les restrictions que font peser sur l'évaluation :

- le nombre de salles et le personnel disponible;
- le nombre de candidats;
- la durée du test;
- la disponibilité d'examinateurs qualifiés compétents;
- le type d'épreuves qu'il semble bon de proposer;
- le moyen choisi pour communiquer les résultats aux candidats;
- les procédures de contrôle de qualité adoptées.

Les **contraintes** peuvent être:

- l'acceptabilité de l'examen pour tous les utilisateurs: les candidats, leurs parents, les enseignants, les directeurs d'école, etc.
- la façon dont l'examen s'accorde avec le système en vigueur en termes d'objectifs de programme et de pratique de classe;
- le niveau de difficulté requis;
- les attentes extérieures de ce qu'un examen de ce type doit être;
- la disponibilité des ressources pour l'élaboration du test, sa passation et la publication des résultats.

Cette liste n'est en aucune façon ordonnée ni exhaustive. Elle a ici pour but de souligner qu'une bonne compréhension des variables et des contraintes est un préalable nécessaire à une conception de test adéquate et raisonnable. Le Chapitre 4 du Cadre de référence expose une vue d'ensemble utile des nombreuses caractéristiques de l'utilisation de la langue et de l'utilisateur/apprenant qu'il faudra prendre en considération à ce moment de la conception du test. Elles comprennent le contexte de l'utilisation de la langue (la situation de communication -Partie 4.1), la nature des tâches et objectifs de communication (4.3) et le choix des thèmes ou sujets (4.2). Tandis que le contenu des Parties 4.1 à 4.4 du Cadre de référence facilitera l'analyse situationnelle et l'identification de certaines des variables professionnelles qui s'y appliquent, les Parties 4.5, 4.6 et le Chapitre 5 posent des bases utiles pour définir plus en détail les caractéristiques du contenu d'un test afin de définir des spécifications.

Comme le montre la Figure 1, une fois les spécifications ébauchées, on peut faire un premier essai de conception d'un test et de production d'un échantillon. On peut alors l'expérimenter et en analyser les résultats. À la lumière de l'expérimentation, il se peut que l'on rejette certains types d'items ou de supports et que l'on change la longueur des différentes parties ou du mode d'administration. En conséquence, les spécifications pourront subir plusieurs révisions avant de trouver leur forme définitive.

Il peut arriver (dans le cas d'un examen scolaire d'établissement, par exemple) que la même personne soit responsable à la fois de la définition des spécifications et de la rédaction du matériel pour l'examen. Néanmoins, il doit aussi rester possible pour ceux qui n'ont participé ni à la conception ni à l'élaboration du test d'obtenir des informations détaillées à son sujet au moyen des spécifications. Certains auront besoin de cette information afin de décider s'ils présentent des candidats (par exemple, dans le cas d'un test disponible pour tous). D'autres peuvent en avoir besoin pour rédiger des items pour le test; un rédacteur d'items qui n'a jamais produit d'items auparavant pour un test donné et qui n'a pas participé aux différentes étapes de son élaboration a besoin d'un bon descriptif pour le guider; on doit s'attacher à ce que les spécifications répondent à ce besoin.

2.2.2 Problèmes de contenu, de technique et de procédures

Les spécifications définitives (le cahier des charges, en quelque sorte) doivent fournir des informations détaillées sur chaque partie du test ou chaque épreuve, notamment au sujet d'au moins trois aspects du test. Il s'agit des **caractéristiques du contenu** du test (ou ce qui est dans le test), des **caractéristiques techniques**

(telles que le nombre d'items, de parties, etc.) et de la **procédure** (où le test sera passé et comment il sera noté).

On trouvera ci-dessous des exemples de ces trois aspects.

Contenu

- le but des tâches; par exemple, prouver la compréhension détaillée d'un texte, etc. (voir Parties 4.4 et 4.5);
- ce qui est testé; par exemple, appliquer des règles de grammaire (voir Chapitre 5);
- les types de textes choisis comme supports (voir Partie 4.6);
- l'origine des textes (voir Parties 4.1 et 4.6);
- quelque information sur les centres d'intérêt exploitables (voir Parties 4.1 et 4.2);
- les types de déclencheurs des épreuves de production orale (voir Parties 4.3 et 4.4);
- les types de tâches demandées pour les épreuves de production écrite (voir Parties 4.3 et 4.4);

Technique

- durée du test;
- nombre de parties;
- nombre d'items dans chaque partie;
- types d'items dans chaque partie;
- nombre total de textes supports et longueur (en nombre de mots) de ces textes;
- présentation et durée des tâches;
- note attribuée à chaque item et note totale;
- pondération, coefficients;
- si la correction est assurée par des examinateurs, détail de l'élaboration des barèmes et de l'organisation des équipes de correction;
- critères de correction des productions orale et écrite;
- nombre d'examinateurs et de correcteurs, par exemple la double correction est-elle automatique ?;
- détail des procédures de notation et publication des résultats.

Procédure

- dates et lieux des sessions;
- disponibilité d'annales ou de spécimens d'épreuves (épreuves banalisées);
- nombre approximatif d'heures d'enseignement/apprentissage nécessaires à la préparation du test.

Toutes ces informations aident les utilisateurs des spécifications à se faire une idée claire de la nature du matériel.

Le Chapitre 4 du Cadre de référence fournit un ensemble de références particulièrement utiles auxquelles on peut confronter, pour clarification, les traits caractéristiques de tout test en cours de production. Pour ce faire, il faut d'abord rédiger un résumé schématique du test sur lequel on travaille. La Figure 2 montre comment on peut présenter, sous forme de tableau, l'information relative à un examen composé de 5 parties (ou "épreuves"). Pour chaque partie de l'examen sont résumés les objectifs, les supports et la nature de la réponse attendue.

Première épreuve - *Compréhension écrite*	Objectifs	Supports	Forme
	- Comprendre le lexique et les structures - Comprendre globalement un texte, sa fonction et l'idée générale - Comprendre les points essentiels malgré des termes inconnus - Sélectionner des informations particulières dans un texte écrit - Reconnaître une opinion ou un point de vue clairement exprimés - Prouver la compréhension détaillée d'un texte	Partie A - Phrases isolées Partie B - 3 ou 4 textes écrits représentant des types discursifs différents: narratif, descriptif, informatif, argumentatif, etc. Origine: textes littéraires de fiction ou autre, journaux, magazines, publicités, prospectus, etc.	Partie A - 25 items discrets sous forme de QCM à 4 options Partie B - 15 QCM à 4 options dans 3 ou 4 textes
Deuxième épreuve *-Production écrite (composition)*	Objectifs	Supports	Forme
	- Utiliser naturellement la langue qui convient pour répondre à des stimuli thématiques ou situationnels variés	4 stimuli situationnels ou questions sur des sujets quotidiens	Deux tâches écrites sur un choix de 5; longueur exigée entre 120 et 180 mots chacune; types de discours: lettres, descriptions, récits
Troisième épreuve *Utilisation de la langue*	Objectifs	Supports	Forme
	- Utiliser la langue au niveau du mot ou de la phrase, y compris l'usage de mots et de formes structurellement corrects et adéquats; reformulations; dérivation lexicale - Résumer une information dans un texte correct et de longueur appropriée	- Exercices en contexte et hors contexte - Supports visuels (cartes, diagrammes, etc.) pour guider des questions écrites	- Textes lacunaires - Transformations - Dérivation - Construction de phrases - Ecrit guidé

Quatrième épreuve *Compréhension orale*	Objectifs	Supports	Forme
	- Comprendre globalement un texte, sa fonction et l'idée générale - Comprendre les points essentiels malgré des termes inconnus - Sélectionner des informations particulières dans un discours oral - Reconnaître l'humeur et l'attitude quand ils sont clairement exprimés - Comprendre des points de détail dans un discours oral.	3 ou 4 documents enregistrés authentiques ou fabriqués Origine: informations, actualités, conversations, exposés, annonces publiques, etc.	- 3 ou 4 tâches pour un total d'environ 30 questions - les types de tâches peuvent inclure des QCM, des textes lacunaires, de la prise de notes, des vrai/faux, oui/non, etc.
Cinquième épreuve *Production orale*	Objectifs	Supports	Forme
	Participer à une conversation en langue cible sur des thèmes allant du quotidien à des notions plus abstraites; le faire avec aisance, des interactions appropriées, une prononciation correcte au niveau du mot et de la phrase, un vocabulaire juste.	Stimuli visuels ou textuels incluant des photographies et des textes courts. Les déclencheurs peuvent avoir un lien avec des textes lus facultativement.	Une conversation en 3 parties, sur un sujet donné, entre le candidat et l'examinateur: 1) Parler d'une photographie 2) Parler d'un texte court 3) Avoir un échange. L'entretien peut avoir lieu seul(e) avec le professeur ou à 2, ou en groupes de 3.

Figure 2: Tableau d'informations sur un examen

Les utilisateurs du Guide qui sont engagés dans la rédaction de spécifications envisageront et expliciteront selon le cas:

- *quels sont le type et le niveau des besoins de performance langagière à évaluer*
- *quels sont les types de tâches qui permettent d'y parvenir*
- *quelles sont les ressources matérielles disponibles, par exemple locaux, personnel, etc.*
- *quels problèmes politiques, sociaux et/ou économiques risquent d'influencer la production du test*
- *à qui devrait-on demander de définir les spécifications des tests et d'élaborer les échantillons, par exemple en termes d'expertise, d'influence, d'autorité, etc.*
- *comment seront décrits dans les spécifications les détails relatifs au contenu, à la technique et à la procédure de passation*
- *quelle sorte d'information sur les tests doit être fournie aux usagers et sous quelle forme, par exemple la publication des spécifications.*

2.3 Le processus d'élaboration

Les spécifications fournissent une définition de ce qui doit être produit pour un examen. Cette partie met l'accent sur le processus de production tel qu'il se déroule, généralement en cinq étapes:

- **appel d'offre, commande**
- **contrôle/révision et mise en forme**
- **expérimentation ou pré-test**
- **analyse et mise en banque du matériel**
- **production des épreuves proprement dites**

La Figure 3 ci-après illustre ce processus de production, mais le niveau de formalisation de ces étapes dans un contexte donné dépendra du public cible et de l'utilisation des résultats de l'évaluation.

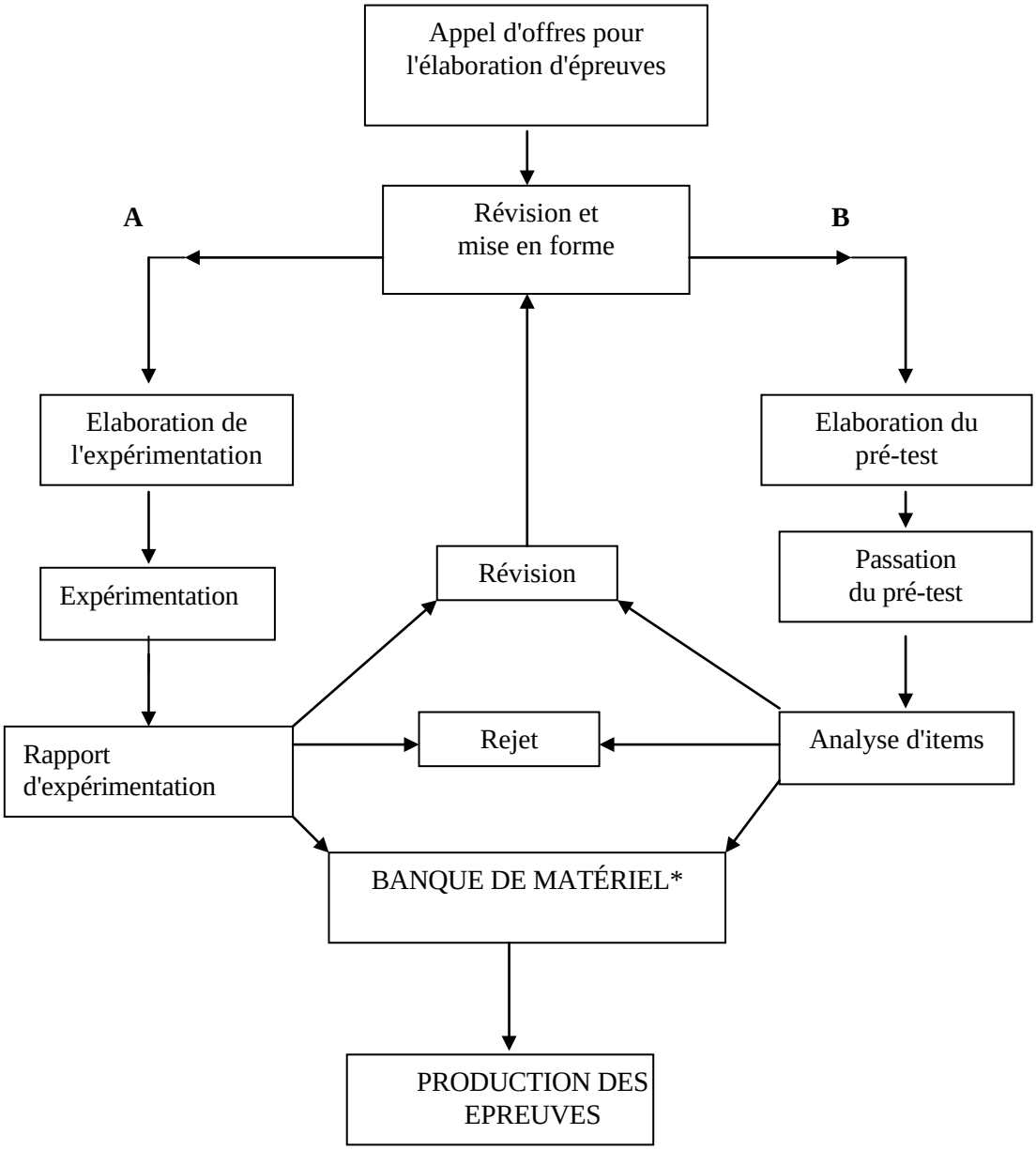

Figure 3: La phase opérationnelle de la production de tests

* banque informatisée de matériel expérimenté, archivage de matériel non expérimenté

La Figure 3 montre comment toutes les épreuves sous-traitées passent d'abord par une étape de révision/contrôle et de mise en forme. C'est à ce niveau que le matériel peut suivre deux voies sensiblement différentes - A ou B - avant d'arriver au point où on le considère acceptable pour la production des épreuves et la mise en banque.
Sur la voie A, le matériel est expérimenté avec un échantillon réduit de population. Bien que cette expérimentation ne puisse déboucher que sur une analyse statistique limitée, elle apporte néanmoins une information tout à fait valable sur l'efficacité de la tâche, le degré de difficulté et la qualité des réponses; elle est donc particulièrement utile en ce qui concerne les épreuves "subjectives" du test, par exemple les tâches de production orale et écrite.

En B, le matériel est pré-testé sur une population plus large et offre ainsi de vraies possibilités d'analyse statistique, y compris l'analyse des items. C'est pourquoi le pré-test est particulièrement approprié pour les épreuves "objectives".

Quel que soit le nombre d'individus engagés dans le processus, deux grands principes doivent être retenus à toutes les étapes:

- le **calendrier** (ce qui signifie que l'on planifie de manière réaliste et que l'on respecte les échéances);

- l'**enregistrement** ou compte rendu (c'est-à-dire le rapport exact et détaillé de toutes les décisions prises et de toutes les modifications apportées au matériel lors des différentes étapes de la production)

Etablir un calendrier convenable est essentiel pour s'assurer que le matériel passe par toutes les étapes du processus de production et devient utilisable en fin de compte pour l'examen réel. L'enregistrement est vital pour permettre révisions et modifications et si l'on doit produire plusieurs versions du même matériel.

2.3.1 Appel d'offres et commande

On appelle **appel d'offres** la démarche qui consiste à recruter des gens pour produire des tests. Comme on l'a dit plus haut, la même personne (que l'on peut appeler le coordinateur) peut avoir l'entière responsabilité de tout le processus de production, y compris celle de la rédaction d'items; il en est souvent ainsi pour les examens dans un établissement scolaire. Toutefois, dans d'autres situations, le coordinateur peut déléguer à un certain nombre d'autres personnes la sélection ou la production de textes ou d'items; il peut s'agir d'autres membres de la même institution ou de personnes extérieures mais ayant à voir avec l'enseignement ou l'évaluation. Il arrive que le membre de l'institution qui produit un test donné ait la responsabilité d'organiser l'appel d'offres et la commande, de suivre toutes les étapes de la production et d'utiliser les items produits pour bâtir les épreuves, tandis que d'autres s'occupent de l'expérimentation, de l'analyse et de la mise en banque des données. La même personne peut avoir la responsabilité de toutes les parties de l'examen ou bien, dans le cas d'un examen comprenant différentes épreuves de compréhension orale et écrite et de production orale et écrite par exemple, chaque épreuve peut être prise en charge par une personne différente.

L'appel d'offres peut se faire à des dates régulières (deux fois par an, par exemple) ou avoir lieu lorsque le coordinateur juge qu'il a besoin de matériel nouveau. On peut demander aux rédacteurs d'items soit un examen complet, soit des items pour telle ou telle partie.

Le coordinateur a pour but d'obtenir une proportion aussi élevée que possible de matériel qui, après traitement, sera jugé globalement acceptable et pourra être finalement utilisé pour des épreuves réelles. Une partie de sa responsabilité consiste donc à trouver et choisir des producteurs d'items compétents et à leur donner les consignes et la formation les plus claires et les plus facilitantes. Les producteurs extérieurs d'items

se recrutent souvent parmi des gens qui connaissent l'examen, soit qu'ils y préparent des étudiants, soit qu'ils en sont correcteurs d'écrit ou examinateurs d'oral. Que le coordinateur travaille seul ou en équipe ou délègue à des rédacteurs extérieurs, les points ci-dessous doivent être clairs:

- **Précisions sur le matériel attendu**

  Cela comprend les précisions sur le nombre de textes, de tâches et d'items requis.

  Dans le cas de textes, il faut savoir si les items doivent être rédigés immédiatement ou seulement après acceptation du texte. Le rédacteur d'items doit fournir la clé de tous les items, y compris les autres solutions correctes possibles.

  Pour la compréhension orale, il faut demander l'enregistrement avec sa transcription. Il ne s'agit pas nécessairement d'un enregistrement professionnel; une cassette enregistrée à la maison peut s'avérer très utile lors de la mise en forme.

  Dans le cas de la production orale, le producteur du test doit être clairement informé s'il doit fournir des déclencheurs visuels ou indiquer seulement quels types de déclencheurs seront nécessaires.

- **Précisions sur la présentation attendue du matériel**

  La meilleure présentation est probablement celle d'un texte dactylographié et l'on peut demander la disquette ainsi qu'une copie papier. Un manuscrit est toujours plus difficile à mettre en forme et peut ne pas être accepté

  Si le rédacteur produit un examen complet, il doit savoir si les items doivent être numérotés en continu et si les parties se suivent ou si chaque partie ou chaque tâche sont présentées séparément sur une nouvelle feuille.

  Il peut s'avérer utile que les producteurs d'items indiquent sur chaque feuille leur nom, la date et l'intitulé de l'examen.

  (Toutes ces précisions peuvent apparaître dans le guide du concepteur; elles sont par ailleurs traitées ultérieurement dans cette partie)

- **Précisions sur les échéances**

  Il est utile que tous les rédacteurs d'items sachent comment leur travail s'intègre au calendrier général de production afin de mettre en évidence l'importance des échéances qui leur sont imposées. Il est bon de leur préciser, lors de la commande, à quel moment la mise en forme aura lieu; on peut alors leur dire si l'on attend d'eux qu'ils y participent ou leur demander s'ils veulent y participer.

- **Précisions sur la rémunération**

  Les conditions financières du travail doivent être claires dès le début. On peut ne rémunérer que le matériel accepté, sans aucune rétribution pour les items rejetés; on peut aussi verser une avance à la commande, le solde étant payé ultérieurement pour tout matériel accepté. On peut également fixer un tarif détaillé selon les différents types d'items ou donner simplement le montant global pour une partie ou l'examen complet. Dans un établissement scolaire, il faudra accorder du temps, dans le cadre de leurs horaires, aux professeurs à qui l'on a demandé de produire des tests.

Lorsque la commande est passée, les rédacteurs recevront les documents suivants:

- le **cahier des charges** ou **spécifications;**
- des **échantillons de matériel** ou des épreuves banalisées;
- des **instructions pour le rédacteur d'items** relatives au test ou à l'épreuve en question.

Pour une expérimentation à grande échelle, il faudra fournir aux rédacteurs différents documents et une information complémentaire, tels que:

- un formulaire d'acceptation de la commande (contrat);
- un formulaire réservant le droit de copyright à l'organisme commanditaire;
- un lexique ou un glossaire des mots et structures utilisables ainsi que des indications sur le niveau;
- un livret d'informations générales sur l'organisme commanditaire.

Pour les tests qui sont diffusés commercialement, il faut prévoir un cahier des charges (ou spécifications) à usage général; il donnera des informations détaillées sur le contenu de l'examen mais n'entrera pas dans les arcanes de la production ni des problèmes qu'elle peut soulever. Il peut cependant y avoir une version plus complète de ce document, en principe confidentielle, qui contienne des lignes directrices et des conseils supplémentaires à l'usage des rédacteurs d'items. Par exemple sur la **sélection** et la **présentation** du matériel; c'est une façon d'éviter une perte de temps aux rédacteurs d'items qui auraient tendance à faire leur propres hypothèses, éventuellement erronées, sur ce qui est acceptable.

Recommandations sur le choix des textes

Selon la définition donnée dans le Chapitre 5 du Cadre de référence, on utilise ici le mot "texte" pour désigner toute manifestation langagière, qu'elle soit écrite ou orale. En conséquence, les recommandations sur le choix des textes s'appliqueront non seulement aux textes écrits mais aussi aux textes oraux utilisés pour la compréhension orale.

Elles recouvriront vraisemblablement les points suivants:

- les meilleures sources de textes (par exemple, articles de journaux de qualité, prospectus);
- les sources qui ne fourniront pas nécessairement des textes acceptables (par exemple, publications spécialisées);
- une mise en garde pour éviter les dérives culturelles;
- une liste des raisons de rejet de textes dans le passé.

Parmi ces raisons:

- un présupposé trop important sur la culture générale ou spécifique des candidats (à moins que ce ne soit l'objet de l'évaluation);
- des sujets inappropriés tels que la guerre, la mort, la politique, la religion qui peuvent bloquer ou choquer certains candidats;
- des sujets qui ne correspondent pas à l'âge des candidats;
- un niveau trop élevé de vocabulaire ou de notions;
- des erreurs techniques ou de style ou des particularités de langue;
- une mauvaise présentation du texte original.

On peut également donner une liste de sujets si souvent traités qu'il vaut mieux les éviter.

En ce qui concerne la recherche de textes, les Chapitres 4 et 7 du Cadre de référence apportent une aide considérable pour situer les textes proposés dans le cadre de la conception générale de l'apprentissage des langues qui est celle du Conseil de l'Europe. La liste des supports de la Partie 4.6.2 (voix en direct, téléphone, radio, etc.) ainsi que celle des types de textes écrits en 4.6.3 fournissent des aide-mémoire utiles et des occasions de varier les types d'items:

Recommandations pour la présentation

Elles couvriront vraisemblablement les points suivants:

- hauteur d'interligne des textes dactylographiés;
- informations à donner en tête de chaque page;
- photocopie ou original des textes;
- références des textes (par exemple, date de publication).

Recommandations détaillées pour chaque question

L'exemple ci-dessous peut illustrer ce point. La tâche à produire est un texte lacunaire sur des éléments grammaticaux plutôt que lexicaux. On donnera les conseils suivants au rédacteur de l'épreuve:

- On trouvera un document authentique d'environ 200 mots avec un titre court. L'accent est mis sur les mots grammaticaux isolés. Le texte ne doit pas comporter trop de vocabulaire peu ou pas connu.

- On produira un minimum de 16 items, plus si possible, afin d'opérer un choix après expérimentation. Le premier item sera utilisé comme exemple et numéroté zéro (0). Les items testeront des pronoms, des prépositions, des conjonctions, des auxiliaires, etc. Ils seront répartis régulièrement dans le texte et l'on veillera à ce que l'incapacité à répondre à l'un d'entre eux n'entraîne pas automatiquement une erreur pour les autres (indépendance des items).

- On évitera d'enlever le premier mot d'une phrase ou une locution figée pour laquelle les candidats ne sauraient pas si elle compte pour un ou plusieurs mots. De même, on évitera une lacune qui ne change pas la grammaticalité d'une phrase (par exemple "tous" dans: *On nous a dit que tous les trains seraient en retard*), ainsi que les items portant sur des structures très peu courantes ou idiomatiques.

On donne également la consigne qui accompagne habituellement ce type de tâche afin d'aider le rédacteur de l'épreuve.

En possession de tous les conseils et informations disponibles, le rédacteur n'a plus qu'à produire le matériel demandé pour l'échéance annoncée. Les rédacteurs expérimentés d'items liés à des textes prennent souvent l'habitude de collecter constamment des textes dans des publications adéquates en vue des commandes futures; lorsque celles-ci arrivent, ils choisissent dans leur stock les textes les plus productifs. Il est recommandé d'avoir un dictionnaire et des ouvrages de référence sous la main pour rédiger certains types d'items (par exemple, ceux qui portent sur la grammaire ou le vocabulaire). Pour produire des épreuves de compréhension orale, il faudra travailler avec un lecteur de cassettes afin de rédiger les items en fonction de l'oral et non de la transcription écrite.

De nombreux rédacteurs de tests trouvent utile de tester leur matériel sur un collègue ou un locuteur fiable non impliqués dans l'évaluation. C'est un moyen de repérer les coquilles, de relever les consignes opaques,

les clés et corrigés erronés et les items trop difficiles ou ceux pour lesquels plusieurs réponses correctes sont possibles.

Le cahier des charges devrait aussi comporter une liste de contrôle que le rédacteur puisse utiliser pour vérifier le texte, les items et la tâche dans son ensemble avant de les soumettre. La liste de contrôle correspondant au texte lacunaire décrit ci-dessus est proposée ici comme exemple. On doit pouvoir répondre "oui" à chaque question si le texte, les items et la tâche sont adéquats.

Texte:

Le texte est-il accessible et culturellement acceptable, etc. ?
Le texte est-il à un niveau de difficulté convenable ?
Est-il approprié pour une tâche centrée sur la structure ?
Est-il assez long pour permettre un minimum de 16 items ?
Lui a-t-on donné un titre convenable ?

Items:

A-t-on produit le nombre d'items exigé ?
Couvrent-ils l'ensemble du texte ?
Leur variété est-elle suffisante ?
A-t-on vérifié que tous les items ont une fonction syntaxique ?
S'est-on assuré qu'ils ne sont pas interdépendants ?
En a-t-on prévu deux ou trois de plus ?
A-t-on évité les items trop idiomatiques ?

Consigne et corrigé:

La consigne a-t-elle été vérifiée ?
A-t-on donné un exemple zéro ?
A-t-on donné un corrigé complet sur une feuille à part ?

Avant de soumettre leur matériel, les rédacteurs de l'épreuve devront en faire une copie complète; s'ils remettent les originaux des articles de journaux ou de magazines au coordinateur, il serait judicieux qu'ils en gardent une photocopie dûment référencée.

2.3.2 Contrôle/révision et mise en forme

Lorsque tous les rédacteurs ont soumis le matériel commandé, on doit d'abord décider ce qui sera immédiatement rejeté, ce qui fera l'objet de révisions et ce qui donnera lieu à une suite. Cette étape est celle du **contrôle**. C'est souvent le coordinateur qui en est responsable, quelquefois assisté d'un autre producteur d'items expérimenté, et c'est à ce niveau que les textes jugés vraiment inacceptables pour l'une ou l'autre des raisons mentionnées plus haut seront rejetés. Si l'on a commandé des textes *sans* items, c'est le moment pour les rédacteurs d'items de se mettre au travail sur les textes acceptés au contrôle. Les producteurs d'épreuves à qui l'on a demandé de soumettre des textes sans items doivent être encouragés à proposer au moins une ébauche des items qu'ils ont l'intention d'écrire de sorte que, aussitôt le texte accepté, ils puissent passer à la production.

Le matériel prêt pour la **mise en forme** peut être étudié par un groupe de travail constitué de rédacteurs et animé par le coordinateur ou un producteur d'épreuves expérimenté. C'est le coordinateur qui décidera:

- de l'organisation des groupes de travail pour la mise en forme;
- du matériel confié à chaque groupe.

Idéalement, le matériel en question sera soumis à l'avance à chaque membre du groupe qui pourra ainsi en prendre connaissance avant la réunion. Lorsqu'il s'agit d'items qui s'appuient sur des textes, on recommande de lire les items *avant* le texte; on repérera ainsi ceux auxquels on peut répondre *sans* se référer au texte (c'est-à-dire par simple bon sens ou culture générale). Ensuite, on travaillera sur les items comme si on passait le test, ce qui permet d'identifier ceux pour lesquels il y a plus d'une réponse juste possible, ou ceux pour lesquels elle est mal formulée et peu claire, ou encore si l'un des distracteurs est si improbable qu'aucun candidat qui le comprendra ne le choisira ou enfin ceux qui sont difficiles ou opaques même pour un locuteur compétent. Pour les tests de compréhension orale ou écrite, on vérifiera leur durée ou leur longueur et que le sujet, le style et le niveau de langue sont convenables. Ce matériel, distribué pour préparation avant la réunion est, de toute évidence, confidentiel.

Au cours de la réunion elle-même, tout problème relevé dans le matériel sera posé et discuté en détail par le groupe. Il est rare que le matériel soumis soit accepté tel quel et, même retenu, il subira probablement des modifications durant la réunion de mise en forme. On portera aussi une attention toute particulière à l'adéquation des consignes et des clés ou corrigés. Le matériel fait généralement l'objet d'une discussion nourrie et il faut que les rédacteurs puissent accepter et formuler des critiques constructives, ce qui s'avère quelquefois difficile. Lorsqu'un producteur d'items se trouve dans la position de défendre et d'expliquer certaines de ses propositions à des collègues expérimentés, il est probable qu'elles ont quelque faiblesse. Lorsque la discussion a été suffisante, il est utile que le coordinateur, ou toute autre personne ayant de l'autorité sur le groupe, soit en mesure de trancher en fin de compte. Chaque groupe de travail doit avoir un rapporteur qui note toutes les décisions prises avec précision et de manière détaillée et rende clairement compte de toute modification. On peut former des rédacteurs nouveaux au travail de révision et mise en forme en les plaçant dans un groupe expérimenté. Un groupe de plus de quatre ou cinq personnes risque d'être assez lent; en revanche, à moins de trois, la variété des points de vue sera peut-être insuffisante.

À l'issue de la réunion, il ne saurait subsister aucun doute quant aux changements décidés. C'est pourquoi on doit garder un compte rendu clair des modifications apportées au matériel accepté. Il arrive que des propositions initiales soient potentiellement intéressantes mais les modifications à y apporter sont trop importantes pour être faites en réunion. On peut alors les rendre à leur rédacteur ou les confier à un producteur d'épreuves expérimenté pour révision et mise en forme. Pour raisons de sécurité, après la réunion, on détruira toutes les copies de travail et les exemplaires supplémentaires du matériel préparé. C'est le coordinateur qui garde les exemplaires révisés du matériel accepté.

Les producteurs d'épreuves sont en droit d'attendre du coordinateur une explication sur le matériel refusé, notamment s'ils n'ont pas participé à la révision ou étaient absents lors du traitement de leur propre matériel. C'est un moyen d'éviter le renouvellement des mêmes erreurs.

Les utilisateurs du Guide qui sont engagés dans l'organisation du processus d'élaboration envisageront et expliciteront selon le cas:

- *comment le processus d'élaboration sera organisé dans leur situation propre, c'est-à-dire les horaires et le calendrier, le personnel, la procédure, etc.*
- *à qui seront commandées les épreuves*
- *quel niveau de connaissance du contenu et d'expérience est exigé*
- *quelle formation et/ou quels conseils recevront les rédacteurs*
- *qui participera au processus de contrôle/révision et de mise en forme des épreuves*
- *comment sera organisé le processus de contrôle/révision/mise en forme*

2.4 Pré-test et expérimentation

Le pré-test et l'expérimentation supposent également que l'on essaie le matériel d'évaluation sur un échantillon représentatif du groupe de candidats afin de recueillir diverses informations sur leur performance et les caractéristiques de la mesure. Le **pré-testage** (ou pré-testing) est la dénomination courante de cette activité mais on utilise aussi ce mot plus particulièrement pour désigner les cas où le matériel d'évaluation est administré à un groupe important de candidats afin de mener à bien un ensemble d'études statistiques sur les résultats obtenus. L'**expérimentation** correspond souvent à une forme de pré-testage qui n'implique que de petits groupes de candidats mais peut renvoyer un feed-back utile sur différents aspects de la performance du matériel d'évaluation.

Les types d'items normalement pré-testés sont les plus objectifs tels que QCM et textes lacunaires. À la suite des étapes de révision/contrôle et de mise en forme, le pré-testage permet un contrôle supplémentaire plus objectif pour vérifier qu'un item fonctionne assez bien pour entrer dans un examen en grandeur nature. On teste les items en tant que tels et non le test dans son ensemble; ainsi un pré-test n'a pas à ressembler trait pour trait à l'examen réel pour lequel le matériel a été produit, ni dans sa longueur ni dans sa composition.

Les épreuves soumises à expérimentation sont présentées comme examen blanc dans une simulation d'examen à des étudiants dont les enseignants estiment qu'ils se trouvent au niveau convenable pour s'y présenter. En le passant, ils bénéficient d'une pratique des épreuves et d'un feed-back sur leur performance fondé sur les résultats qu'ils obtiennent. Afin de réaliser les études statistiques nécessaires et d'avoir des résultats fiables, on recommande une population de 100 à 150 candidats ou plus. L'expérimentation constitue une alternative convenable au pré-testage lorsque ce dernier est irréalisable.

On ne peut pré-tester de la même façon des tests de production orale ou écrite notés subjectivement car il n'y a pas une seule (ou un nombre limité de) réponse(s) juste(s). Malgré tout, on peut vérifier le fonctionnement des tâches avant de les inclure dans un examen. On peut également les expérimenter auprès d'étudiants qui sont au niveau adéquat et les réponses obtenues peuvent être notées par les examinateurs suivant les critères qui seront appliqués à l'examen réel. Ce type d'expérimentation révélera au coordinateur si les étudiants ont compris la tâche, si elle convenait à leur niveau d'expérience et à leur groupe d'âge, s'ils avaient assez d'information pour l'exécuter convenablement et si elle leur a donné l'occasion de manifester les connaissances discursives, syntaxiques et lexicales attendues d'un candidat se présentant à un examen à ce niveau.

Le pré-testage à grande échelle et l'expérimentation à petite échelle permettent également de recueillir des informations importantes sur les aspects pratiques de la passation ainsi que sur les réactions des candidats au matériel d'évaluation.

L'analyse statistique des résultats apporte au coordinateur des informations extrêmement utiles sur la productivité des items et peut éviter que l'on n'inclue des items erronés ou de qualité médiocre dans un examen en grandeur nature. Toutefois il ne faut pas oublier qu'un item de qualité médiocre peut toujours avoir un rendement statistique acceptable; c'est pourquoi on ne considérera les résultats de ce type d'analyse que comme un facteur parmi d'autres déterminant ce qui sera utilisé en fin de compte. L'Annexe 1 présente les résultats d'une analyse d'items accompagnée d'un commentaire explicatif.

Les utilisateurs du Guide qui sont engagés dans l'élaboration de tests envisageront et expliciteront selon le cas:

- *dans quelle mesure ils sont en situation de pré-tester ou d'expérimenter leur matériel d'évaluation*
- *quelles peuvent être les conséquences d'une absence d'expérimentation et comment y remédier*
- *quel type d'analyse subiront les données sur la performance recueillies par le pré-testage et/ou l'expérimentation*
- *comment seront utilisés les résultats de toute analyse, par exemple en vue de l'élaboration de matériel, pour la formation des rédacteurs, etc.*

2.5 Elaboration des tests

De toute évidence, la production du matériel est une activité clé dans l'élaboration des épreuves pour s'assurer qu'elles sont conformes aux normes de difficulté, de contenu et de couverture (linguistique et/ou culturelle). La manière d'aborder la production, la nature de l'information recueillie et le niveau de détail, ainsi que la façon d'enregistrer cette information, peuvent varier d'un test à l'autre. Une seule personne dans une institution donnée peut entreprendre la production de certains tests; d'autres tests exigeront la mobilisation d'une équipe constituée de membres de l'institution et de gens de l'extérieur dont certains peuvent jouer le rôle de consultants.

L'étape de production suppose que l'on prenne en compte un certain nombre de variables différentes qui doivent s'équilibrer pour que soit produit un test au niveau et au contenu exigés et couvrant ce que l'on veut évaluer. On peut fixer certaines caractéristiques d'un test (par exemple, le nombre d'items et de tâches) tandis que d'autres resteront souples (par exemple, le thème ou des accents différents). Si l'on dispose de données expérimentales, elles seront naturellement versées dans le processus d'élaboration. On veillera, dans la plupart des tests, à l'équilibre entre:

- le niveau de difficulté (en termes de difficulté moyenne des items et des tâches du test et de l'étendue de la difficulté couverte);
- le contenu (en termes de sujets ou de domaines):
- la couverture (en termes de représentativité des tâches et de la centration du test);

- la progression (en termes de progression de la difficulté du test);
- de types d'items ou de tâches (en termes de fonctionnements cognitifs variés demandés aux candidats).

Des réflexions particulières s'appliquent à certains tests. Par exemple, pour un test de compréhension écrite qui comprend plusieurs textes et items, il faudra éventuellement vérifier que le même sujet ne soit pas traité plusieurs fois ou éviter que les textes ne soient trop longs en nombre total de mots. De même, pour un test de compréhension orale, on veillera à l'équilibre des voix d'homme et de femme et à celui des accents régionaux.

Lorsque l'élaboration du test est achevée, il est bon d'en faire faire un contrôle indépendant. Il peut être assuré par un consultant extérieur qui connaisse la forme générale de ce type de test mais qui n'ait pas été impliqué dans l'élaboration de celui en question; on peut solliciter ses commentaires sur des points relatifs à l'adéquation du contenu, à la cohérence de la présentation, etc. Si l'on soumet le test à un lecteur extérieur, ce dernier pourra donner un feed-back utile sur la clarté des consignes, la mise en page, etc.

Il est important d'enregistrer précisément toutes les décisions prises pendant l'étape d'élaboration du test; on peut utiliser une grille d'analyse pour saisir l'information descriptive, les données pertinentes du pré-test, la nature de toute modification du matériel et la justification de toutes les décisions prises. À ce niveau, il faut aussi regarder de très près les consignes et la numérotation ainsi que l'établissement d'un corrigé complet et d'un barème.

Lorsque les tests appartiennent à un examen plus important ou à une série d'examens, la phase d'élaboration doit prendre l'ensemble en compte et pas seulement les épreuves ou parties isolées. Il est important d'avoir une vue exacte de l'ensemble d'un examen donné et d'avoir les moyens de comparer des versions parallèles au même niveau ainsi qu'à des niveaux différents et à travers des passations différentes. Une réunion de synthèse donnera une bonne occasion d'échanger des informations transversales sur les épreuves et les examens et permettra d'avoir une vue d'ensemble cohérente de la qualité de l'examen suffisamment tôt pour apporter les modification de forme ou de contenu qui s'avéreraient nécessaires.

Les utilisateurs du Guide qui sont engagés dans l'élaboration de tests envisageront et expliciteront selon le cas:

- *qui, dans leur propre situation, sera impliqué dans l'élaboration des tests*
- *quelles sont les variables à prendre en compte et à équilibrer (par exemple, le niveau de difficulté, le contenu thématique, la gamme de types d'items, etc.)*
- *quel rôle aura l'analyse statistique, par exemple pour la définition de l'indice moyen de difficulté et de l'étendue du test*
- *quelle sera l'importance de l'analyse statistique par rapport aux autres considérations*
- *si le test élaboré devra faire l'objet d'un contrôle indépendant*
- *comment le test élaboré sera apparié à des versions parallèles du même test ou s'intégrera dans une série plus importante de tests*
- *comment sera saisi le profil descriptif du test élaboré, par exemple relevé du contenu thématique, types d'items et de tâches, caractéristiques de mesure, etc. transversalement dans l'ensemble du test.*

2.6 Problématique de la production des items

Dans cette partie, nous traiteront certains des problèmes qui surgissent lors de la rédaction des items et nous proposerons quelques lignes directrices afin d'aider concrètement les rédacteurs de tests. Les problèmes considérés ici sont:

- **la planification de la tâche;**
- **le choix des textes (authenticité, difficulté, etc.);**
- **le choix des types d'items;**
- **les consignes;**
- **les grilles ou clés de correction et corrigés, les barèmes et les échelles de notation.**

Une fois encore, on trouvera dans les Chapitres 4 et 7 du Cadre de référence de précieuses indications sur ces points.

2.6.1 Planification de la tâche

Il est important de noter tout d'abord que le type de tâche doit être conçu en fonction du type de capacité langagière que l'on teste et du but du test.

Lorsqu'on produit du matériel d'évaluation, il est essentiel de relier de façon appropriée le stimulus et la réponse, sinon il est probable que des difficultés surgiront. Par exemple, il est possible de rédiger des items qui s'appuient sur un texte et auxquels on puisse apporter une réponse correcte sans avoir compris le texte. Un stimulus peut provoquer une réponse "correcte" sans que l'on ait testé quoi que ce soit d'utile. De même, il se peut qu'un stimulus se prête facilement à un certain item mais que cet item ne corresponde pas à l'objectif du test.

On ne peut pas présumer simplement que la difficulté d'un item résulte de la relation linguistique entre le texte et la réponse. Le stimulus et la réponse ont aussi leurs caractéristiques linguistiques et la tâche qui les relie peut, outre la demande en langue, entraîner des opérations cognitives complexes. La culture générale (ou connaissance du monde) aura aussi un rôle à jouer ainsi que d'autres aspects du modèle d'utilisation de la langue proposé dans le Cadre de référence. Quand il aborde la rédaction d'items, le rédacteur doit avoir une idée claire de l'objectif d'un item, de la raison pour laquelle ce type d'item a été choisi et des domaines de la compétence du candidat que teste chaque item. Les Parties 7.2 et 7.3 du Cadre de référence examinent relativement en détail de quelle manière les compétences, les caractéristiques et les stratégies de l'apprenant interagissent avec les conditions et les contraintes pour affecter la productivité de la tâche et, notamment, la difficulté de la tâche.

Un test doit comprendre un certain nombre de tâches. Les types de tâches le plus étroitement contrôlés (ceux, par exemple, utilisés pour tester la compréhension écrite, la compétence grammaticale, la compréhension orale et la production écrite au niveau de la phrase) se composent des éléments suivants:

- une **consigne** (ou instructions pour réaliser la tâche);
- un **support** qui sert d'appui à un **stimulus** (un texte, par exemple);
- la **réponse du candidat** aux items de types différents (qu'ils soient choisis ou produits);
- un **corrigé**, une clé ou un **barème de notation**.

On peut distinguer les types de tâches fondées sur les items de celles mises en œuvre dans les épreuves de production orale ou écrite qui comprennent une consigne, un support et une réponse évaluée sur une échelle de notation ou un ensemble de critères et non sur une clé ou un barème.

2.6.2 Choix du texte

Dans la préparation d'un matériel d'évaluation, les producteurs d'items doivent affronter la tâche de sélection des textes, notamment pour les épreuves de compréhension orale ou écrite, et nous examinerons ici un certain nombre de points importants qui gouvernent ces choix. Lorsqu'on sélectionne des textes pour une tâche donnée, il est essentiel en premier lieu d'utiliser des textes convenables pour l'objectif de l'évaluation et la population de candidats concernée. Le niveau de difficulté de la langue doit être adéquat et le thème approprié pour le groupe d'âge prévu et la formation antérieure des candidats. En règle générale, il vaut mieux éviter les sujets qui n'appartiennent pas à l'expérience des candidats ou qui, pour une raison quelconque, peuvent les perturber ou les blesser. Le Cadre de référence apporte une contribution inestimable à tout examen de ce sujet car il sera de plus en plus difficile que les tests évitent le débat qui progresse sur une évaluation au niveau européen, et ceci malgré le souci de s'adapter aux conditions locales. On trouvera dans les Chapitres 4 et 7 du Cadre de référence les parties les plus pertinentes sur le choix des textes. La Partie 4.6 propose une liste utile d'exemples de types de textes et de leurs supports; la Partie 4.6.4 traite plus étroitement de la nature et de la fonction des textes en relation avec les activités et le support.

Deux points relatifs au choix des textes méritent qu'on s'y arrête. Il s'agit d'une part de l'authenticité et, d'autre part, de ce qui rend un texte difficile.

Authenticité

C'est un point sur lequel on débat depuis les années 70 lorsqu'il s'agit de choisir des textes, que ce soit pour l'enseignement ou pour l'évaluation. Dans un examen qui comporte un texte (en compréhension écrite par exemple) répond-on mieux aux besoins du candidat en utilisant un document authentique tiré d'un journal ou d'un magazine, ou un texte fabriqué par le concepteur du test ou le rédacteur d'items ? *A priori*, le document authentique est plus approprié. Il reflète l'usage courant de la langue et a été écrit pour le locuteur natif et non dans un but d'évaluation. On peut avancer que l'interprétation des textes qu'un locuteur natif est capable de faire constitue le but de l'apprenant en langue cible et que c'est donc le type de discours auquel il doit être exposé et sur lequel il sera testé. On peut ajouter qu'un texte écrit dans le seul but de tester certains points de langue ne ressemble en rien à la langue utilisée par les locuteurs natifs qui ne se préoccupent pas d'évaluation.

Cependant, la définition de l'authenticité peut être plus large que cela. On a pu affirmer que l'authenticité est la résultante de l'interaction entre le lecteur et le texte et pas seulement une caractéristique du texte. Si l'on tient compte de cette vision des choses, même un coup d'œil rapide à la grande variété de discours utilisés dans la presse conduit à conclure que tous les textes écrits ne sont pas authentiques pour tous les lecteurs. L'identité du lecteur, son projet de lecture, l'intention du scripteur et le degré de proximité sociale et culturelle entre le lecteur et le texte ont une incidence sur la nature de l'interaction entre un lecteur et un texte donné. S'il y a peu de points communs entre le contenu factuel et culturel du texte et les connaissances du lecteur (imaginer un vieil amateur d'opéra essayant de lire un magazine de rock pour adolescents !), il n'y aura guère d'interaction possible. En tant que locuteurs natifs, nous choisissons plus volontiers les textes qui répondent à nos besoins et à nos intérêts et évitons les autres.

Comment le rédacteur d'items peut-il puiser des textes dans les sources courantes que sont les journaux et magazines, en étant sûr qu'ils conviennent aux apprenants d'une langue qui ne sont peut-être jamais allés dans aucun des pays où on la parle et dont on ne sait pas s'ils partagent le moindre savoir culturel ou social avec les locuteurs natifs pour qui l'article a été écrit ? De toute évidence, il ne suffit pas de découper des articles ou des publicités dans des journaux et de considérer qu'ils seront utiles pour l'enseignement de la langue, simplement parce qu'ils proviennent de sources authentiques au lieu d'avoir été créés par les

évaluateurs. Parce qu'il ne dispose pas du savoir partagé que l'on supposait chez le lecteur cible originel, l'apprenant en langue est renvoyé à une interprétation du texte au mot à mot; l'expérience de la lecture risque alors d'être faussée et artificielle. Cependant, il faut un lien entre les activités d'évaluation et les situations communicatives d'une part et, d'autre part, les tâches langagières réelles dans lesquelles le candidat compte être capable d'utiliser la langue et que l'évaluateur souhaite généraliser. Se pose aussi la question de la validité apparente, le degré auquel le matériel d'évaluation paraît convaincant aux utilisateurs du test comme illustration du type de discours auquel ils veulent accéder.

Depuis la fin des années 70, la notion d'authenticité a été largement explorée afin de mettre en place une approche argumentée de l'utilisation des textes pour l'enseignement et l'évaluation des capacités langagières. Widdowson (1978) et Bachman (1990) ont mis en évidence deux niveaux d'authenticité: **situationnelle** et **interactionnelle**.

i. authenticité situationnelle

L'authenticité situationnelle peut se définir comme le degré auquel les caractéristiques d'un mode d'évaluation d'une activité langagière reflètent celles des situations de la vie réelle dans lesquelles la langue est utilisée.

Pour concevoir une activité authentique en termes situationnels, il faut d'abord identifier les traits significatifs qui définissent la tâche dans le domaine d'utilisation de la langue cible. On est alors en mesure de produire les activités d'évaluation qui incluent ces traits significatifs.

ii. authenticité interactionnelle

L'authenticité interactionnelle peut se définir comme l'interaction entre l'activité d'évaluation (la tâche) et le candidat; elle suppose que les rédacteurs et les concepteurs de tests devraient:

- proposer des textes, des situations et des tâches qui simulent la "vraie vie" sans essayer de la reproduire à l'identique;
- essayer de proposer des situations et des tâches qui ont des chances d'être pertinentes pour le candidat potentiel à un niveau donné;
- clarifier la *finalité* de chaque tâche ainsi que le *public cible* en mettant en contexte adéquat;
- expliciter les *critères de réussite* de la tâche.

Il est donc important, au moment du choix des textes et de la conception des items, de se pencher sur l'authenticité situationnelle des tâches et de voir si les opérations sur les textes que l'on demande aux candidats d'exécuter correspondent à ce que l'on ferait naturellement de ces textes. On ne sait généralement pas assez comment les gens lisent ou écoutent pour être sûr de l'authenticité d'un test dans ce sens. Néanmoins, il est souvent possible de distinguer si l'appariement du texte et des items est inadéquat ou trompeur pour les candidats. Il faut que les rédacteurs de tests soient sensibilisés aux éventuels problèmes et en aient conscience au moment de la préparation du matériel. Le traitement de cette question en Partie 7.3 du Cadre de référence est particulièrement pertinent.

La difficulté des textes

Le second point important à prendre en considération est la difficulté du texte et les divers caractères qui peuvent en influencer la difficulté. Que l'on parle de textes écrits ou de document oraux, différents facteurs peuvent affecter le degré de difficulté que le lecteur ou l'auditeur rencontre en les traitant; et il en est ainsi pour tous les lecteurs ou auditeurs, qu'ils soient ou non dans la situation de candidat à un examen.

Il est clair qu'une partie de la difficulté provient de la structure linguistique. Par exemple, on entrera plus aisément dans un texte composé de courtes phrases simples à la forme active que dans un texte composé de phrases longues et complexes avec une utilisation fréquente du passif.

Au-delà des traits de la structure linguistique, d'autres facteurs qui concernent le contexte dans lequel le texte se trouve ont une incidence sur son degré de difficulté. Que le texte soit écrit ou parlé, il sera plus facile à comprendre s'il s'adresse directement au lecteur ou à l'auditeur plutôt que de le mettre en position de troisième partie, de spectateur qui observe les échanges entre les personnages principaux. Le support visuel fourni par des dessins ou des schémas (ou par la vidéo dans un test de compréhension orale) peut faciliter la compréhension, de même que l'absence de pression à traiter le texte en temps limité. Si le texte est en contexte de "rupture d'information" et donne ainsi aux candidats une raison impérative d'en extraire des informations, la situation en sera sans doute facilitée; autrement dit, la stimulation de l'intérêt du lecteur ou de l'auditeur peut en augmenter l'accessibilité.

Certains traits du contenu d'un texte peuvent aussi avoir un effet sur sa difficulté. On comprendra mieux un récit si les personnages sont peu nombreux et clairement différenciés. Par exemple, l'histoire de deux femmes et de deux hommes, d'âge différent, avec des noms différents et une personnalité nettement marquée sera perçue comme plus facile que celle qui présenterait un plus grand nombre de personnages moins distincts et moins typés. La suite des événements dans un récit est d'autant plus facile à saisir qu'elle est chronologique et sans retours en arrière; si, en outre, ils sont reliés entre eux (dans un rapport de causalité, par exemple), le texte sera plus compréhensible que s'ils n'ont aucun lien. L'auditeur ou le lecteur qui connaît déjà la structure narrative du récit le trouvera moins difficile que celui qui l'ignore.

Enfin, le type d'interaction et la relation qu'il crée entre le texte et l'auditeur ou le lecteur affecte le degré de difficulté du texte. Des textes très officiels, avec un haut degré de formalisme ou, au contraire, très informels, voire familiers, risquent de poser plus de problèmes aux auditeurs ou aux lecteurs que ne le ferait un style relativement neutre ou modérément informel.

Les difficultés mentionnées ci-dessus ont d'autant plus d'importance pour le producteur de tests de compréhension orale; en effet, la relation entre les différentes parties du document, la possibilité d'y revenir et celle de voir le document dans son ensemble comme ont peut le faire à l'écrit, sont ici impossibles. En plus de tenir compte du niveau de difficulté linguistique en termes de complexité de la structure et du lexique, un producteur d'épreuves de compréhension orale devra être attentif aux facteurs suivants pour écrire ou choisir des textes; tous ces facteurs ont un effet sur l'interprétation que l'on demande au-delà du niveau de compréhension simple, et cela a alors un impact sur le niveau de difficulté du texte.

- Le type de discours oral le plus facile à comprendre est le monologue, notamment dans le cas où le locuteur semble s'adresser directement à l'auditeur. Deux voix différentes (homme/femme ou adulte/enfant) viennent ensuite comme facilitateurs. Une conversation entre deux personnes du même âge ou entre plus de deux personnes est souvent plus difficile. Le texte est plus facile à comprendre si les locuteurs jouent des rôles clairement distincts tels que parent/enfant; à l'inverse, une conversation entre des locuteurs ayant un statut semblable comme, par exemple, des collègues de même sexe et de statut proche discutant au sujet du travail est, en général, plus difficile.

- Un texte qui comporte des changements de lieu, de temps et un grand nombre d'événements sera plus difficile que celui dans lequel n'est rapporté qu'un petit nombre d'événements, dans un même lieu et au même moment.

- Si la situation est claire dès le début, le texte sera plus facile à suivre.

- Un texte court et dense, accompagné d'un nombre relativement élevé d'items, est difficile à traiter même si le niveau de langue est adapté.

- L'ajout de matériel redondant tel que explications, reformulations et répétitions aide à baisser le niveau de difficulté.

- La langue familière, avec son débit, ses ellipses et ses formes propres, son absence apparente d'une organisation cohérente et les changements fréquents de tours de parole exige souvent une activité d'écoute plus soutenue qu'une langue plus soignée, en général plus lente, moins hachée et plus proche de l'écrit.

- Un locuteur au débit naturellement lent et à la voix expressive est plus facile à comprendre que celui qui parle vite ou sur un ton monocorde. Un débit régulier ou en relation directe avec la densité d'information du texte est aussi un élément facilitateur.

La Partie 7.3 du Cadre de référence commente dans le détail certaines des caractéristiques d'un texte et des tâches qui l'accompagnent qui peuvent conduire à une difficulté accrue.

2.6.3 Choix des types d'items

L'un des points les plus importants relatif aux types d'items est de savoir lequel est le mieux adapté pour tester une capacité donné dans un test donné. La décision est généralement prise au moment de la conception du test.

On peut classer de différentes façons le grand nombre de types d'items différents utilisés en évaluation des langues. Certains sont considérés comme objectifs en ce sens qu'aucun jugement n'est nécessaire pour les corriger, tandis que d'autres demandent une réponse élaborée et des méthodes subjectives de correction. Certains contrôlent la compréhension, d'autres la production. Certains s'appuient sur un texte, d'autres sont indépendants ou discrets. Même si certains types d'items sont plus utilisés que d'autres, il ne faut pas croire qu'ils sont meilleurs pour autant. Le meilleur critère de mesure de la qualité d'un type d'item est sa pertinence à évaluer la langue dans un but précis et une situation donnée. Le type d'item qu'il faut retenir est celui qui donne le moyen le plus direct de mesurer la réponse voulue.

Il existe quelques règles générales à suivre lors de la production d'un item quel qu'il soit:

- un item doit toujours viser à contrôler une information significative plutôt qu'une information secondaire ou périphérique

- les conventions grammaticales courantes doivent y être respectées

- quand on utilise un nouveau type d'item, il faut toujours l'accompagner d'un exemple, à moins que sa simplicité ne le rende inutile

- lorsque les items s'appuient sur un texte, il ne doit pas être possible d'y apporter une réponse juste seulement fondée sur la culture générale sans avoir lu et compris le texte

- un item qui s'appuie sur un texte devrait être rédigé clairement et simplement afin que ceux qui comprennent le texte n'échouent pas sur une incompréhension de l'item

- les items qui s'appuient sur un texte peuvent être placés avant ou après; on placera plutôt avant ceux qui testent un traitement superficiel du texte, tandis que ceux placés après peuvent exiger une lecture détaillée ou déboucher sur des conclusions.

Un classement possible des types d'items en deux grands groupes consiste à distinguer ceux pour lesquels le candidat doit choisir une réponse et ceux pour lesquels il doit fournir la réponse. On les appellera items à sélection de réponse et items à réponse libre. La forme la plus courante de l'item à sélection de réponse est la question à choix multiple (QCM), bien que d'autres types d'items tels que vrai/faux et différentes sortes d'appariements puissent être regroupés avec les QCM car ils demandent au candidat le même type d'activité. En règle générale, du point de vue de la notation, les tests qui reposent sur des QCM sont considérés comme plus objectifs que ceux pour lesquels le candidat doit fournir la réponse. Dans l'idéal, on ne devrait pas utiliser de QCM sans les pré-tester et les analyser. La Partie 7.3 du Cadre de référence traite de nombreux points à prendre en compte lorsqu'on choisit des modalités de réponse différentes pour des activités d'évaluation.

Il est important de réaffirmer qu'aucun type d'item n'est, en soi, plus ou moins utile qu'un autre. Le choix d'un item approprié dépend de l'objectif particulier du concepteur de test et de ses priorités. Par exemple on peut, pour l'évaluation de la production écrite et orale, mettre en place des tests basés sur des items ou des tâches plus globales. La production orale et la production écrite peuvent être divisées en sous-savoir-faire intitulés "grammaire", "vocabulaire", "orthographe", "prononciation", etc. Considérée au niveau de ces éléments discrets, la compétence de production orale ou écrite pourrait se mesurer à l'aide d'épreuves fondées sur des items et intitulées soit "écrit", soit encore "grammaire et usage" ou "compétence structurale". En revanche, la production orale ou écrite qui suppose l'organisation des idées et des arguments, l'articulation, le plan et la construction d'un récit cohérent doit être contrôlée par des tâches qui ne reposent pas sur des items discontinus. L'analyse de la compétence générale et langagière des utilisateurs/apprenants dans le Chapitre 5 du Cadre de référence, ainsi que la présentation du processus de la compétence communicative en 4.5, proposent un paradigme utile dans le cadre duquel on peut replacer l'analyse de l'adéquation des types d'items.

2.6.4 Consignes

On peut définir la consigne comme "les instructions données aux candidats pour réagir devant un support donné". Ces instructions doivent préciser où et comment la réponse sera enregistrée, par exemple: cocher la réponse qui vous paraît appropriée ou écrire quelques mots, et si cela doit être fait sur la copie elle-même ou sur une feuille de réponse à part. Les consignes sont importantes car elles disent au candidat ce qu'il doit faire et comment; c'est pourquoi elles doivent être rédigées avec soin. La Partie 7.3 du Cadre de référence examine l'importance de l'aide donnée pour exécuter une tâche, en fonction des conditions et des contraintes que l'on peut manipuler pour les tâches productives comme pour les tâches réceptives.

La consigne doit présenter le plus clairement possible la tâche que l'examinateur donne au candidat. Il ne doit y avoir ni ambiguïté ni besoin de clarification qui pourrait perturber le candidat: l'angoisse d'un candidat peut porter atteint à sa performance et affecter la fidélité et la validité du test. Souvent, pour les épreuves de compréhension orale, la consigne est non seulement imprimée sur la feuille de réponse mais également enregistrée sur la cassette. Dans les épreuves de production orale en situation d'entretien en face à face, les conditions de passation sont assez différentes des autres; en effet, plutôt qu'une consigne, c'est l'examinateur/interlocuteur/interviewer qui donne des instructions orales. Le candidat peut même solliciter une clarification de la tâche à exécuter, demande que l'on peut évaluer comme faisant partie des échanges.

Voici, par exemple, la consigne d'une épreuve de transfert d'information:

> Pour les items 1 à 8, lisez la note officieuse qu'un(e) collègue vous a communiquée. En utilisant les renseignements qu'elle donne, complétez l'annonce officielle en écrivant les mots manquants sur votre feuille de réponse. Les mots dont vous avez besoin n'apparaissent pas dans la note officieuse. L'exercice commence par un exemple (0). N'utilisez pas plus de deux mots dans chaque espace.

Les questions clés à se poser pour rédiger des consignes sont:

- est-ce clair ? (c'est-à-dire, est-il possible de faire un contresens sur la nature de la tâche ?);
- est-ce facilement compréhensible ? (c'est-à-dire, la langue utilisée est-elle au niveau convenable ? Ceci est particulièrement important pour les tests de débutants);
- est-ce adéquat ? (c'est-à-dire, *toute* l'information nécessaire est-elle donnée?);
- est-ce pertinent ? (c'est-à-dire, ne donne-t-on *que* l'information nécessaire ?);
- les consignes sont-elles conséquentes ?

En ce qui concerne ce dernier point on recommandera que la langue utilisée dans les consignes soit normalisée pour un test donné afin que le candidat puisse, autant que faire se peut, suivre des consignes données sur le même modèle.

La consigne est une partie importante de la production d'épreuves et il faut encourager les rédacteurs à mettre autant de soin à la rédaction des consignes qu'à celle des items. Un aide-mémoire semblable au suivant peut s'avérer utile:

- la consigne et les instructions générales pour l'examen sont-elles cohérentes?
- si la consigne est nouvelle pour le candidat, est-elle accompagnée d'un exemple?
- la langue utilisée est-elle grammaticalement correcte et adaptée au niveau de l'examen ? (cela signifie que le niveau de langue de la consigne doit être inférieur à celui de la langue évaluée dans le test);
- le vocabulaire appartient-il aux ressources du candidat ?
- la langue est-elle simple et claire ?
- y a-t-il un discours superflu ?
- y a-t-il des doubles négations ?
- peut-il y avoir malentendu ou équivoque ?
- la consigne contient-elle toute l'information nécessaire et précise-t-elle les contraintes ?

Il est souvent difficile pour un rédacteur de prendre une distance qui lui permette de repérer les difficultés; c'est pourquoi il est utile de demander à un collègue d'expérimenter l'item ou la tâche.

Des détails importants à inclure dans les consignes sont:

- précisément où trouver le support (par exemple, numéro de page);
- le nombre de mots attendus dans la réponse;
- si la même réponse peut être utilisée plusieurs fois;
- si les réponses peuvent être données dans n'importe quel ordre;
- le nombre approximatif de mots pour une production écrite;

- des indications claires sur l'étendue des choix possibles de tâches;
- le nombre d'auditions d'un document sonore;
- les contraintes d'utilisation du texte support;
- l'indication des critères de succès.

2.6.5 Grilles de correction, barèmes et échelles de notation

Toute activité d'évaluation qui utilise des types d'items objectifs doit être accompagnée non seulement d'une consigne convenable mais aussi d'une grille de correction ou clé (c'est-à-dire des réponses correctes) ou d'un barème; dans le cas d'une tâche évaluée plus subjectivement, il faut avoir une échelle de correction, une liste des tâches exigées et des critères de notation.

Le rédacteur fournira toujours une grille de correction lorsqu'il n'y a qu'une réponse possible comme dans les QCM et autres items de ce type. Pour les types de tâches qui demandent de la production plutôt que de la sélection, le rédacteur du test doit fournir un barème et donner une liste aussi exhaustive que possible de réponses acceptables. L'établissement du barème est un moment essentiel de la production de tests car c'est souvent à ce niveau qu'apparaît la solidité ou la faiblesse d'un item et qu'on doit le rejeter ou le récrire alors qu'il avait paru acceptable.

L'aide-mémoire du producteur de tests comportera les questions suivantes;

- a-t-on produit une réponse type appropriée?
- s'il y a plus d'une réponse possible, a-t-on donné toutes les possibilités dans le corrigé?
- la grille ou clé de correction est-elle facile à utiliser?
- le nombre de points pour les réponses justes par item ou par tâche est-il clairement indiqué?
- le nombre de réponses possibles est-il assez réduit?
- toutes les contraintes sont-elles précisées? (par exemple, les candidats doivent choisir 2 options sur 5; l'épreuve ne sera pas notée s'ils en proposent plus de 2).

Il faut aussi savoir si l'orthographe et la grammaire compteront dans une réponse au demeurant correcte. Il ne faut pas oublier qu'un test peut être noté par un non spécialiste de langue; s'il y a trop peu de contraintes sur les réponses possibles, la notation risque de devenir problématique parce que le correcteur ne saura plus ce qui est acceptable ou pas.

Il existe des méthodes de notation différentes pour la production écrite longue ou la production orale. En règle générale, elles mettent en œuvre un barème qui dissocie les capacités testées lors d'un test de production orale en domaines tels que prononciation, aisance et utilisation correcte des structures et note chacune sur une échelle. Pour aider l'examinateur, on lui fournit l'exemple d'une performance type à chaque niveau. La note finale pour une épreuve de production orale ou pour une composition écrite sera obtenue en attribuant une note sur chaque échelle de telle sorte que la note totale de la tâche sera la somme des notes éclatées. Le Chapitre 3 du Cadre de référence et les annexes proposent une information utile sur les différentes approches d'élaboration des échelles de correction et sur la formulation des descripteurs; le Chapitre 9 examine quelques-uns des problèmes soulevés par l'évaluation subjective, parmi lesquels celui du besoin d'accompagner les descripteurs de performance d'exemples de travaux de candidats qui correspondent à l'échelle des notes et des descripteurs. L'Annexe B du Cadre de référence – Les échelles de démonstration - fournit un premier exemple de la façon de décrire différents niveaux de performance dans des activités communicatives et des capacités langagières, et les Annexes E et D décrivent des projets pour l'élaboration de descripteurs, y compris les « Can Do » du projet ALTE.

Il est bon qu'un rédacteur d'items écrive un échantillon de réponse pour tout item, que ce soit exigé ou pas. Même si le producteur ne donne rien d'autre qu'un titre pour la composition écrite, il est important de vérifier que le sujet peut être traité convenablement dans le nombre de mots indiqué et au niveau de langue des candidats. On peut relever les erreurs de ce type d'épreuve au moment de l'expérimentation mais il vaut mieux faire tout son possible pour les éliminer plus tôt.

Les utilisateurs du Guide engagés dans la conception des épreuves envisageront et expliciteront selon le cas:

- *quelle est la nature, dans leur situation, de la relation entre l'objectif du test et la conception de la tâche*
- *quels types d'items et de tâches seront les plus adéquats pour évaluer les capacités langagières en question*
- *quelle sorte de conseils donner aux rédacteurs d'items sur la sélection des textes, par exemple: sources probables, thèmes inappropriés, etc.*
- *quelles sont les caractéristiques des textes qui peuvent provoquer des difficultés*
- *quelles sont les opérations cognitives exigées par les types d'items et de tâches choisis pour atteindre l'objectif du test*
- *jusqu'à quel point est-il souhaitable de normaliser le discours de consigne*
- *quels sont les types de barème et/ou d'échelle de notation les plus appropriés*
- *comment ces barèmes seront-ils élaborés*

3.0 ÉVALUATION DES TESTS

La validation de l'évaluation fait partie intégrante du modèle de processus d'élaboration de tests. Le cycle d'élaboration commence par une réflexion sur la fonction du produit (c'est-à-dire le but du test); cela doit comporter des réflexions sur la manière d'utiliser le test, sur sa pertinence et son utilité en termes de conséquences sociales et de rentabilité et les effets possibles qu'il risque de provoquer (y compris des retombées imprévues).

Afin d'élaborer et de fournir un test de qualité, il faut mettre en place des systèmes et des procédures, non seulement pour le produire, mais également pour l'évaluer; ces systèmes et procédures doivent entrer en jeu dans les phases d'élaboration comme d'opérationalisation et ont essentiellement pour but de:

- valider le test;
- mesurer l'impact du test;
- fournir une information pertinente aux utilisateurs du test;
- assurer un service suivi de grande qualité.

On s'accorde généralement à dire que l'évaluation a un effet sur le processus éducatif et sur la société en général. Cet effet fonctionne sur deux niveaux au moins en termes de:

i) éducation et société en général;
ii) individus directement concernés par les tests et leurs résultats.

Par principe, les évaluateurs devraient avoir pour but de faire que leurs tests n'aient pas un effet négatif et, autant que possible, de s'efforcer qu'ils aient un effet positif. En termes généraux, on peut y parvenir par l'élaboration et la présentation de spécifications et d'un projet détaillé de programme, ainsi que par l'apport de programmes professionnels d'assistance pour des institutions comme pour des professeurs/étudiants individuellement.

L'effet positif sur l'enseignement et l'apprentissage est un impact important qui fonctionne aussi bien au niveau général que particulier. On peut parvenir à un impact éducationnel positif par les pratiques suivantes:

- l'identification d'experts compétents dans tous les domaines pour travailler sur tous les aspects de l'élaboration de tests;
- la formation et le recrutement d'experts compétents qui deviendront des rédacteurs d'items au cours de l'élaboration d'items;
- la formation et le recrutement d'experts pour jouer le rôle d'examinateurs.

Il est important d'être en mesure d'évaluer l'impact éducationnel que les tests ont dans les situations dans lesquelles ils sont utilisés, et le recueil régulier de données fournit l'essentiel de l'information nécessaire pour analyser l'impact et l'utilité d'un test donné. Il pourrait être souhaitable de recueillir des données relatives à:

- qui se présente au test (c'est-à-dire le profil des candidats);
- qui utilise les résultats du test et dans quel but;
- qui enseigne pour la préparation du test et dans quelles conditions;
- quelles sortes de cours et de matériels sont conçus et utilisés pour préparer les candidats;
- quel est l'effet du test sur la perception du public en général (par exemple en ce qui concerne les normes scolaires);

- comment le test est-il perçu par ceux qui sont directement engagés dans le processus éducatif (par exemple, les étudiants, les candidats, les enseignants, les parents, etc.);
- comment le test est-il perçu par ceux qui n'appartiennent pas au système éducatif (par exemple, les politiques, les hommes d'affaires, etc.)

En résumé, une bonne pratique de l'évaluation repose sur l'adoption d'un modèle de processus d'élaboration de tests puisque c'est ce qui donne les conditions nécessaires à l'élaboration de tests utiles et permet leur validation.

Les utilisateurs du Guide engagés dans l'évaluation envisageront et expliciteront selon le cas:

- *quels seront les systèmes et les procédures nécessaires dans leur situation pour contrôler et évaluer la performance du test une fois qu'il est mis en service*
- *quelles seront les procédures particulières d'analyse les plus appropriées*
- *comment sera évalué l'impact social et éducatif de leur test*
- *quels seront les systèmes et les procédures nécessaires pour maintenir un service de grande qualité, par exemple l'élaboration d'une charte (ou code de déontologie)*
- *comment les utilisateurs des tests recevront-ils l'information pertinente, par exemple, la documentation, les programmes professionnels d'assistance), etc.*

RÉFÉRENCES

ALTE Code of Practice in ALTE Handbook of European Language Examinations and Examination Systems: ALTE 1998

Bachman, L.F.: *Fundamental Considerations in Language Testing*. Oxford University Press, Oxford, 1990.

Canale, M. et Swain, M.: A theoratical framework for communicative competence. *In* A.S. Palmer, P.J. Groot et S.A. Trosper (sous la direction de) *The Construct Validation of Tests of Communicative Competence.* TESOL,Washington, DC, 1981.

Cadre européen commun de référence pour les langues: Apprendre, Enseigner, Evaluer – Editions DIDIER ISBN 227805075-3

van Eck, J.A.: *The Treshold Level in a European unit/credit system for modern language learning by adults.* Conseil de l'Europe, Strasbourg, 1975.

van Eck, J.A.: *Threshold Level English*. Pergamon Press, Londres, 1980.

van Eck, J.A. et Trim, J.L.M.: *Threshold Level 1990*. Conseil de l'Europe, Strasbourg, 1990.

Widdowson, H.G.: *Language Teaching as Communication*. Oxford University Press, Oxford, 1978.

AUTRES SOURCES

Alderson, J.C., Clapham, C. et Wall, D.: *Language Test Construction and Evaluation.* Cambridge University Press, Cambridge, 1995.

Alderson, J.C. et Hughes, A. (sous la direction de) : *Issues in Language Testing*. ELT Documents 111, British Council, Londres, 1981.

Alderson, J.C. et North, B. (sous la direction de) : *Language Testing in the 1990s: The Communicative Legacy.* Modern English Publications et le British Council, Londres, 1991.

Alderson, J.C,, Krahnke, K. et Stansfield, C. (sous la direction de): *Reviews of English Language Proficiency Tests.* TESOL, Washington, DC, 1987.

Bachman, L.F. et Palmer, A.: *Language Testing in Practice.* Oxford University Press, Oxford, 1996.

Barlow, M.: *Formuler et évaluer ses objectifs en formation*. Chronique sociale, 3e tirage, Lyon, 1989.

Bolton, S.: *Évaluation de la compétence communicative en langue étrangère.* CREDIF-Hatier, Coll. LAL, Paris, 1987.

Carroll, B.J.:*Testing Communicative Performance.* Pergamon, Londres, 1980.

Delorme, C.: *L'évaluation en questions.* ESF éditeur, 2e édition, Paris, 1987.

Henning, G.: *A Guide to Language Testing.* Newbury House, Cambridge, Mass.,1987.

Hill, C. et Parry, K.: *From Testing to Assessment*. Longman, 1994.

Lienert, G.A. et Raatz, U.: *Testaufbau und Testanalyse (5. neubearb. und erw. Auflage).* Beltz, Psychologie Velags Union, Weinheim, 1994.

Luissier, D.: *Evaluer les apprentissages dans une approche communicative.* Hachette, Paris, 1992.

Mager, R.F.: *Comment mesurer les résultats de l'enseignement*. Bordas, Paris, 1986.
Underhill, N.: *Testing Spoken Language.* Cambridge University Press, Cambridge, 1987.

Weir, C.: *Communicative Language Testing.* Prectice Hall, 1990.

Weir, C.: *Understanding and Developping Language Tests*. Prenctice Hall, 1993.

ANNEXE 1: ANALYSE D'ITEMS

L'analyse statistique des notes d'un test fournit au concepteur/élaborateur du test des informations utiles sur la productivité des items isolés et permet d'éviter l'utilisation d'items erronés ou de médiocre qualité dans la version finale de l'examen. Toutefois, il est important de se rendre compte qu'un item médiocre peut produire des statistiques acceptables; c'est pourquoi les résultats de ce type d'analyse ne sont que l'un des facteurs parmi d'autres qui permettent de déterminer quels matériels utiliser dans les épreuves d'examen.

Les données recueillies lors du pré-testage peuvent être analysées selon des méthodes statistiques classiques ainsi que selon la méthode de Rasch. Pour une analyse statistique classique, on utilise un logiciel comme MicroCAT. Ce type d'analyse fournit des informations sur la productivité des items traités individuellement, telles que l'**indice de facilité de l'item**, sa **valeur discriminante** et le **pointage des distracteurs.**

Facilité de l'item

Le fait de connaître l'indice de facilité des items permet au concepteur/élaborateur de s'assurer que le matériel d'évaluation est au niveau convenable de difficulté pour les candidats concernés. La facilité est la proportion de réponses correctes à un item, transcrite sur une échelle de 0 à 1 ou exprimée en pourcentage.

Sur le listage présenté dans la Figure 4, l'indice de facilité de chaque item figure dans la colonne "Proportion de réponses correctes". L'item 8, par exemple, a un indice de facilité de 0.38 (ce qui signifie que 38% des candidats au pré-test ont obtenu la note attribuée à cet item). Le niveau convenable pour un test se situe au point milieu de l'amplitude de difficulté mais on peut situer une amplitude acceptable de facilité d'un item de 33 à 67 ou de 20 à 80. En fait, le niveau convenable peut varier d'un test à l'autre en fonction du but du test; un test de niveau de capacité donné à la fin d'un cycle d'études peut exiger un niveau de facilité différent de celui qu'il faut pour un test d'aptitude.

Un test devrait contenir quelques items aux deux extrémités de l'amplitude. Dans certains tests on place quelques items faciles au début pour permettre aux candidats de "s'échauffer"; il arrive que l'on ne compte pas ces items dans la note finale.

Les items qui n'entrent pas dans l'amplitude acceptable sont rejetés à ce niveau; il ne s'agit pas de les gaspiller pour autant. Si une banque d'items existe, on les y déposera et ils pourront servir pour un autre test à un autre niveau.

Figure 4

Feuille de résultats de MicroCAT Analysis
(Statistiques d'items)

MicroCAT (TM) Testing System

Programme d'analyse d'items et de test - ITEMAN (TM) Version 3.50

Statistiques d'items					Autres statistiques					
Numéro d'ordre	Échelonnage de l'item	Proportion de rép. correctes	Indice de discrimination	Coef. de corélation bisériale de point	Options	Proportion totale	Bas	Haut	Coef. de corélation bisériale de point	Clé
8	2-1	.38	.52	.48	A	.00	.00	.00		
					B	.38	.13	.66	.48	*
					C	.12	.11	.12	-.01	
					D	.49	.74	.23	-.44	
					Autre	.01	.00	.00	-.11	
9	2-2	.71	.42	.42	A	.07	.11	.01	-.16	
					B	.11	.18	.04	-.22	
					C	.10	.16	.00	-.22	
					D	.71	.53	.95	.42	*
					Autre	.01	.00	.00	-.13	
10	2-3	.68	.56	.56	A	.68	.39	.96	.56	*
					B	.21	.36	.04	-.37	
					C	.03	.08	.00	-.24	
					D	.07	.14	.00	-.22	
					Autre	.01	.00	.00	-.13	
11	2-4	.57	.49	.49	A	.18	.28	.08	-.27	
					B	.15	.19	.09	-.12	
					C	.08	.16	.01	-.31	
					D	.57	.33	.83	.49	*
					Autre	.01	.00	.00	-.13	
12	2-5	.61	.63	.54	A	.09	.18	.00	-.22	
					B	.20	.28	.03	-.27	
					C	.61	.32	.96	.54	*
					D	.09	.18	.01	-.28	
					Autre	.02	.00	.00	-.09	
13	2-6	.81	.35	.48	A	.11	.20	.04	-.29	
					B	.01	.03	.00	-.11	
					C	.81	.61	.96	48	*
					D	.07	.17	.00	-.34	
					Autre	.00	.00	.00		
14	3-1	.93	.19	.39	A	.93	.81	1.00	.39	*
					B	.07	.18	.00	-.39	
					Autre	.01	.00	.00	-.03	

Valeur discriminante de l'item

Ces statistiques traitent de la capacité de l'item à distinguer les candidats forts des faibles. Ceux dont la note finale est élevée devraient répondre correctement à n'importe lequel des items dans une proportion plus élevée que ceux dont la note est faible. On utilise couramment deux méthodes pour mesurer la valeur discriminante de l'item:

i) l'indice de discrimination;
ii) le coefficient de corrélation bisériale de point.

On peut les trouver dans les colonnes intitulées Indice de discrimination et Coefficient de corrélation bisériale de point dans la Figure 4 (MicroCAT).

i. Indice de discrimination

Une fois qu'un certain nombre de candidats ont passé un test on peut les classer (ou les placer en ordre) selon les notes qu'ils ont obtenues. On peut alors distinguer deux groupes dans l'échantillon: les premiers 30% qui représentent le groupe de capacité élevée, et les derniers 30% qui représentent le groupe de faible capacité.

Le nombre de candidats dans chacun de ces deux groupes est identique et on le représentera par N. On compte le nombre de candidats qui ont réussi à l'item dans chaque groupe pour obtenir:

nH (nombre des candidats appartenant au groupe de capacité élevée qui ont répondu correctement à l'item);

et

nL (nombre des candidats appartenant au groupe de faible capacité qui ont répondu correctement à l'item).

On peut alors définir comme suit l'indice de discrimination di:

$$d_i = \frac{nH - nL}{N}$$

d_i peut prendre n'importe quelle valeur entre -1 et +1.

Un indice de discrimination d_i de +1 indique que tous les "bons" étudiants répondent correctement à cet item et que tous les candidats "faibles" y échouent.

Un indice de discrimination d_i de -1 indique que tous les "bons" étudiants échouent à cet item et que tous les candidats "faibles" y répondent correctement.

Les items dont le d_i est égal ou supérieur à 0.30 sont, en principe, considérés comme appropriés pour ce groupe particulier. Il faut souligner que l'indice de discrimination est en rapport aux capacités de ce groupe donné. Par exemple, l'item 8 a un indice de discrimination de 0.52 ce qui permet de penser qu'il distingue les candidats forts des faibles dans ce groupe, tandis que l'item 14 ne fait qu'une distinction médiocre.

ii. Coefficient de corrélation bisériale de point

La corrélation bisériale de point, $r_{pb,}$ est donnée par la formule suivante:

$$r_{pb} = \frac{\overline{X_p} - \overline{X_q}}{s_X} \sqrt{pq}$$

dans laquelle

$\overline{X_p}$ est la note moyenne totale de tous les candidats qui ont répondu correctement à cet item.

$\overline{X_q}$ est la note moyenne totale de tous les candidats qui n'ont pu répondre à cet item

p est la proportion du nombre total de candidats qui ont répondu correctement à cet item

q est la proportion du nombre total de candidats qui n'ont pu répondre à cet item

s_X est l'écart standard des notes du test pour tous les candidats.

En règle générale, les items dont la valeur du coefficient de corrélation bisériale de point est égale ou supérieure à 0.30 sont considérés comme acceptables. Lorsqu'une corrélation bisériale de point apparaît avec une valeur négative, cela signifie que les candidats forts n'ont pas su choisir la bonne réponse à cet item. Cela peut faire penser qu'une option autre que celle prévue comme correcte peut apparaître très légitimement comme juste; on appelle distracteur positif une option de ce type. On ne peut pas utiliser un tel item dans un test mais il est possible de le réviser en enlevant le distracteur positif et en pré-testant de nouveau.

Pointage du distracteur
L'analyse statistique des items à choix multiple indiquera si les distracteurs fonctionnent ou non de manière adéquate, autrement dit, si chacun est assez plausible pour attirer quelques candidats mais pas si près de la réponse juste qu'un plus grand nombre de candidats le choisissent au lieu de choisir la clé (la réponse juste)

La feuille de résultats du traitement MicroCAT montrera la proportion de candidats qui choisissent chacun des distracteurs dans la colonne "Proportion totale". Regardons, par exemple, l'analyse suivante d'un item à choix multiple à 4 options dont la clé est C:

A	.15
B	.10
C	.63
D	.12

Dans ce cas précis, les statistiques révèlent un item où la clé et les autres distracteurs fonctionnent tous de manière satisfaisante. Dans l'idéal, chaque distracteur pour un item devrait attirer au moins 5% des candidats (c'est-à-dire que chaque distracteur devrait avoir une valeur égale ou supérieure à 0.05).

Dans un autre cas, cependant, la clé de l'item est A et les indications de la colonne "Proportion totale" sont les suivantes:

A	.95
B	.04
C	.01
D	.00

Il apparaît à l'évidence que cet item était si facile que presque tous les candidats y ont répondu correctement et que le distracteur D était si faible que personne ne l'a choisi.

Les colonnes intitulées "Numéro d'ordre" et Échelonnage de l'item" apparaissent aussi sur la feuille du traitement MicroCAT. "Numéro d'ordre" indique le numéro d'ordre de l'item dans l'ensemble des données; "Échelonnage des données" indique le numéro de l'échelle sur laquelle l'item a été placé et sa position sur cette échelle. Par exemple l'item 8 , dans l'ordre général des items de cet ensemble de données, est le premier dans un sous-ensemble de 6 items à avoir été placé sur l'Echelle 2

On peut aussi tirer des informations sur la productivité du pré-test dans son ensemble avec ce groupe particulier de candidats. La Figure 5 en donne un exemple. Le sens des termes utilisés sous "Statistiques d'échelle" est donné ci-dessous:

Nombre d'items	Nombre d'items traités dans l'analyse
Nombre de candidats	Nombre de candidats inclus dans l'analyse
Moyenne	Pour la notation dichotomique d'items - le nombre moyens d'items qui ont fait l'objet d'une réponse juste; pour des items à choix multiple - la note moyenne des candidats de l'échantillon
Variance	Mesure de la dispersion des notes autour de la moyenne
Déviation standard	Racine carrée de la variance
Distribution symétrique	Forme d'une distribution
Voussure	Sommet de la distribution
Minimum	Note du candidat le plus faible
Maximum	Note du candidat le plus fort
Médiane (ou Médian Md)	Note du candidat moyen
Alpha	Coefficient alpha de fidélité pour chaque échelle allant de 0.0 à 1.0; c'est un indice de l'homogénéité d'une échelle et, dans l'idéal, la valeur devrait être aussi près que possible de 1.
Erreur standard de mesure	Indique "l'erreur" prévisible dans une note donnée. ESM = DS Ö (l-r(test)) DS = déviation (écart) standard f (test) = fidélité du test Nous pouvons être sûrs que 70% des notes se situeront à l'intérieur d'une déviation standard de la moyenne (+- 1 ESM) et sûrs à 95% que les notes se situeront à l'intérieur de 2 déviations standard (+- 2 ESM). Exemple: un étudiant obtient une note de 67 à un test avec une déviation standard de 9 et un coefficient de fidélité de 0.9

	SEM = 9√(1-0.9) = 2.8 Nous pouvons être sûrs à 70% que la note du candidat se situe entre 64.2 et 69.8. Nous pouvons être sûrs à 90% que la note du candidat se situe entre 61.4 et 72.6.
Indice moyen de difficulté à un item (moyenne P)	Proportion moyenne de réponses justes à un item (pour les items dichotomiques seulement)
Item moyen - total	Moyenne de corrélation bisériale de point transversalement à tous les items de l'échelle (pour les items dichotomiques seulement)
Corrélation bisériale moyenne de point	Moyenne de corrélation bisériale transversalement à tous les items de l'échelle
Note maximum (groupe faible)	Note maximum qu'un candidat placé dans le groupe faible peut obtenir (les derniers 27%)
N (groupe faible)	Nombre de candidats du groupe faible (les derniers 27%)
Note minimum (groupe fort)	Note minimum qu'un candidat placé dans le groupe fort peut obtenir (les premiers 27%)
N (groupe fort)	Nombre de candidats du groupe fort (les premiers 27%)

Figure 5

Feuille de résultats de MicroCAT Analysis
Statistiques d'échelles

MicroCAT (TM) Testing System
Copyright: Assessment System Corporation, 1982, 1984, 1986, 1988, 1993.
Programme d'analyse d'items et de test - ITEMAN (TM) Version 3.50
15 heures 59
Option de données absentes: traitement informatique des statistiques sur toutes les données disponibles
Il y avait 270 candidats pour ces données.

Statistiques d'échelles

Echelle	1	2	3	4
Nombre d'items	5	10	10	10
Nombre de candidats	270	270	270	270
Moyenne	3.230	6.633	8.422	8.163
Variance	0.725	3.321	1.755	2.588
Déviation standard	0.851	1.822	1.325	1.609
Distribution symétrique	0.047	-0.348	-0.361	-0.709
Voussure	-0.491	-0.202	3.043	-0.148
Minimum	1.000	1.000	2.000	3.000
Maximum	5.000	10.000	10.000	10.000
Médiane (ou Médian Md)	3.000	7.000	9.000	8.000
Alpha	0.091	0.431	0.318	0.499
Erreur standard de mesure (ESM)	0.812	1.375	1.094	1.138
Indice moyen de difficulté à un item (moyenne P)	0.646	0.663	0.842	0.816
Item moyen - total	0.428	0.406	0.378	0.415
Corrélation bisériale moyenne de point	0.676	0.547	0.602	0.621
Note maximum (groupe faible)	3	6	8	7
N (groupe faible)	168	116	115	89
Note minimum (groupe fort)	4	8	9	9
N (groupe fort)	102	85	155	132

ANNEXE 2: GLOSSAIRE

administration
Date ou période durant laquelle un examen a lieu. Certains examens sont administrés à dates fixes plusieurs fois par an, d'autres ont lieu à la demande.

analyse de contenu
Moyen permettant de décrire et d'analyser le contenu du matériel d'un test. L'objet de cette analyse est de s'assurer que le contenu du test est pertinent par rapport à ses spécifications. Elle est essentielle dans l'établissement de la validité de contenus et de la validité de construct.

analyse d'items
Description de la performance des items de tests individuels, employant généralement des indices statistiques classiques tels que la facilité ou la discrimination. On utilise pour cette analyse des logiciels tels que MicroCAT Iteman.

appariement
Type d'épreuve où le candidat doit relier entre eux des éléments apparaissant dans deux listes séparées. Une épreuve d'appariement consiste à sélectionner la phrase correcte qui complétera chacune des phrases tronquées proposées. Lors des épreuves de compréhension écrite on peut, par exemple, faire choisir dans une liste le type de vacances ou de livres correspondant à la description des goûts ou des besoins d'un personnage précis.

attribut
Caractéristique physique ou psychologique d'un individu (comme la capacité langagière, par exemple) ou échelle de mesure servant à décrire cette caractéristique.

banque d'items; syn**.: itémothèque**
Gestion des items qui permet de stocker des informations afin de pouvoir élaborer des tests aux contenu et difficultés connus. Une base de données informatisée est généralement utilisée à cet effet. Elle met en œuvre la théorie de l'attribut latent, ce qui signifie que les items peuvent être mis en relation les uns avec les autres au moyen d'une échelle de difficulté commune.

barème de notation
Liste de toutes les réponses acceptables aux items d'un test. Le barème permet au correcteur d'accorder la note appropriée.

calibrage
Détermination de l'échelle pour un ou plusieurs tests. Le calibrage peut impliquer des items d'ancrage de différents tests sur une sur une échelle de difficulté commune (échelle thêta). Quand un test est élaboré à partir d'items calibrés, les notes, en fonction de leur localisation sur l'échelle thêta, indiquent la capacité du candidat.

calibrer
Dans la théorie item-réponse: estimer la difficulté d'un ensemble questions.

candidat
Individu qui prend part à un examen ou à un test. Appelé aussi sujet.

caractéristiques des méthodes de test
Caractéristiques précises des différentes méthodes de test. Elles peuvent inclure l'environnement, la consigne, la langue dans laquelle sont données les instructions, la forme, etc.

classement
Conversion des notes obtenues en niveaux.

compétence structurale
Connaissance qu'un individu a des structures grammaticales d'une langue et sa capacité à les utiliser.

composante
Partie d'un examen souvent présentée comme un test à part entière, comportant un livret de consignes et une limite de temps. Les composantes sont souvent des épreuves basées sur les aptitudes langagières telles que la compréhension ou la production orale. Egalement appelé sous-test.

composition discursive
Tâche écrite dans laquelle le candidat doit, soit produire un discours à propos d'un sujet sur lequel il peut y avoir différentes prises de position, soit argumenter pour défendre son propre point de vue.

consigne
Instructions données aux candidats afin de guider leurs réponses à une tâche précise.

correcteur
Personne qui attribue une note aux réponses d'un candidat à un test écrit. Cette activité peut demander un jugement expert ou, dans le cas d'une notation mécanique, la simple application d'un barème de notation.

correction d'épreuves
Tâche où le candidat doit relire un texte en cherchant des erreurs, par exemple d'orthographe ou de structures. On peut également lui demander de noter les erreurs et de fournir les formes correctes.

correction collective
Méthode de correction des épreuves qui consiste à réunir tous les correcteurs pendant un temps limité, plutôt que de leur envoyer les tests à corriger chez eux.

déclencheur
Support graphique ou écrit qui permet d'obtenir une réponse du candidat dans les tests d'expression orale ou écrite.

descripteur
Brève description accompagnant un graphique en bande sur une échelle de notation. Elle résume le degré de compétence ou le type de performance attendu pour qu'un candidat atteigne une note précise.

descripteur d'échelle
Se réfère à la définition de **descripteur.**

discrimination
Le fait qu'un item puisse établir une distinction entre des candidats en les classant selon un degré allant du plus faible au plus fort. On utilise plusieurs indices de discrimination. Certains (comme le point bisérial) sont basés sur la corrélation entre la note obtenue à un item et un critère. Celui-ci peut être la note totale obtenue à ce test ou une autre mesure externe de niveau de capacité. D'autres critères sont basés sur la différence de

difficulté de l'item pour des groupes de capacité faible et élevée. Dans la théorie item-réponse, les modèles de paramètre 2 et 3 désignent l'item de discrimination comme paramètre-A.

document semi-authentique
Texte authentique dont le vocabulaire ou la grammaire a été adapté au niveau des candidats pour les besoins de l'évaluation.

double notation
Méthode d'évaluation où la performance du candidat est validée de façon indépendante par deux personnes.

échelle
Ensemble de catégories destinées à mesurer quelque chose. On en distingue quatre sortes: échelle nominale, ordinale, d'intervalle et de rapport.

échelle commune
Façon de reporter les notes obtenues à deux ou plusieurs tests sur une échelle commune, permettant une comparaison directe des résultats. Cela est faisable si les notes brutes ont été au préalable transformées par une procédure statistique comme, par exemple, le calibrage.

échelle de notation; syn.: **échelle d'évaluation**
Echelle composée de plusieurs catégories qui permettent d'exercer un jugement subjectif. Ce type d'échelle est fréquemment accompagné de descripteurs qui permettent d'interpréter les catégories.

élaborateur
Personne qui conçoit un nouveau test.

élaboration
L'ensemble du processus de production de matériel d'évaluation et de rédaction d'épreuves.

élaboration de test
Action de sélectionner des items ou des tâches en vue de la production d'un test. Souvent précédée du pré-testage ou de l'expérimentation du matériel. Les tâches ou les items nécessaires à l'élaboration du test peuvent être sélectionnés dans une banque de matériel.

entrée; syn. : **apport**
Matériel donné dans un test afin que le candidat produise une annonce appropriée. Dans une épreuve de compréhension orale par exemple, le texte enregistré pourra être accompagné d'un questionnaire écrit.

erreur standard de mesure
Dans la théorie de la note vraie, l'erreur standard de mesure (Se) indique l'imprécision de la mesure. La grandeur de l'erreur standard de mesure dépend de la fidélité (f) et de la déviation standard des notes (Sx). Pour calculer Se, la formule est:
$Se = Sx \sqrt{1-f}$

Si, par exemple, un candidat obtient une note vraie T, et si une déviation standard de mesure de Se revient fréquemment dans le test, cela signifie que, 68% des fois, la note observée sera dans le rang T+-Se, et que 95% du temps elle sera dans le rang T+-2Se.

essai; syn.: **expérimentation**
Etape de l'élaboration des tâches d'un test servant à vérifier que le test fonctionne de la façon attendue. Souvent utilisé dans le cas de tâches à notation subjective telles que la composition ou l'essai et administré à une population limitée.

examen réel (en grandeur nature)
Un test prêt à être utilisé et qui, pour cette raison doit être gardé en sécurité.

examinateur; syn: **évaluateur.**
Personne chargée de noter, de façon subjective, la performance du candidat à un test donné. Les évaluateurs sont généralement qualifiés dans leur domaine. On attend d'eux qu'ils se soumettent à un processus de formation et de standardisation. À l'oral, on distingue parfois les rôles d'examinateur et d'interlocuteur.

fidélité
Uniformité, constance ou stabilité des mesures. Plus un test est fidèle, moins il contient d'erreurs accidentelles. Un test présentant une erreur systématique, par exemple une distorsion qui désavantagerait certains groupes, peut être fidèle mais pas valide.

formes équivalentes; syn.: **formes parallèles, formes alternées**
Différentes versions du même test considérées comme équivalentes car basées sur les mêmes spécifications et mesurant la même compétence. Dans la théorie classique du test, pour répondre aux exigences d'une véritable équivalence, les différentes formes du test doivent avoir le même type de difficulté, la même variance, la même covariance et avoir un critère concordant lorsqu'ils sont administrés aux mêmes personnes. Dans la pratique, l'équivalence est très difficile à atteindre.

impact
Effet produit par un examen, à la fois en termes d'influence sur le processus éducatif en général et pour les individus intéressés par les résultats de cet examen.

indice de facilité
Proportions de réponses correctes à un item, transcrite sur une échelle de 0 à 1. Egalement appelée valeur-p.

item; syn.: **question**
Chaque point particulier d'un test auquel on attribue une ou plusieurs notes séparées. Exemples: un "blanc" dans un test de closure, une des questions dans un questionnaire à choix multiple à quatre options, une phrase donnée pour une transformation grammaticale, une question dont la réponse attendue est une phrase complète.

item à choix multiple; syn.: **question à choix multiple (QCM)**
Type d'item qui consiste en une question ou une phrase incomplète, accompagnée d'un choix de réponses ou de propositions pour compléter la phrase (options). Le candidat devra choisir l'option correcte (clé) parmi trois, quatre ou cinq possibilités. Aucune production langagière ne lui est demandée. C'est pour cette raison qu'on utilise habituellement les items à choix multiples dans les tests de compréhension écrite et orale. Ils peuvent être discrets ou basés sur du texte.

item basé sur un texte
Item qui s'appuie sur un discours suivi par exemple items à choix multiple basés sur une compréhension de texte.

item discret
Item contenant en lui-même tous les éléments de la question. Il n'est lié ni à un texte, ni à d'autres items, ni à un quelconque matériel complémentaire. Le choix multiple est un exemple de ce type d'item.

item discret spécifique
Item discret évaluant un point spécifique, par exemple une structure ou du lexique, et n'ayant aucune relation avec d'autres items. La vulgarisation de ce type d'item dans les tests de langue est due à Robert Lado dans les années 60.

item de construction de mots
Type d'item dans lequel le candidat doit produire une forme d'un mot à partir d'un mot de la même famille qui lui est donné comme entrée.

item de liaison
Renvoie à la définition d'item d'ancrage ou de référence.

item de référence ou d'ancrage
Item intégré à deux ou plusieurs tests et permettant d'estimer soit la différence du degré de difficulté entre les tests, soit la différence de performance entre les différents groupes de candidats.

item de transformation
Se réfère à la définition de **transformation de phrase.**

jeu de rôle
Type de tâche parfois utilisée dans les tests d'expression orale et dans laquelle les candidats doivent se projeter dans une situation de communication précise ou jouer un rôle particulier.

langue sur objectifs spécifiques
Enseignement ou évaluation de la langue centré sur un domaine particulier de la langue utilisée dans des activités ou une profession particulière; par exemple, anglais des contrôleurs aériens, espagnol du commerce.

lecteur optique; syn.: **scanner**
Appareil optique utilisé pour scanner l'information directement recueillie à partir des feuilles de notes ou des feuilles de réponse. Les candidats ou les examinateurs marquent les réponses aux items sur une feuille de notes et cette information est automatiquement lue par l'ordinateur.

lexique
Terme utilisé pour désigner le vocabulaire.

mesure
D'une façon générale, il s'agit du processus qui permet de trouver la somme de quelque chose par comparaison avec une unité fixe, comme lorsqu'on utilise une règle pour mesurer la longueur. En sciences sociales, la mesure se réfère souvent à la quantification des caractéristiques des individus comme, par exemple, la compétence langagière.

mise en forme
Procédure qui consiste à modifier le matériel d'évaluation soumis par des producteurs d'items et à le mettre dans la forme définitive qu'il aura pour l'examen.

modèle de Rasch

Modèle mathématique, connu également comme le modèle de la logistique simple, qui postule qu'il existe une relation entre la probabilité qu'un individu réalise une tâche et la différence entre la capacité de l'individu et la difficulté de la tâche. Equivalant mathématiquement au modèle à paramètre unique dans la théorie de l'item réponse. Le modèle de Rasch a été appliqué de différentes façons, par exemple pour traiter les réponses échelonnées ou les différentes facettes à prendre en compte dans la "difficulté" d'une tâche.

moyenne
La moyenne est la mesure de la tendance centrale. On obtient la note moyenne à un test en additionnant toutes les notes obtenues et en divisant ce total par le nombre de candidats.

niveau
a) La note obtenue à un test peut être communiquée au candidat sous forme de niveau, par exemple sur une échelle de A à E, où A représente le niveau le plus élevé, B un bon niveau, C un niveau passable et D et E des niveaux insuffisants.
b) On fait souvent référence à une série de niveaux pour désigner le niveau de capacité requis pour qu'un étudiant soit classé dans tel ou tel groupe ou lorsqu'il a réussi à un test donné. Les termes les plus utilisés pour désigner ces niveaux sont "élémentaire", "intermédiaire", "avancé", etc.

niveau de survie (Waystage Level**)**
Référentiel d'un niveau élémentaire de compétence en langue étrangère, publié pour l'anglais en 1977 par le Conseil de l'Europe et revu en 1990. Moins exigent que le Niveau seuil, il ne couvre qu'environ la moitié des apprentissages définis par ce dernier.

notation
a)Attribution d'une note aux réponses d'un candidat à un test. Cette activité peut demander un jugement professionnel ou l'application d'un barème où sont indiquées toutes les réponses acceptables
b)Note accordée et qui représente le résultat du processus d'évaluation.

notation par ordinateur
Différentes méthodes utilisent l'informatique afin de minimiser les erreurs dans les notations des tests objectifs. On peut, par exemple, scanner les feuilles de notes des candidats à l'aide d'un lecteur optique afin d'analyser les données.

notation standardisée (mécanique)
Méthode de notation dans laquelle on n'attend pas des évaluateurs qu'ils exercent quelque compétence ou jugement subjectif que ce soit. La bote est établie d'après un relevé de toutes les réponses acceptables pour chaque question du test.

notes
Le résultat d'un examen, souvent exprimé en pourcentage. À cause des réajustements dus au jeu des coefficients, la note ne correspond pas toujours au total des points.

phrase à compléter
Type d'item dans lequel seule une moitié de la phrase est donnée. La tâche du candidat consiste à compléter la phrase, soit en fournissant les mots convenables (éventuellement d'après un texte), soit en choisissant ces mots parmi différentes possibilités.

pondération; syn.: **coefficient**

Action d'assigner un nombre maximum différent de points à un item, une tâche ou une épreuve afin de changer sa contribution relative au total des points en fonction des autres parties du test. Si, par exemple, on attribue une note double à tous les items de la tâche n° 1 d'un test, la tâche n° 1 sera proportionnellement plus importante que les autres tâches dans le total des points obtenus.

pré-testing; syn.: **pré-testage**
Etape de l'élaboration du matériel des tests pendant laquelle on essaie les items sur des échantillons représentatifs de la population cible afin de déterminer leur niveau de difficulté. Suivant une analyse statistique, les items considérés comme satisfaisants pourront être utilisés dans des tests réels.

production de tests
Procédure de sélection des items qui figureront dans la version finale de l'examen auxquels on ajoute les consignes et la grille de correction.

question
Parfois utilisé pour désigner une tâche ou un item dans un test.

question intégrée
Se réfère à des questions ou des tâches à réaliser qui mettent en jeu plus d'une habileté ou sous-habileté. Exemple: compléter un test de closure, participer à un entretien oral, lire une lettre et y répondre.

question lacunaire
Tout type d'item qui demande au candidat d'insérer du matériel écrit - des lettres, des chiffres, un mot isolé, plusieurs mots, des phrases ou des paragraphes - dans les espaces blancs aménagés dans un texte. La réponse peut être produite par le candidat ou bien sélectionnée dans une liste.

question ouverte; syn.: **question à réponse construite, question à réponse libre**
Type d'item ou de tâche dans un test écrit qui demande au candidat de produire une réponse (et non de la sélectionner). L'objectif de ce type d'item est de faire produire une réponse relativement libre et dont la longueur peut aller de quelques mots à un grand nombre de phrases. Le barème proposera alors tout un choix de réponses acceptables.

registre
Différentes variétés de langue correspondant à des activités particulières ou à un formalisme plus ou moins grand.

réponse
Comportement du candidat manifesté par les entrées données dans un test. Par exemple, la réponse donnée à un item à choix multiple ou le travail produit dans un test d'expression écrite.

réponse clé
a) Choix correct dans un item à choix multiple (voir: item à choix multiple)
b) Plus généralement, un ensemble de réponses correctes ou acceptables.

révision; syn.: **contrôle**
Etape de l'élaboration d'un test pendant laquelle les élaborateurs évaluent le matériel produit et décident de rejeter ce qui ne convient pas aux spécifications du test et de poursuivre la mise en forme de ce qui convient.

script

Feuille contenant les réponses du candidat à un test, dans les tâches de type réponse ouverte.

situation de communication réelle
Point de vue selon lequel les tests devraient inclure des tâches ressemblant le plus possible à des activités réelles. Le contenu d'un test évaluant si un candidat est capable de suivre un cours de langue étrangère devrait, par exemple, être basé sur une analyse de la langue et des activités langagières particulières à ce cours.

spécification
Description des caractéristiques d'un examen indiquant ce qui est testé, comment ainsi que le nombre et la longueur des épreuves, les types d'item utilisés, etc.

spécifications du Threshold Level **ou du** Niveau seuil
Description détaillée d'un niveau de langue en anglais ou en français conçue par le Conseil de l'Europe. On estime qu'un débutant a besoin d'environ 375 heures d'apprentissage pour l'atteindre.

syllabus
Document détaillé où sont listés tous les domaines d'un programme d'études particulier et l'ordre dans lequel le contenu est présenté.

tâche
Combinaison de consignes, d'entrées et de réponses. Exemple: texte à lire accompagné d'items à choix multiple auxquels on peut répondre en suivant une seule consigne.

tâche d'appariement multiple
On propose au candidat un certain nombre d'items à compléter sous forme de questions ou de phrases, généralement à partir d'un texte écrit. Les réponses sont fournies dans une banque de mots ou de phrases qui peuvent être utilisés plusieurs fois. L'avantage de cette présentation est que les options ne disparaissent pas au fur et à mesure que le candidat progresse dans le test (comme c'est le cas dans d'autres formes d'appariement); l'exercice ne devient donc pas de plus en plus facile.

tâche d'écriture dirigée
Se réfère à la définition de production écrite guidée, à savoir 311: le candidat doit produire un texte écrit, dans lequel des informations graphiques ou textuelles, telles que des images, des lettres, des cartes postales ou des modes d'emploi, sont utilisés pour contrôler et standardiser la réponse attendue.

test de closure
Type de tâche lacunaire dans laquelle des mots entiers sont supprimés d'un texte. Dans un test de closure traditionnel, on supprime un mot tous les x mots. On appelle également test de closure l'exercice dans lequel des phrases courtes sont supprimées d'un texte, ou lorsque l'élaborateur choisit les mots qui seront supprimés, comme c'est le cas dans le test de closure rationnel. Les candidats devront fournir les mots manquants (test ouvert) ou les choisir dans une liste (choix multiple ou test de closure à lacunes sélectives). Le corrigé d'un test ouvert peut comporter soit le mot exact (le mot supprimé du texte original étant seul accepté comme réponse correcte), soit les mots acceptables (dans ce cas une liste de mots acceptables est donnée au correcteur).

test lacunaire à choix multiple
Type d'item d'un test pour lequel le candidat doit choisir parmi plusieurs options la phrase ou le mot correct à insérer dans une lacune du texte.

test objectif
Test auquel on peut appliquer un barème de notation et qui ne fait pas appel à une opinion d'expert ou à un jugement subjectif.

texte
Discours suivi, écrit ou oral, utilisé pour élaborer un ensemble d'items dans un test.

théorie de l'item-réponse TQR
Groupe de modèles mathématiques permettant de mettre en rapport la performance d'un candidat à un test avec son niveau de capacité. Ces modèles se fondent sur la théorie fondamentale qui spécifie que la performance attendue d'un individu à une question ou à un item donné d'un test est fonction à la fois du niveau de difficulté de la question et du niveau de capacité de l'individu.

transfert d'information
Technique d'évaluation qui implique qu'une information donnée sous une certaine forme soit présentée d'une façon différente. Par exemple, reporter les informations d'un texte sur un diagramme; transformer une note informelle en annonce officielle.

transformation de phrases
Type d'item dans lequel l'amorce donnée est une phrase complète, suivie par le premier ou les deux premiers mots d'une seconde phrase qui reprend le contenu de la première mais sous une forme grammaticale différente. La première phrase peut être, par exemple, à la forme active et le candidat devra la présenter à la forme passive.

validité
Degré auquel les notes d'un test permettent de tirer des conclusions appropriées, significatives et utiles, en relation avec l'objet du test. On distingue différents aspects de la validité tels que la validité de contenu, la validité critérielle et la validité de construct; elles donnent différentes sortes de preuves permettant de juger la validité globale d'un test en fonction de ses objectifs.

validité apparente
Qualité d'un test ou de toute autre mesure qui semble correcte et adéquate à l'objet mesuré. Il s'agit là d'un jugement subjectif plus que d'un jugement basé sur une analyse objective du test. La validité apparente est souvent considéré comme une fausse forme de validité. On l'appelle également attrait d'un test *(test appeal).*

validité concourante
On dit d'un test qu'il a une validité concourante si les notes obtenues sont en corrélation élevée avec un critère externe reconnu qui mesure le même domaine de connaissance ou de capacité.

validité convergente
On dit d'un test qu'il a une validité convergente lorsqu'il y a une corrélation élevée entre les notes obtenues à ce test et celles obtenues à un autre test mesurant le même construct (indépendamment de la méthode). Il s'agit là d'un autre aspect de la validité de construct.

validité critérielle
On dit d'un test qu'il a une validité critérielle si on peut démontrer le rapport entre les notes obtenues et un critère externe qui est censé mesurer la même capacité. Lors de l'absence de critère, l'information fournie indique jusqu'à quel point le test peut prédire le comportement futur.

validité de construct; syn.: **validité hypothético-déductive; validité conceptuelle.**

On dit d'un test qu'il a une validité de construct si les notes obtenues peuvent être interprétées comme une théorie sur la nature d'un construct ou sur le rapport de ce construct avec d'autres. On pourrait prédire, par exemple, que deux tests valides de compréhension orale classent les apprenants de la même façon, mais chacun d'eux aurait un rapport plus éloigné avec les notes obtenues à un test de compétence grammaticale.

validité de contenu
On dit d'un test qu'il a une validité de contenu si les items ou les tâches dont il est composé constituent un échantillon représentatif des items ou des tâches pour une capacité ou un domaine de connaissances précis.

validité discriminante
On dit d'un test qu'il a une validité discriminante si la corrélation qu'il entretient avec des tests évaluant différents attributs est plus faible que celle qu'il a avec des tests évaluant le même attribut, sans tenir compte de la méthode d'évaluation. Cela peut être considéré comme un aspect de la validité de construct.

validité prédictive
Indique la façon dont un test peut prédire la future performance dans une aptitude donnée.

(Ce glossaire est extrait du glossaire multilingue de l'évaluation produit par ALTE - Association of Language Testers in Europe et publié par Cambridge University Press dans la série *Cambridge Studies in Language Testing*) :
Reliure cartonnée 0 521 65099 2
Reliure papier glacé 0 521 65877 2
CD-ROM 0521 658241