MANUEL pour L'ELABORATION et la PASSATION DE TESTS et d'EXAMENS DE LANGUE

A utiliser en liaison avec le CECR

ALTE,

pour le Conseil de l'Europe,
Division des politiques linguistiques

Division des Politiques linguistiques
DG II – Service de l'éducation
Conseil de l'Europe, Strasbourg

www.coe.int/lang/fr

© Conseil de l'Europe, avril 2011

Les opinions exprimées dans cet ouvrage n'engagent que leurs auteurs et ne reflètent pas nécessairement la politique officielle du Conseil de l'Europe.

Toute correspondance concernant cette publication, la reproduction ou la traduction de tout ou partie de ce document doit être adressée au directeur d'Education et Langues (Division des politiques linguistiques) (F-67075 Strasbourg cedex, ou à decs-lang@coe.int)

La reproduction d'extraits est autorisée, excepté à des fins commerciales, à condition que la source soit citée.

TRADUCTION

Gilles BRETON Christine TAGLIANTE

SOMMAIRE

Contenu

NOTE PRÉLIMINAIRE

Ce Manuel complète la « boîte à outils » qui propose une aide à l'utilisation du *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer (CECR)*. Nos remerciements vont à l'Association des organismes certificateurs en Europe (ALTE) chargée par le Conseil de l'Europe de la préparation ce Manuel. Cette association contribue une fois de plus à une utilisation efficace du CECR, dans le respect de l'esprit des statuts participatifs dont l'Organisation internationale non gouvernementale (INGO) jouit auprès du Conseil de l'Europe.

L'objectif du CECR était de fournir aux Etats membres du Conseil de l'Europe, un point de départ commun pour la réflexion et les échanges entre les différents partenaires du champ, incluant les personnels impliqués dans la formation d'enseignants ainsi que dans l'élaboration des programmes de langues, des directives concernant les cursus, des manuels d'apprentissage, des examens, etc. Le CECR propose un outil descriptif qui permet aux utilisateurs de réfléchir à leurs décisions et à leurs pratiques, de bien placer leurs efforts et de les coordonner, en tant que de besoin, au profit des apprenants de différents contextes. Le CECR est donc un outil souple, adaptable aux différents contextes d'utilisation – l'illustration parfaite de cet aspect fondamental est le système de niveaux, qui peut être adapté et exploité souplement pour l'élaboration de différents objectifs d'enseignement / apprentissage ainsi que pour l'évaluation, et pour la « Description des niveaux de référence » (DNR) pour des langues et des contextes particuliers.

Les descripteurs, créés à partir de ceux « qui ont été reconnus clairs, utiles et pertinents par des groupes de professeurs enseignant ou non leur langue maternelle dans des secteurs éducatifs variés et avec des profils de formation et une expérience professionnelle très différents » (CECR, p. 30), ne prétendent pas être détaillés de façon exhaustive, ni, en aucune façon, normatifs. Les utilisateurs sont invités à les adapter ou les compléter en fonction du contexte et des besoins. Le présent Manuel fournit de précieux conseils pour construire dans cet esprit des tests de compétence liés aux niveaux du CECR d'une manière à la fois guidée et non prescriptive.

La nécessité de garantir la qualité, la cohérence et la transparence dans les prestations liées aux langues ainsi que l'intérêt croissant dans l'aspect porteur des examens, ont créé un grand intérêt pour les niveaux du CECR, perçus en Europe et au-delà comme un outil de référence et un instrument de calibrage. Partageant ce point de vue, nous souhaitons également encourager les utilisateurs à explorer et partager des expériences sur la façon dont le CECR, dans ses différents aspects, peut être encore davantage utilisé pour favoriser l'évolution, tout au long de la vie, du profil plurilingue (irrégulier et changeant) des apprenants qui, au final, devront prendre la responsabilité d'organiser et d'évaluer leur apprentissage en fonction de leurs besoins évolutifs et des changements de circonstances. L'initiative du Conseil de l'Europe de promouvoir l'éducation plurilingue et interculturelle, ainsi qu'une approche globale de toutes les langues dans et pour l'éducation, présente de nouveaux défis pour l'élaboration des programmes, pour l'enseignement et l'évaluation, le moindre d'entre eux n'étant pas celui d'évaluer la compétence des apprenants à l'aide de leurs répertoires plurilingues et interculturels. Nous attendons beaucoup de la contribution essentielle d'associations professionnelles telles qu'ALTE pour nous aider à promouvoir les valeurs du Conseil de l'Europe dans le domaine de l'éducation aux langues.

Joseph Sheils

Division des politiques linguistiques

Conseil de l'Europe

INTRODUCTION

Contexte

Depuis sa publication dans sa version finalisée, en 2001, *le Cadre européen commun de référence pour les langues (CECR)* n'a cessé de connaître un intérêt toujours croissant non seulement en Europe, mais également à l'échelle mondiale. Son impact a dépassé les attentes et il ne fait aucun doute qu'il a contribué à éveiller l'attention sur d'importants problèmes liés à l'apprentissage, l'enseignement et l'évaluation en langues. Le Conseil de l'Europe a également encouragé la création d'une « boîte à outils » comportant des ressources pour l'information et l'utilisation du CECR par les décideurs politiques, les enseignants, les organismes certificateurs et les autres partenaires du domaine.

Comme l'a signalé Daniel Coste, l'un des auteurs du CECR, l'influence du *Cadre* sur l'évaluation a été particulièrement remarquable, et le processus d'ancrage des examens de langue aux niveaux de référence a reçu plus d'attention que tout autre de ses aspects (2007). Un certain nombre d'outils sont désormais disponibles à l'intention des organismes certificateurs et des praticiens intéressés par les tests de langue :

- *Manuel pour relier les examens de langue au Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer* (Conseil de l'Europe, 2009).
- *Supplément* technique *de référence au Manuel* (Banerjee 2004 ; Verhelst 2004 a, b, c, d ; Kaftandjieva 2004 ; Eckes 2009).
- *Illustrations des niveaux de compétences en langues.*
- *Grilles d'analyses de contenus pour la production orale et écrite ainsi que la réception orale et écrite.*
- *Description des niveaux de référence pour l'anglais et d'autres langues.*

Le Conseil de l'Europe a également organisé des forums (Réflexions sur l'utilisation de l'avant-projet du *Manuel pour relier les examens de langue au CECR,* Cambridge, 2007 ; séminaire pré-conférence, Conférence d'EALTA, Athènes, 2008) au cours desquels les praticiens ont échangé leurs réflexions sur l'utilisation du manuel ainsi que sur leurs expériences de mise en pratique des différentes étapes d'ancrage suggérées dans le Manuel.

L'Association des organismes certificateurs en Europe (ALTE), en tant qu'Organisation internationale non-gouvernementale (INGO) ayant un statut consultatif au sein du Conseil de l'Europe, a contribué aux ressources composant la boîte à outils, y incluant le Portfolio européen des langues (PEL) d'EAQUALS/ALTE ainsi que les grilles d'analyse de contenus d'ALTE. L'association était également représentée par le Dr Piet van Avermaet, du groupe d'auteurs du *Manuel pour relier les examens de langue au CECR.* En accord avec la Division des politiques linguistiques du Conseil de l'Europe, ALTE tient à ce que les utilisateurs de la boîte à outils se servent efficacement du *Cadre* dans leur propre contexte et afin de satisfaire leurs propres objectifs.

Le but de ce Manuel

Le *Manuel pour relier les examens de langue au CECR* mentionné ci-dessus a été spécifiquement conçu pour aborder l'ancrage des tests et examens au *Cadre*, et, avec le *Supplément de référence*, il présente et propose une approche générale ainsi qu'un certain nombre de choix, y compris sur la définition des points de césure.

Le *Manuel pour l'élaboration et la passation de tests et d'examens de langues* est conçu comme un complément du *Manuel pour Relier les examens de langues au CECR*. Il met l'accent sur les aspects de l'élaboration et de la passation de tests et d'examens qui ne sont pas couverts par l'autre Manuel. Il s'agit, en fait, d'une version actualisée d'un document antérieur produit par le Conseil de l'Europe connu sous le nom de *CECR : Evaluation de compétences en langues et conception de tests* (1996), l'un des *Guides pour les utilisateurs* accompagnant le premier projet du CECR en 1996/7, commandités par le Conseil de l'Europe.

ALTE était l'auteur de la première version de ce Manuel sur l'évaluation. Au cours de la dernière décennie, des évolutions de la théorie de la validité ainsi que l'utilisation et l'influence grandissantes du CECR ont montré la nécessité d'une réelle actualisation du document. ALTE a accepté avec plaisir de coordonner ces révisions en 2009/10 et de nombreuses personnes, membres et associés d'ALTE ont contribué à la rédaction de ce document.

Lors des révisions, il a été utile de se souvenir des origines et des buts du CECR et de les faire apparaître dans la structure et les objectifs de ce Manuel destiné aux utilisateurs.

En tant que cadre commun de référence, le CECR se voulait tout d'abord un « outil pour la réflexion, la communication et la prise de décision » (Trim, 2010). Il a été conçu pour permettre une même compréhension des domaines de l'apprentissage,

de l'enseignement et de l'évaluation en langues et, dans le débat sur l'éducation aux langues, il permet un langage commun sur chacun de ces aspects. Il fournit également un ensemble de niveaux de référence pour identifier les niveaux de compétence en langues, depuis le faux débutant (A1) jusqu'à un niveau très avancé (C2), et ceci dans toute une série de capacités différentes et de domaines d'utilisation.

Grâce à tout cela le CECR permet la comparaison des pratiques dans des contextes très différents, en Europe et au-delà. En tant qu'outil de référence, il doit cependant, dans certaines situations, être adapté au contexte et aux objectifs locaux.

Ce point a été très clairement décrit par les auteurs du CECR. Dans l'avertissement destiné aux utilisateurs (p.4), ils précisent notamment « Soyons clairs : il ne s'agit aucunement de dicter aux praticiens ce qu'ils ont à faire et comment le faire », et ils le réitèrent à plusieurs reprises. Parmi les ressources de la boîte à outil, le *Manuel pour Relier les examens de langues au CECR* suit la même démarche. Ses auteurs indiquent sans ambigüité que ce manuel n'est pas le seul qui permette d'ancrer un test ou un examen au CECR et qu'aucune institution n'est obligée d'entreprendre ce processus d'ancrage (p.1).

Dans un forum politique intergouvernemental du Conseil de l'Europe sur l'utilisation du CECR à Strasbourg en 2007, Coste a souligné combien les utilisations contextuelles prises comme des interventions délibérées dans un environnement donné peuvent « prendre des formes variées, concerner des niveaux différents, avoir différents objectifs et impliquer des types de partenaires distincts ». Il déclare « Chacune de ces application contextuelles est légitime et significative, mais, alors que le *Cadre* lui-même propose une série de choix intégrés, certaines de ces applications contextuelles les exploitent à fond, alors que d'autres les élargissent ou les dépassent ». C'est pourquoi, lorsqu'on envisage la question de l'ancrage, il est important d'avoir présent à l'esprit que le CECR n'a pas été conçu pour être utilisé de façon prescriptive et qu'il n'y a donc pas une façon unique de justifier l'ancrage d'un examen dans un contexte et un but d'utilisation particuliers.

Comme l'ont souligné Jones et Saville (2009 : 54-55) :

« … certaines personnes disent appliquer scrupuleusement le CECR à un contexte particulier. Nous préférons plutôt rapporter le contexte au CECR. L'autre façon d'agir est la transitivité. Le débat en faveur de l'ancrage est encore à construire, la base de comparaison est à établir. C'est le contexte spécifique qui détermine la signification définitive de l'affirmation d'ancrage. En posant le problème ainsi, nous replaçons le CECR dans son rôle de point de référence et contribuons à son évolution future. »

Alors que le *Manuel pour Relier les examens de langues au CECR* met l'accent sur « les procédures engagées pour présenter les preuves de l'affirmation que tel test ou examen est ancré au CECR » et « ne donne pas de conseils généraux sur la façon de concevoir de bon tests ou de bons examens » (p.2), l'approche complémentaire adoptée dans le présent Manuel part du processus d'élaboration du test et montre comment il est possible d'établir un lien avec le CECR à chaque étape de ce processus, de façon à :

- spécifier le contenu du test ou de l'examen
- cibler des niveaux spécifiques de compétence langagière
- interpréter la performance au test de langue en termes qui se réfèrent à la langue réelle utilisée hors situation de test.

Ce Manuel a par conséquent un objectif plus vaste que les trois principales utilisations du CECR, qui sont :

- La spécification des contenus des tests et examens.
- L'établissement des critères permettant d'atteindre un objectif d'apprentissage, en liaison à la fois avec l'évaluation d'une performance orale ou écrite particulière et avec l'évaluation continue de l'enseignant, l'évaluation par les pairs et l'auto-évaluation.
- La description des niveaux de compétence dans les tests et les examens existants qui permet des comparaisons entre les différents systèmes de certification.

Son souhait est de fournir un guide cohérent pour l'élaboration de tests et d'examens généraux, qui peut être utile pour concevoir des tests et examens à objectifs spécifiques, en présentant cette élaboration sous la forme d'un cycle, chaque étape réussie étant due au travail fourni à l'étape précédente. La totalité du cycle doit obligatoirement être traitée pour que chaque étape fonctionne correctement. La section 1.5 montre une vue d'ensemble du cycle, qui est par ailleurs détaillé aux chapitres suivants :

Chapitre 1 – Présentation des concepts fondamentaux liés à la compétence langagière : validité, fiabilité et équité.

Chapitre 2 – Elaboration - depuis la décision de concevoir jusqu'à la rédaction des spécifications définitives.

Chapitre 3 – Assemblage - traite de la rédaction des items et de la construction des tests.

Chapitre 4 – Passation - s'applique à l'administration des tests, depuis l'inscription des candidats jusqu'au retour du matériel de test.

Chapitre 5 – Correction, notation et délivrance des résultats – à la fin du cycle opérationnel.

Chapitre 6 – Contrôle et révision – montre comment le cycle peut être répété au fil du temps afin d'améliorer la qualité et l'utilité du test ou de l'examen.

Pour qui a été conçu ce Manuel ?

Il est destiné à tous ceux qui sont impliqués dans l'élaboration et l'utilisation de tests et d'examens de langues liés au CECR. Il a été conçu pour être utile aussi bien aux concepteurs débutants qu'aux plus expérimentés. C'est pourquoi il présente des principes communs, qui s'appliquent aux tests de langues en général, que l'organisme certificateur soit une grande institution préparant des tests pour des milliers de candidats dans le monde, ou qu'il s'agisse d'un enseignant isolé souhaitant évaluer ses élèves en classe. Les principes sont les mêmes pour des tests à fort ou à faible enjeu, seules les étapes pratiques varieront.

Nous partons du principe que les lecteurs sont déjà familiarisés avec le CECR, ou seront prêts à l'utiliser conjointement avec ce Manuel lors de l'élaboration et de l'utilisation de tests ou d'examens.

Comment utiliser ce Manuel ?

Bien que les principes présentés ici soient généraux, le certificateur doit décider de leur application dans son contexte particulier. Ce Manuel donne des exemples et des conseils sur la façon de mener certaines activités. Ces conseils pratiques seront toutefois plus pertinents dans certains contextes que d'autres en fonction de l'objectif du test ou de l'examen et des ressources disponibles pour les mettre au point. Cela ne signifie pas que le Manuel est moins utile pour certains : si les utilisateurs comprennent les principes, ils peuvent se servir des exemples pour les appliquer à leur contexte particulier.

Outre le CECR, il existe de nombreuses autres ressources utiles pour relier un test ou un examen de langue au CECR. Ce Manuel n'est qu'un outil parmi ceux proposés dans la boîte à outils conçue et mise à disposition par le Conseil de l'Europe. C'est pourquoi on n'y trouvera pas d'informations ou de théories disponibles ailleurs. Comme nous l'avons déjà signalé, cet ouvrage est complémentaire du *Manuel pour Relier les examens de langues au CECR*, il ne reprend pas les informations qui y sont données.

Il n'est pas nécessaire de le lire de A à Z. Chacun, en fonction de ses besoins d'élaboration et de passation de test ou de l'examen, peut lire uniquement les parties qui lui conviennent. Cependant, même pour ceux qui se sont spécialisés dans l'un des champs des examens de langue, la lecture complète du Manuel permet d'avoir un bon aperçu de l'ensemble du cycle.

A la fin de chaque chapitre, des « Lectures complémentaires » guident le lecteur soit vers des ressources pour approfondir un domaine, soit vers des outils pratiques. Ces lectures sont suivies de questions clés destinées à renforcer la compréhension de ce qui a été lu.

Cet ouvrage est non prescriptif, son objectif est de mettre l'accent sur les grands principes et les approches concernant la conception de tests et d'évaluation de façon à ce que l'utilisateur puisse y faire référence lors qu'il élabore un test ou un examen répondant à son contexte particulier. Ce n'est pas un livre de recettes pour placer les questions d'un test sur les échelles du CECR : les six niveaux de référence, suffisamment clairs et détaillés, fournissent un outil commun de référence et n'ont pas été conçus à l'origine dans ce but.

En réalité, dans l'une des premières versions du *Cadre* (Strasbourg, 1998), les échelles descriptives se trouvaient en annexe, à titre d'exemples, et n'apparaissaient pas dans le corps du texte. Seuls les niveaux communs de référence étaient présents dans le texte. La mise en page originale de la version de 1998 renforçait les différents statuts et fonctions des niveaux de référence généraux, dont certains n'étaient pas calibrés et étaient sous-représentés aux niveaux C.

Dans la version de 1998 du CECR, le statut provisoire des échelles de descripteurs était décrit de façon explicite dans le texte (p.25) :

« L'élaboration d'un ensemble de points de référence communs ne limite en aucune façon les choix que peuvent faire des secteurs différents, relevant de cultures pédagogiques différentes, pour organiser et décrire leur système de niveaux. On peut aussi espérer que la formulation précise de l'ensemble des points communs de référence ainsi que la rédaction des descripteurs évolueront avec le temps, au fur et à mesure que l'expérience des Etats membres et des organismes compétents dans le domaine sera prise en compte. »

Le risque d'utiliser les échelles de façon exagérément prescriptive est de laisser croire qu'on peut mesurer la compétence langagière par une approche « universelle ». Les échelles fonctionnelles et linguistiques sont plus conçues pour illustrer la nature générale des niveaux que pour en donner une définition précise. C'est pourquoi, étant donné la grande diversité des contextes démographiques, des besoins et des types d'apprentissage et d'enseignement, il est par exemple impossible de donner les caractéristiques d'un étudiant de « type B1 ». Le corolaire est qu'il est difficile de concevoir un programme ou un test convenant à tous les contextes, que ce soit pour B1 ou pour n'importe quel autre niveau.

Afin que le CECR ait un impact durable et positif, les organismes certificateurs doivent intégrer ses principes et ses pratiques dans leurs procédures. Cela permettra, au fur et à mesure, aux systèmes professionnels d'établir des argumentaires en faveur de l'ancrage afin d'appuyer leur affirmation, et impliquera de travailler en s'appuyant sur le texte du CECR éventuellement adapté aux contextes et aux applications particuliers.

Puisqu'il n'est pas possible, à partir d'un seul atelier de définition des points de césure, de mettre en évidence un ancrage stable et constant, il est important que les organismes certificateurs fournissent des preuves variées, accumulées dans le temps. Les recommandations du *Manuel pour Relier les examens de langues au CECR* ainsi que celles des autres ressources de la boîte à outil doivent donc faire partie intégrante des procédures standard que les organismes certificateurs mettent en œuvre pour leurs besoins d'ancrage, et ne doivent pas être traitées à la légère.

C'est ce à quoi ce Manuel encourage le lecteur, en mettant l'accent sur l'importance de concevoir des systèmes qui permettent d'établir des normes et de les suivre au fil du temps.

Conventions utilisées

Les conventions suivantes sont appliquées tout au long du Manuel :

- Les termes *ce Manuel* sont utilisés pour désigner le *Manuel pour l'élaboration et la passation de tests et d'examens de langues*.
- Le sigle CECR désigne le *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*.
- L'institution chargée de développer le test est appelée *organisme certificateur*. L'expression *le concepteur de test* est parfois utilisée pour désigner ceux qui ont une fonction particulière dans le cycle d'élaboration du test.
- Lors de leur première apparition dans ce Manuel, et lorsqu'il nous semble qu'il est utile de les signaler au lecteur, les mots indexés au glossaire (annexe VIII) apparaissent en PETITES MAJUSCULES.

Dr Michael Milanovic

Directeur d'ALTE

1. Considérations essentielles

Les conseils pratiques donnés dans ce Manuel pour élaborer des tests ou des examens de langue nécessitent de bonnes bases en principes et théorie. Ce chapitre traite des questions suivantes :

- Comment définir la compétence langagière
- Pourquoi la *validité* est-elle la qualité-clé d'un bon test
- Qu'est-ce que la *fiabilité*
- L'équité dans les tests et examens.

Cette dernière section présente également les grandes lignes des processus d'élaboration d'un test, détaillées dans les chapitres ultérieurs.

1.1. Comment définir la compétence langagière

1.1.1. Modèles d'utilisation du langage et de la compétence

Le langage en cours d'utilisation est un phénomène très complexe qui fait appel à un grand nombre de capacités ou de compétences différentes. Lors du démarrage d'un projet de test ou d'examen, il est important de disposer d'un modèle explicite de ces compétences et de la façon dont elles interagissent les unes avec les autres. Il n'est pas utile que ce modèle soit représentatif d'un grand courant concernant la façon dont la compétence langagière est organisée dans nos têtes ; son rôle est d'identifier des aspects de la compétence significatifs pour notre propos. C'est un point de départ qui permet de décider quels aspects de l'utilisation du langage ou de la compétence peuvent ou devraient faire l'objet d'un test ou d'un examen et cela aide à s'assurer que les résultats seront utiles et interprétables. La caractéristique mentale identifiée par les modèles est également appelée TRAIT ou CONCEPT.

1.1.2. Le modèle d'utilisation du langage du CECR

Des modèles de compétence langagière déterminants ont été proposés par différents auteurs (Bachman en 1990, Canale et Swain en 1981, Weir en 2005).

Il était logique que ce Manuel commence avec le modèle général d'utilisation du langage et de l'apprentissage proposé par le CECR. Cette APPROCHE ACTIONNELLE est présentée ainsi :

« L'usage d'une langue, y compris son apprentissage, comprend les actions accomplies par des gens qui, comme individus et comme acteurs sociaux, développent un ensemble de **compétences générales** et notamment une compétence à **communiquer langagièrement**. Ils mettent en œuvre les compétences dont ils disposent dans des **contextes** et des **conditions** variés et en se pliant à différentes **contraintes** afin de réaliser des **activités langagières** permettant de traiter (en réception et en production), des **textes** portant sur des thèmes, à l'intérieur de **domaines** particuliers, en mobilisant les **stratégies** qui paraissent le mieux convenir à l'accomplissement des **tâches** à effectuer. Le contrôle de ces activités par les interlocuteurs conduit au renforcement ou à la modification des compétences. » (CECR p.15, caractères gras dans le texte original).

Ce paragraphe identifie les éléments essentiels du modèle, qui sont présentés de façon plus détaillée dans le CECR. De fait, on peut considérer qu'un modèle hiérarchique comprenant des éléments emboîtés dans des éléments plus vastes est défini dans les titres et sous-titres des chapitres 4 et 5 du CECR.

A titre d'illustration, la figure 1 présente quelques titres et sous-titres du chapitre 5, *Les compétences de l'apprenant/utilisateur*. Elle montre la déclinaison des compétences : *Compétences générales* (telles que *Savoir* et *Savoir-être*, non présentés ici) et *Compétences communicatives langagières*, qui sont déclinées en trois : *compétences linguistiques, sociolinguistiques* et *pragmatiques*. Chaque entrée est ensuite subdivisée.

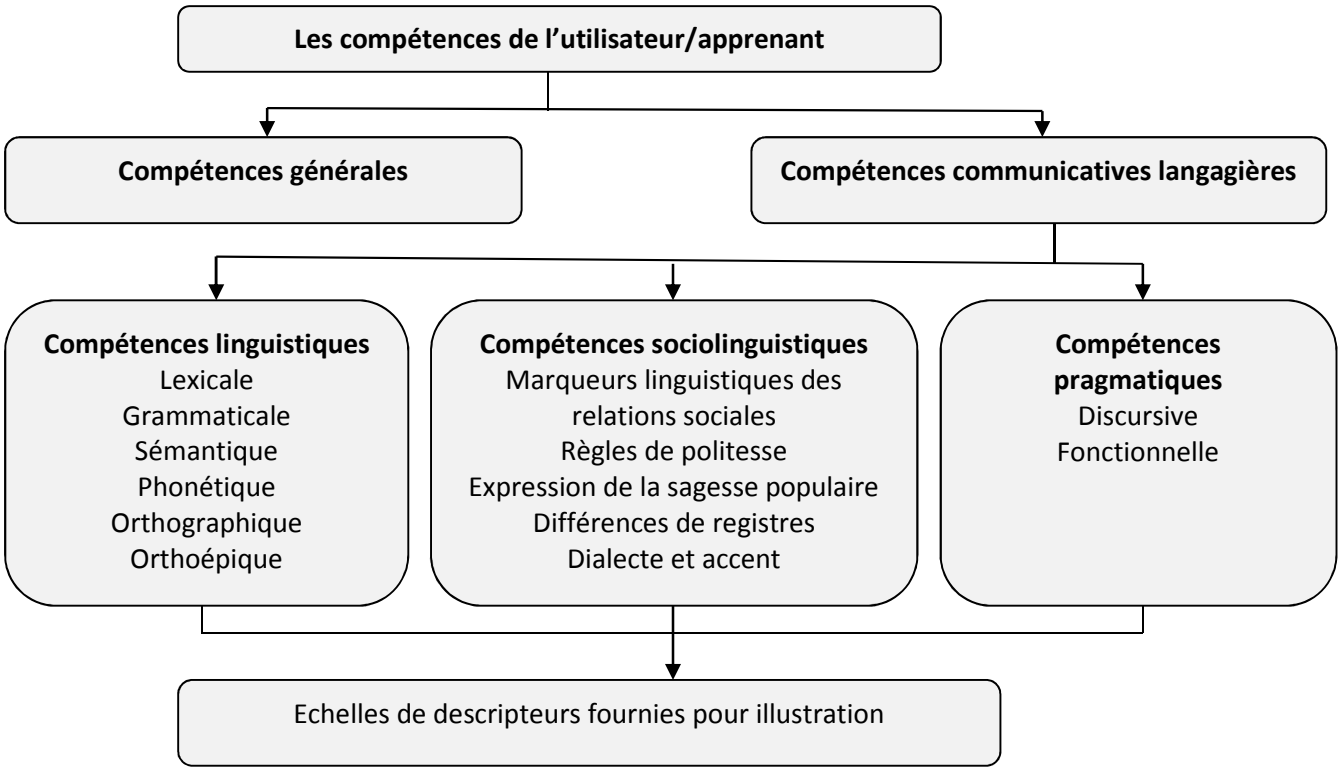


Figure 1. ***Vue partielle du chapitre 5 du CECR : Les compétences de l'utilisateur/apprenant***

Le chapitre 4 quant à lui, examine les objectifs communicatifs et les façons d'utiliser le langage. La figure 2 indique que cela implique de prendre en compte *ce* qui est communiqué (les thèmes, les tâches et les objectifs) ainsi que les activités de communication langagière et les stratégies et, partant, les capacités fonctionnelles du langage que les apprenants mettent en œuvre lorsqu'ils communiquent. Pour plus de clarté, la figure 2 n'illustre qu'une partie de cette hiérarchie complexe.

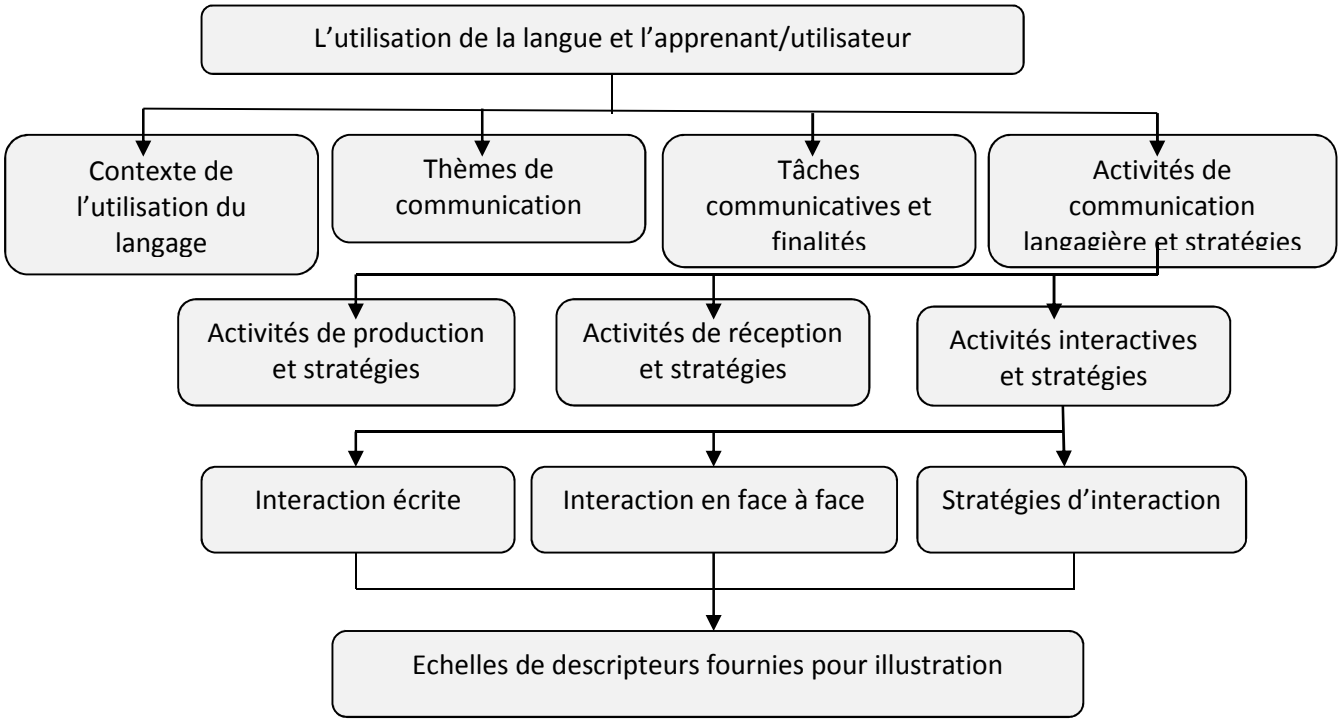


Figure 2. ***Vue partielle du chapitre 4 du CECR : L'utilisation de la langue et l'apprenant/utilisateur***

1.1.3. Rendre le modèle opérationnel

Lorsqu'on cherche à opérationnaliser le MODÈLE D'UTILISATION DU LANGAGE, deux paramètres importants ayant une influence considérable sur l'aspect final du test sont à prendre en compte : l'AUTHENTICITÉ des ITEMS et des TÂCHES et le caractère discret avec lequel les compétences sont évaluées.

Authenticité

Deux aspects importants de l'authenticité dans l'évaluation en langue sont l'authenticité *situationnelle* et l'authenticité *interactionnelle*. L'authenticité *situationnelle* se réfère à l'exactitude avec laquelle les tâches et les items représentent des activités langagières telles qu'on les trouve dans la vie quotidienne. L'authenticité *interactionnelle* se réfère au caractère naturel qu'il peut y avoir dans l'interaction que mène le candidat en accomplissant une tâche et les processus mentaux qui entrent en jeu. Un test de compréhension d'information spécifique, fondé sur des tâches, peut être rendu plus authentique au niveau situationnel si un contexte quotidien est créé, de type bulletin météo à la radio. Il peut gagner de l'authenticité au niveau interactionnel si on donne au candidat un objectif d'écoute, par exemple choisir, dans la semaine, le jour qui convient pour organiser un pique-nique.

Dans les tests de langue, lors de la création d'une tâche, il faut souvent composer entre les différents aspects de la fidélité. Il faut par exemple adapter des supports et des activités au niveau de compétence langagière de l'apprenant dans la langue cible. Cette adaptation signifie qu'alors que les supports peuvent ne pas être entièrement authentiques, les situations dans lesquelles les apprenants s'engagent ainsi que leur interaction avec les textes et entre eux peuvent, elles, être authentiques.

Pour que l'item ou la tâche soit le plus authentique possible, il faut identifier les caractéristiques de la tâche dans la vie réelle et les reproduire autant que possible. On peut obtenir une plus grande authenticité interactionnelle en :

- utilisant des situations ou des tâches vraisemblablement familières et pertinentes pour le candidat visé, à un niveau donné
- rendant claires pour le *public* visé, les raisons de mener une tâche spécifique à bien, par une contextualisation bien choisie
- rendant clairs les critères de réussite dans l'accomplissement de la tâche.

Compétences intégrées

Les compétences peuvent paraître indépendantes les unes des autres quand on définit un modèle d'utilisation du langage. Dans des tâches authentiques, il est cependant très difficile de les isoler. En effet, tout acte de communication implique l'utilisation de plusieurs compétences en même temps. Par exemple, lorsqu'un apprenant essaie de comprendre quelqu'un qui vient de l'arrêter dans la rue pour demander son chemin, plusieurs compétences entrent en jeu : les compétences grammaticales et textuelles pour décoder le message, la compétence sociolinguistique pour comprendre le contexte social de la communication, la compétence illocutionnaire pour mener à bien ce qu'il souhaite exprimer.

Lors de la conception d'une tâche destinée à un examen, il est essentiel de voir clairement les compétences requises pour une REPONSE correcte. Certaines compétences seront plus importantes que d'autres – ce sont celles-ci qui seront mises en avant dans la tâche. L'accomplissement de la tâche devra susciter suffisamment de réalisation langagière pour qu'un jugement puisse être porté sur la capacité du candidat dans la ou les compétences choisies. Il faut également prendre en compte la façon dont la réponse est CORRIGEE et notée (sections 2.5 et 5.13) : la correction doit porter uniquement sur la capacité dans la ou les compétences choisies.

1.1.4. Les niveaux du CECR

Accompagnant le modèle présenté ci-dessus, le CECR décrit un ensemble de six niveaux de compétence langagière communicative, qui permettent de fixer des objectifs d'apprentissage et de mesurer les progrès de l'apprentissage ou du niveau de compétence. Une série d'échelles de descripteurs affirmant *est capable de / peut*, illustre ce Cadre conceptuel.

Exemple d'affirmation pour le premier niveau (A1), en compréhension écrite :

Peut comprendre des noms familiers, des mots et des phrases très simples, par exemple sur des pancartes, des affiches ou des catalogues.

A comparer avec le descripteur du dernier niveau (C2) :

Peut comprendre aisément toute forme de langage écrit, y compris des textes abstraits, structurellement et linguistiquement complexes, comme des manuels et des articles spécialisés ainsi que des ouvrages littéraires.

Les six niveaux de compétence sont intitulés ainsi :

C2	Maîtrise	Utilisateur expérimenté
C1	Autonome	
B2	Indépendant	Utilisateur indépendant
B1	Niveau seuil	
A2	Intermédiaire	Utilisateur élémentaire
A1	Introductif	

Le concepteur de tests de langues doit avoir une bonne compréhension des affirmations *est capable de / peut*. Elles sont :

- Illustratives.

Elles ne sont donc pas :

- Exhaustives
- Prescriptives
- Une définition
- Un programme
- Une liste de contrôle.

Les descripteurs donnent des conseils aux éducateurs afin qu'ils puissent reconnaître les niveaux de compétence et en parler. On peut considérer qu'ils sont une indication pour élaborer un test mais les adopter ne signifie en aucun cas que le travail de définition des niveaux pour ce test a été achevé.

Il appartient aux organismes certificateurs de décider quels descripteurs correspondent le mieux à leur contexte. Ils doivent, par exemple, décider du DOMAINE de leur test : pour enseigner aux personnels d'un hôtel et les évaluer, les descripteurs de la « Coopération à visée fonctionnelle » peuvent être utiles (CECR 4-4.3.1) alors que les descripteurs ayant trait à « Comprendre des émissions de télévision et des films » (CECR 4-4.2.3) ne le seront probablement pas. Si les échelles descriptives disponibles ou si d'autres matériels de la boîte à outils du CECR ne conviennent pas suffisamment au contexte, il est possible de les compléter avec des descripteurs provenant d'autres sources ou d'en rédiger de nouveaux destinés à ce contexte.

Ancrer des tests ou des examens sur le CECR
Travaillant de cette façon, il est aisé de voir que le travail d'ancrage d'un test ou d'un examen sur le CECR débute par l'adaptation du CECR au contexte du test. Il est possible de faire cela parce que le CECR est à la fois « *hors contexte* afin de prendre en compte les résultats généralisables provenant de situations spécifiques différentes» et en même temps « *pertinent par rapport au contexte*, rattachable ou transposable dans chaque contexte pertinent » (CECR, p.23).

L'ancrage ne doit pas consister en une tentative d'appliquer de façon rigide et mécanique le CECR à n'importe quel contexte. Les organismes certificateurs doivent pouvoir justifier la façon dont ils ont rattaché ou transposé le CECR à leurs contextes, en partie en expliquant les caractéristiques de ces contextes.

Les caractéristiques des candidats sont d'autres points importants à prendre en compte. Les apprenants peuvent, par exemple, être très différents en termes d'âge et de développement cognitif, d'objectifs d'apprentissage, etc. En fait, quelques-unes de ces différences déterminent les caractéristiques des différents groupes d'apprenants. Les tests de langue sont souvent conçus pour l'un de ces groupes en particulier, par exemple pour de jeunes apprenants, ou pour des adultes. Les deux groupes peuvent être reliés au CECR, mais un B1 pour jeunes apprenants et un B1 pour adultes seront deux types différents de B1, car des descripteurs différents auront été appliqués.

Le profil de capacités des apprenants est souvent variable (certains seront meilleurs en réception orale qu'en réception écrite, d'autres seront le contraire). C'est pourquoi il est difficile de les comparer à l'aide d'une seule échelle. Deux candidats peuvent être placés en B1, mais pour des qualités et des points faibles différents. Il faut distinguer les aptitudes dans les différentes capacités, certaines pourront être évaluées à part et dans ce cas-là on utilisera les descripteurs spécifiques comme base pour définir les niveaux de compétence dans cette capacité particulière.

Il y a cependant une limite importante à l'adaptation du CECR à un contexte particulier. Le CECR a été uniquement prévu pour décrire la compétence langagière en fonction du modèle de l'utilisation du langage décrit au paragraphe 1.1.2 de ce Manuel. On ne doit pas essayer de RELIER des connaissances ou des capacités non prévues par ce modèle, comme, par exemple, la compréhension de la littérature en langue étrangère.

1.2. La validité

1.2.1. Qu'est-ce que la validité ?

On peut la définir de façon simple : un test est valide s'il mesure ce qu'il a l'intention de mesurer. Ainsi, par exemple, si notre test a l'intention de mesurer la compétence communicative en italien, et que les scores obtenus par les candidats soient systématiquement plus élevés ou plus faibles en fonction de leur compétence en italien, alors, notre test est valide. Cette définition plutôt étroite a été élargie ces dernières années afin d'inclure la façon dont les tests sont *utilisés*, ainsi, la validité se rapporte au : « degré de preuves et de théorie sous-tendant l'interprétation des scores entraînée par les utilisations données des tests » (AERA, APA, NCME 1999).

Cette définition élargie met l'accent sur l'IMPACT social des tests et la nécessité de fournir aux candidats des informations satisfaisantes afin qu'ils puissent éventuellement prendre des décisions importantes. De ce point de vue, il est impossible de dire d'un test qu'il est valide, au sens absolu. On dirait plutôt que la validité se rapporte à la façon dont les résultats à un test sont utilisés pour des besoins particuliers : c'est l'interprétation de la signification des résultats au test, par le candidat, qui le rend valide ou invalide.

Bachman rapporte cela au cas particulier du langage (1990), en déclarant que les tests devraient être adossés à un domaine de *l'utilisation de la langue cible*. Cela signifie que pour juger de la validité des résultats à un test, nous devons tout d'abord déterminer ce que nous attendons d'un candidat lorsqu'il utilise la langue dans la vie réelle, puis décider si le test apporte ou non la preuve de cette compétence. Le CECR propose une approche utile pour définir la réussite dans des domaines particuliers d'utilisation de la langue. Ses descripteurs sont un point de départ.

1.2.2. La validité et le CECR

Pour le CECR, lorsqu'on délivre les résultats obtenus à un test, cela signifie que l'on prétend être capable d'interpréter des performances en termes de définition des candidats à des niveaux particuliers du CECR. La validité permet de démontrer que ce que nous prétendons faire est la réalité : un apprenant évalué en B1 est réellement du niveau B1 conformément aux preuves que nous pouvons fournir.

Le type de preuve peut varier en fonction du contexte du test. Le modèle d'utilisation/apprentissage du langage du CECR présenté ci-dessus peut être appelé *sociocognitif* : le langage est à la fois un ensemble intériorisé de compétences et un ensemble externalisé de comportements sociaux. Selon le contexte, un test de langue mettra plus l'accent sur l'un ou l'autre aspect, et cela a une incidence sur la preuve de la validité :

- Si l'accent porte sur l'utilisation, la preuve de la validité se rapportera à l'utilisation réelle de la langue pour différents objectifs de communication.
- Si l'accent est mis sur la compétence, alors la preuve de la validité portera sur les capacités cognitives, les stratégies et la connaissance de la langue, qui étayent la preuve de la capacité potentielle pour l'utilisation de la langue.

Dans ce dernier cas, il est important de montrer que la réalisation des tâches sollicitent les mêmes capacités, les mêmes stratégies et les mêmes connaissances de la langue que celles dont on aurait besoin dans le domaine d'utilisation de la langue cible – ce qui signifie qu'elles ont une *authenticité interactionnelle* (cf. 1.1.3).

Les deux types de preuves peuvent étayer la validité liée au CECR. L'équilibre entre les deux dépend des exigences du contexte particulier. Le poids de l'utilisation de la langue pèsera certainement assez lourd dans la balance pour un test de langue destiné à des vendeurs alors qu'un test de langue pour des élèves mettra sans doute plus l'accent sur les compétences.

1.2.3. La validité dans le cycle d'élaboration du test

La validité, on l'a vu, fait un lien entre la performance dans la réalisation de la tâche et la preuve de la compétence langagière du candidat dans le monde réel. Concevoir et élaborer des tâches est manifestement une étape cruciale, mais d'autres étapes sont tout autant décisives.

Cette section traite de la validité dans le cycle de production des tests (cf. 1.5), afin que l'on puisse observer l'influence des phases de production. Cela signifie que l'ensemble des étapes est décrit de façon séquentielle, et que, si l'objectif final est que le test soit valide, chaque étape doit être finalisée de façon satisfaisante.

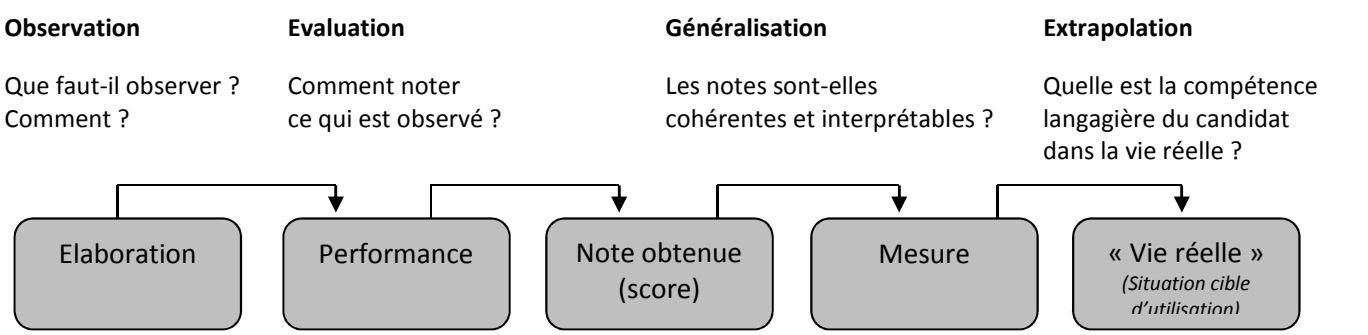


Figure 3. ***Chaîne du raisonnement pour une déclaration de validité (adapté de Kane, Crooks et Cohen 1999 ; Bachman 2005)***

La figure 3 illustre schématiquement ces étapes :

1. Le test ou l'examen est conçu de façon à obtenir un échantillon interprétable de la performance, fondée sur un modèle de compétence d'apprentissage. On peut par exemple demander à un candidat d'écrire une lettre à un ami sur un sujet particulier.
2. La performance au test est notée (elle obtient un score). Quels aspects de la performance seront valorisés ou au contraire pénalisés ? Dans l'exemple précédent, ces aspects seront liés à la compétence communicative décrite dans le modèle d'utilisation de la langue, incluant le REGISTRE (compétence sociolinguistique), la compétence lexicale, grammaticale et orthographique (compétences linguistiques), etc.
3. Jusqu'à ce point, les notes obtenues (ou les scores) sont des nombres qui représentent uniquement une performance isolée dans la réalisation d'une tâche spécifique. Comment peut-on les généraliser – le candidat obtiendrait-il le même résultat lors d'une autre passation, sur une version du test différente ? Cette question concerne la fiabilité (cf. Section 1.3). Un second aspect de la généralisation concerne l'ancrage à une échelle de compétence plus large, une version du test pouvant se révéler plus facile qu'une autre, il est nécessaire d'identifier et de compenser cela (cf. annexe VII).
4. Jusqu'à présent, nous avons décrit la performance en situation de test, mais nous souhaitons extrapoler aux situations hors test. A ce point, nous mettrons en relation une mesure avec un niveau du CECR, en décrivant ce que le candidat devrait être capable de faire dans la vie réelle, à l'aide des descripteurs appropriés.
5. En s'appuyant sur cela, il sera possible de prendre des décisions au sujet du candidat.

Il est clair, après ce bref exposé, que la validité, incluant une déclaration d'ancrage au CECR, dépend de chaque étape du cycle d'élaboration et de passation du test. La validité se construit tout au long de l'ensemble du processus.

L'annexe I propose des conseils pour élaborer une déclaration de validité.

1.3. La fiabilité

1.3.1. Qu'est-ce que la fiabilité ?

En évaluation, la fiabilité est synonyme de cohérence : un test qui a des résultats fiables produit les mêmes résultats ou des résultats similaires lors de différentes sessions. Cela signifie que le test classera un groupe de candidats de pratiquement la même façon. Cela *ne* signifie *pas* que les mêmes personnes réussiraient ou échoueraient, parce que le seuil de passation peut être modifié. On utilise généralement le terme de *fiabilité* lorsqu'on porte de l'intérêt à la cohérence et à l'exactitude des notes ou des résultats.

Une grande fiabilité n'implique pas nécessairement que le test soit bon ou que l'interprétation des résultats soit valide. Un mauvais test peut produire des notes (ou des scores) extrêmement fiables. Le contraire n'est pas vrai, bien que pour une interprétation valide des résultats, les notes *doivent* avoir une fiabilité acceptable, car sans cela, les résultats ne peuvent jamais être ni sûrs ni significatifs.

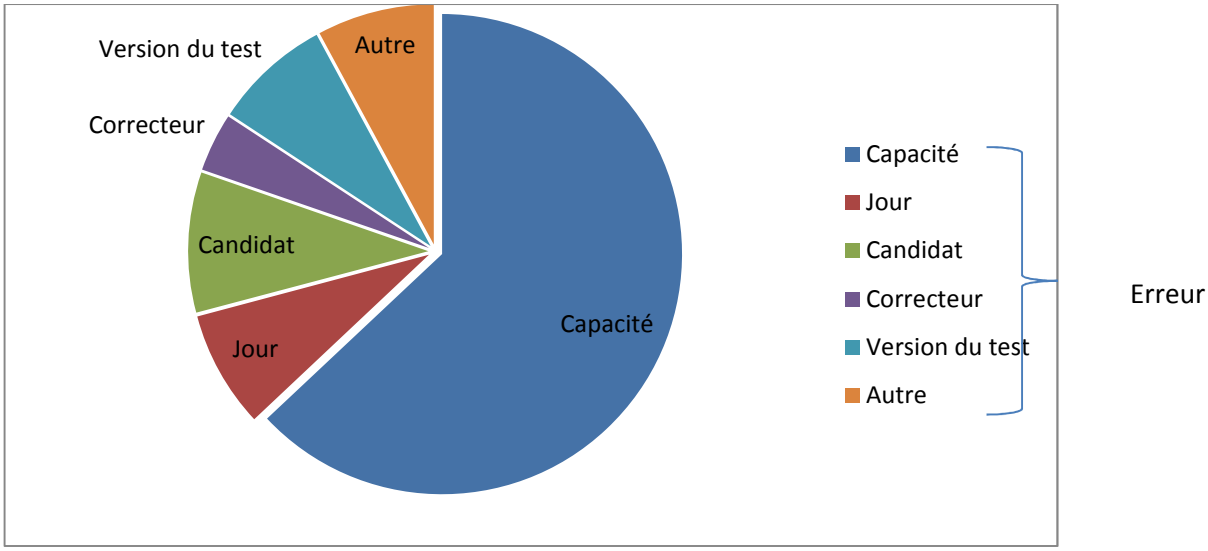


Figure 4. ***Sources d'erreur possible dans la notation d'un test.***

Les notes (ou le score) obtenues à un test varient en fonction des candidats. On définit la fiabilité comme la proportion de variabilité du score à ce test, due à la capacité mesurée et non à d'autres facteurs. La variabilité due à d'autres facteurs est appelée ERREUR. Il est à noter que cet emploi du terme « erreur » est différent de son usage habituel qui signifie souvent que quelqu'un est coupable de négligence. Tous les tests sont sujets à un certain degré d'erreur.

La figure 4 illustre quelques sources courantes d'erreur :

- Le jour de la session (le temps qu'il fait, la façon d'administrer, etc., peuvent varier).
- Le candidat peut être plus ou moins performant le jour du test.
- Les correcteurs de la version du test peuvent exécuter leur tâche de façon différente.
- Il peut y avoir d'autres facteurs incontrôlables.

Notre objectif est de produire des tests dans lesquels la proportion globale de la variabilité du score due à la capacité l'emporte sur celle due à une erreur.

1.3.2. La fiabilité en pratique

Le concepteur doit connaître les sources probables d'erreur et faire en sorte de les minimiser. Suivre les procédures et les principes décrits dans ce Manuel l'y aidera. Se servir de la statistique pour estimer la fiabilité des scores à un test constitue toutefois une importante étape post session. L'annexe VII précise ce qu'est l'estimation de la fiabilité.

On ne peut fixer d'objectif de fiabilité des scores pour tous les tests car les estimations de fiabilité dépendent du degré de variation des scores des candidats. Un test pour un groupe d'apprenants qui ont d'ores et déjà passé une procédure de sélection produira typiquement des estimations de fiabilité plus faibles que celles d'un test destiné à une population très variée. Les estimations de fiabilité peuvent également dépendre de l'item, de la question ou du type de tâche et de la façon dont elle est notée. Les scores des tâches évaluées (cf. Section 5) sont typiquement moins fiables que ceux des ITEMS DICHOTOMIQUES car davantage de variance (erreur) est introduite dans le processus d'évaluation que dans le processus administratif de notation.

Le fait d'étudier systématiquement la fiabilité sera utile pour identifier les tests qui ont bien marché par rapport à ceux qui ont moins bien marché ainsi que pour contrôler, au fil du temps, l'amélioration de la qualité. La plupart des estimations de fiabilité, telles que celles de l'Alpha de Cronbach ou le KR-20 avoisinent le 1. On considère souvent, de façon empirique, qu'une estimation située dans le tiers supérieur de l'amplitude (de 0.6 à 1) est acceptable.

L'estimation statistique de la fiabilité est généralement impossible lorsque le nombre de candidats et/ou d'items est faible. Dans ces cas, il est impossible d'estimer si la fiabilité convient aux objectifs du test. Dans ces situations, une bonne stratégie d'évaluation consiste à décider que le test n'est qu'un élément de preuve parmi ceux qui vont permettre de prendre des décisions. Un portfolio de travaux, d'autres tests passés pendant une période donnée ainsi que d'autres sources peuvent apporter des preuves supplémentaires.

1.4. Ethique et équité

1.4.1. Les conséquences sociales de l'évaluation : éthique et équité

Messick (1989) plaide en faveur du rôle critique des valeurs et des conséquences des tests comme étant partie intégrante de la fiabilité. Son influence a conduit à une plus grande attention envers la valeur sociale des tests ainsi qu'envers leurs conséquences pour les PERSONNES CONCERNÉES. Les effets et les conséquences des tests comprennent les résultats prévus (et heureusement positifs) de l'évaluation ainsi que les effets secondaires imprévus et parfois négatifs. L'apparition d'un nouveau test peut par exemple affecter (positivement ou négativement) la façon dont les enseignants enseignent (« l'impact »).

Il se peut que les organismes certificateurs conduisent des recherches sur les effets et l'impact afin d'en apprendre plus sur les conséquences sociales de leur test. On peut faire ce type de recherche à une toute petite échelle. En situation de classe, on peut voir si les étudiants privilégient certains aspects du programme aux dépends d'autres aspects, car ils mettent l'accent sur la passation du test. Il peut y avoir d'autres pistes pour stimuler le travail sur les aspects négligés, y compris en changeant l'objectif du test.

1.4.2. L'équité

Les organismes certificateurs ont comme objectif de rendre leur test le plus juste possible. Voir le *Code de pratiques pour une évaluation équitable en éducation* (JCTP 1988) et les *Standards pour une évaluation en éducation et psychologie* (AERA et al. 1999).

Les Standards de 1999 mentionnent trois aspects de l'équité : *l'équité en tant qu'absence de biais*, *l'équité en tant que traitement équitable dans le processus de l'évaluation* et *l'équité en tant qu'égalité dans les résultats de l'évaluation*.

L'ouvrage de Kunnan *Cadre de référence sur l'équité des tests* (Kunnan 2000a, 2000b, 2004, 2008) met l'accent sur cinq aspects de l'évaluation en langue, incontournables pour obtenir l'équité : *la validité* (cf. Section 1.2), *l'absence de biais* (cf. annexe VII), *l'accès*, *l'administration* (cf. Section 4) et *les conséquences sociales*.

De nombreux organismes ont rédigé des Codes de pratiques ou des Codes d'équité, pour aider les organismes certificateurs à gérer les aspects pratiques permettant d'assurer l'équité des tests.

Lors de la conception des tests et des examens, les organismes certificateurs peuvent essayer de minimiser les biais. Certains sujets (par exemple les coutumes locales) peuvent avantager ou désavantager certains groupes de candidats (par exemple ceux qui viennent de pays où les coutumes sont très différentes). On peut donner aux rédacteurs d'items une liste de sujets à éviter. Des groupes significatifs de candidats peuvent comprendre ceux qui sont définis par l'âge, le sexe ou la nationalité bien que cela dépende de la situation d'évaluation (cf. 3.4.1).

1.4.3. Préoccupations éthiques

On a commencé à s'intéresser aux préoccupations éthiques depuis le début des années 80. Spolsky en particulier (1981), a mis en garde contre les conséquences négatives que les tests de langue à forts enjeux pouvaient avoir pour des individus et a affirmé que les tests de langues devaient, au même titre que les médicaments, porter la mention « à utiliser avec précaution ». Il a en particulier mis l'accent sur un usage spécifique des tests de langues, par exemple dans le contexte de l'immigration, où les décisions prises sur la base des résultats à un test peuvent avoir de graves et radicales conséquences pour la personne.

L'Association internationale des tests de langue (ILTA) a publié son *Code d'éthique* en 2000 ; il propose des conseils généraux sur la façon dont les organismes certificateurs doivent se comporter à ce sujet.

Les organismes certificateurs doivent s'assurer de la bonne diffusion et compréhension des bons principes parmi les membres de leurs organisations. Cela permettra de s'assurer que l'organisme applique bien les directives proposées. D'autres mesures peuvent également convenir pour certains aspects de l'équité d'un test (cf. Section 4 et Annexe VII).

1.5. Organisation du travail

Les étapes de l'élaboration et de l'utilisation d'un test se présentent sous la forme d'un cycle dans lequel la réussite à une étape dépend des conclusions de l'étape précédente. C'est pourquoi il est important de bien gérer l'ensemble du cycle. Il faut également prendre en considération la collecte des preuves, puisqu'elles entreront en jeu dans les décisions importantes qui seront prises au cours du processus.

1.5.1. Les étapes du travail

La figure 5 illustre les étapes de l'élaboration d'un nouveau test ou d'un nouvel examen. Tout commence avec la décision de concevoir un test, prise par l'organisme certificateur ou quelqu'un d'autre, comme un directeur d'école, un bureau administratif ou un ministère. Vient ensuite l'étape d'élaboration du test, suivie par les étapes en liaison avec l'utilisation du test. La réalisation de chaque étape repose sur l'achèvement d'un grand nombre de petites tâches à l'intérieur de cette étape. L'ensemble de ces tâches est conçu pour répondre aux objectifs listés dans les cases de droite du diagramme. Une flèche de durée indique que les étapes se suivent de façon consécutive, car les données de sortie d'une étape sont nécessaires au démarrage de l'étape suivante. Une fois le test élaboré, les phases de l'utilisation peuvent être répétées un grand nombre de fois, en utilisant les données (les spécifications du test) de la phase d'élaboration. C'est ce qui permet l'élaboration de différentes FORMES EQUIVALENTES du même test.

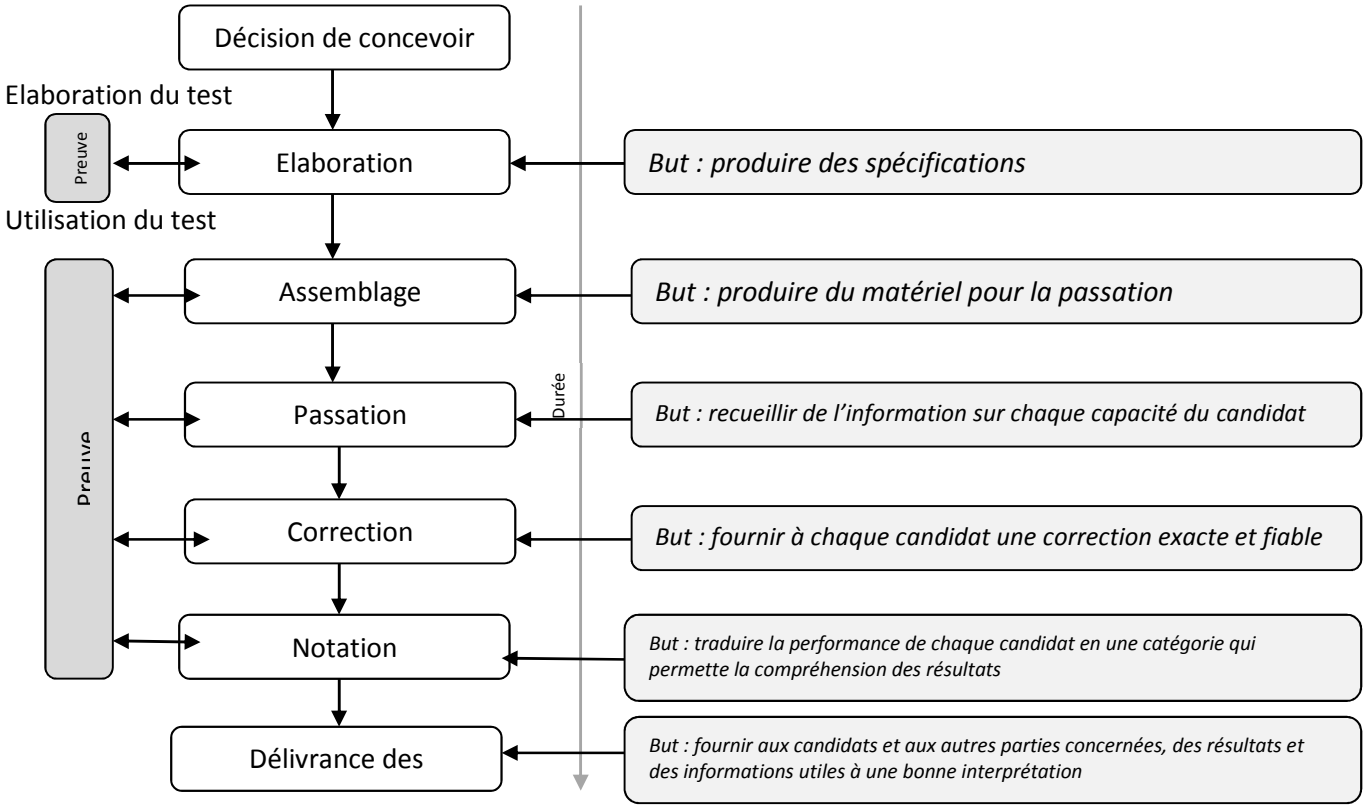


Figure 5. L'essentiel du cycle d'élaboration d'un test

Les étapes illustrées par la figure 5 s'appliquent à tout projet d'élaboration de test, quelle que soit la taille de l'organisme certificateur.

Chacune des étapes représentées dans la figure 5 comporte de nombreuses « micro-tâches » et de nombreuses activités. Elles sont détaillées dans les sections suivantes du Manuel. Le processus d'élaboration des tâches doit être standardisé afin de s'assurer que chaque test produit soit sensiblement semblable aux versions précédentes.

Le recueil et l'utilisation de preuves apparaît dans les cases de gauche du diagramme. Les preuves, qui peuvent être une information contextuelle concernant le candidat, un retour d'information de la part des personnes impliquées, les réponses des candidats aux tâches et items, le temps qu'ils ont mis à réaliser certaines tâches, sont importantes en tant que vérification continue du bon processus d'élaboration ainsi que, par la suite, pour faire la preuve de la validité des utilisations indiquées des résultats du test.

Veillez à systématiquement réunir et utiliser de telles preuves faute de quoi cette activité importante risque d'être oubliée au cours du processus d'élaboration.

1.6. Questions-clés

- Quels aspects du modèle d'utilisation du langage du CECR conviennent le mieux à votre situation ?
- Quels niveaux de compétence du CECR conviennent le mieux ?
- De quelle façon aimeriez-vous que les résultats obtenus à votre test soient compris et interprétés ?
- Dans votre situation, qu'est-ce qui mettrait la fiabilité le plus en danger ?
- De quelle façon pouvez-vous assurer que votre travail est à la fois éthique et équitable pour les candidats ?
- Quels défis devez-vous relever lors de l'organisation de votre cycle d'évaluation ?

1.7. Lectures complémentaires

MODÈLES D'UTILISATION DU LANGAGE
Fulcher et Davidson (2007 :36-51) débattent plus avant des concepts et des modèles.

VALIDITÉ
ALTE (2005 :19) propose un résumé utile des types de validité ainsi que le contexte de l'acception moderne de la validité.

Kane (2004, 2006, Mislevy, Steinberg et Almond (2003) examinent les questions liées aux argumentaires sur la validité (présentée dans l'annexe I de ce manuel) et donnent de plus amples conseils sur la façon de les développer.

FIABILITÉ
Traub et Rowley (1991) et Frisbie (1998) abordent tous deux simplement la fiabilité des notes (du score) obtenues. Parkes (2007) illustre quand et comment l'information venant d'un seul test peut être complétée avec d'autres preuves pour pouvoir prendre des décisions concernant les candidats.

ÉTHIQUE ET ÉQUITÉ
Depuis le début des années 90, des *Codes de pratique* spécialisés pour les évaluateurs ont également été rédigés par des associations professionnelles sur l'évaluation en langue, par exemple :

- Le *Code de pratique* d'ALTE (1994)
- Les *Conseils pour une pratique* d'ILTA (2007)
- Les *Conseils pour une bonne pratique dans les tests de langue et l'évaluation* d'EALTA (2006)

Dans les années 90, une édition spéciale des *Tests de langue*, avec la contribution d'Alan Davies (1997) mettait l'accent sur l'éthique dans les tests de langue et une conférence sur *L'éthique dans l'évaluation en langues* était organisée en 2002 à Pasadena. Les actes de cet événement ont permis une édition spéciale de la revue trimestrielle *L'évaluation en langues* (également avec la contribution d'Alan Davies en 2004). McNamara et Roever (2006) ont listé les revues sur l'équité et les Codes d'éthique pour les examens.

Plusieurs articles de *L'évaluation en langues* d'avril 2010 mettent l'accent sur la collecte des preuves pour l'équité des tests et sur la façon de les présenter sous forme d'argumentaire (Davies 2010, Kane 2010, Xi 2010).

2. L'élaboration du test ou de l'examen

2.1. Le processus d'élaboration

L'élaboration du test ou de l'examen a pour objectif de produire des **spécifications** qui serviront à concevoir les **examens passés par les candidats (l'épreuve finale)**. Cette élaboration commence quand une personne ou une organisation (le commanditaire) décide qu'un nouvel examen est nécessaire. La figure 6 décrit le processus d'élaboration de l'examen qui comprend trois étapes indispensables (la planification, la conception, l'expérimentation) et une autre étape (l'information auprès des parties prenantes) qui peut s'avérer nécessaire selon les contextes. C'est que la diffusion de l'information ne fait pas partie, comme les autres étapes, de l'élaboration des spécifications. Son objectif est avant tout d'informer les intéressés de l'existence du nouvel examen.
Un diagramme plus détaillé est disponible en Annexe II.

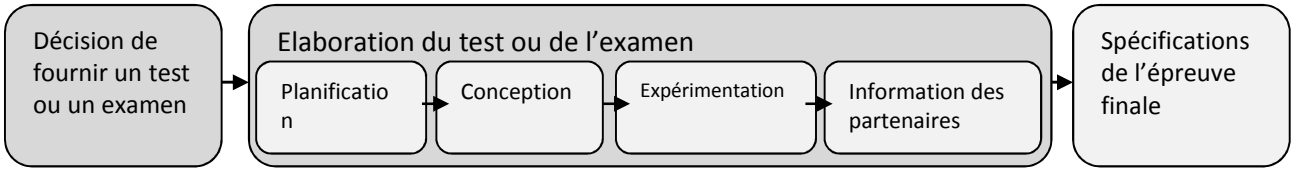


Figure 6 ***Processus d'élaboration de l'examen***

2.2. La décision de produire un test ou un examen

Cette décision ne fait pas vraiment partie du processus d'élaboration mais elle fournit des informations importantes dans la mesure où les besoins exprimés par le commanditaire auront une influence déterminante sur la conception de l'examen et son utilisation.

Qui décide de la nécessité d'un nouvel examen ? Dans certains cas la décision vient de l'organisme certificateur qui se charge du processus d'élaboration. Elle peut aussi venir d'un commanditaire qui a besoin d'un nouvel examen.

Dans les deux cas, il faut que les besoins soient clairement identifiés, ce qui suppose un travail supplémentaire de la part de ceux qui vont élaborer l'examen. Il est souvent plus difficile de comprendre les intentions d'un commanditaire qui ne fait pas partie de l'organisation qui produit l'examen ou qui n'a aucune expertise en évaluation ou dans l'enseignement des langues. Dans ce dernier cas, il ne connait pas les informations dont un concepteur a besoin.

2.3. La planification

Cette étape consiste à rechercher les informations nécessaires dans les étapes ultérieures. En principe la plupart de ces informations devraient être fournies par le commanditaire. Il est cependant recommandé de s'adresser aux parties prenantes telles que les différents ministères concernés, les éditeurs, les établissements scolaires, les parents, les experts, les employeurs, les centres d'enseignement et les administrations. Si un nombre important de personnes doit être consulté, il faut préparer des questionnaires et organiser des séminaires pour transmettre l'information désirée. Par contre, en situation de classe, la connaissance personnelle du contexte et des candidats est suffisante.

Les concepteurs de test ou d'examens doivent impérativement poser les questions suivantes :

- Quelles sont les caractéristiques des candidats qui vont passer le test ou l'examen (âge, genre, situation sociale, niveau d'études, langue maternelle, etc.) ?
- Quel est l'objectif du test ou de l'examen? (certificat de fin de scolarité, admission à un programme d'enseignement, minimum requis dans un domaine professionnel, évaluation formative ou diagnostic, etc.) ?
- Quel est le contexte éducatif dans lequel s'inscrit le test ou l'examen? (un programme, une approche méthodologique, des objectif d'apprentissage, etc.)
- Quelle est la norme requise par l'objectif proposé? (un niveau du CECR dans certaines capacités langagières, dans toutes les compétences, une norme reliée à un domaine spécifique, etc.) ?
- Comment les résultats du test ou de l'examen seront-ils utilisés ?

Le concepteur du test ou de l'examen pourra, grâce aux réponses apportées aux questions précédentes, commencer à définir les capacités langagières à évaluer, décider des points de césure (cf. partie 5) et de la façon de présenter et d'expliquer les résultats aux utilisateurs (voir partie5).

Les questions portant sur l'impact du test ou de l'examen peuvent être utiles :

- Qui sont les parties prenantes ?
- Quel type d'impact est recherché?
- A quel impact peut-on s'attendre ?

Et enfin, il ne faut pas oublier des questions d'ordre pratique :

- Combien de candidats sont attendus ?
- A quel moment le test ou de l'examen doit-il être prêt ?
- Comment le test ou de l'examen sera-t-il financé et quel est le budget alloué ?
- Combien de fois l'examen sera-t-il passé ?
- Où va-t-il être passé ?
- Sous quelle forme doit-il être livré ? (par exemple, papier ou électronique)
- Quel sera le ou la responsable de chaque étape d'élaboration du test ou de l'examen? (par exemple la production du matériel et la conception du test ou de l'examen, la passation, la notation, la communication des résultats)
- Quelles seront les implications en termes de sécurité (par exemple sera-t-il nécessaire d'utiliser une ou plusieurs versions du test ou de l'examen?)
- Comment le suivi à long terme va-t-il être assuré ?
- Est-ce qu'il sera possible de faire des prétests ?
- Quelles sont les implications en termes logistiques ? (par exemple, l'organisme certificateur devra-t-il prendre en compte la situation d'autres institutions telles que les centres d'examens ?)

2.4. La conception

L'étape de la conception commence une fois que toutes les informations de la précédente étape ont été recueillies. C'est le moment de prendre des décisions importantes sur la nature du test et d'élaborer les premières spécifications. Ces spécifications décrivent la structure d'ensemble du test et les différentes parties du contenu. Les spécifications détaillées qui concernent les rédacteurs d'items ainsi que les personnes impliquées dans la distribution des tests et l'organisation de leur passation peuvent être rédigées une fois que les premières spécifications ont été agréées.

2.4.1 Premières préoccupations

Le premier défi à relever dans cette étape est d'avoir une idée précise du contenu du test ou de l'examen et de son format. On partira des informations recueillies concernant les besoins et le contexte : caractéristiques des candidats, objectif du test et le niveau de capacité langagière requis.

Le CECR et tout particulièrement les chapitres du CECR consacrés à l'évaluation sont une source précieuse d'informations pour définir les caractéristiques du test ou de l'examen.

- Le chapitre 6 sur l'apprentissage et l'enseignement des langues concerne les objectifs d'apprentissage et la méthodologie de l'enseignement, deux aspects qui ont un impact sur le type, le contenu et la fonction des tests ou des examens.
- Le chapitre 7 sur les TACHES et leur rôle dans l'enseignement des langues influe sur la façon de les utiliser dans l'évaluation.
- Le chapitre 9 sur l'évaluation traite de la façon d'utiliser le CECR en fonction des différents objectifs d'évaluation.

Les chapitres 4 et 5 qui traitent du contenu du test et des capacités langagières à évaluer sont les plus pertinents. Ils offrent au concepteur de test ou d'examen un large éventail d'options à choisir dans l'approche actionnelle et le modèle de langue en usage (cf. 1.1) proposés dans le CECR. Cela concerne par exemple:

- l'objet principal de la tâche : la compréhension détaillée d'un texte, etc. (cf. chapitre 4.4 et 4.5 du CECR) ;
- l'objet de l'évaluation : les capacités langagières, les compétences et stratégies (cf. CECR chap. 5) ;
- les genres et les types de textes utilisés comme supports (cf. CECR chap. 4.1 et 4.6) ;
- des propositions de thèmes (cf. CECR chap. 4.1 et 4.2) ;
- des types de déclencheurs utilisés dans des tests de production orale (cf. CECR chap. 4.3 et 4.4) ;
- des types de situations de la vie quotidienne familières aux candidats (cf. CECR chap. 4.1 et 4.3) ;
- le niveau de performance correspondant à ces situations (voir les nombreux niveaux de « savoir-faire » (can dos) du CECR ;
- des critères pour évaluer des tâches d'écriture créative et des tests de production orale (voir les niveaux correspondant représentatifs de « savoir-faire » (can dos) du CECR par exemple pages 58 et 74, etc.).

L'organisme certificateur doit également préciser les caractéristiques techniques du test ou de l'examen, à savoir :

- la durée. Un candidat moyen devrait disposer d'assez de temps pour répondre à tous les items du test ou de l'examen sans avoir à se presser. L'essentiel est que les candidats aient l'occasion de montrer leur capacité réelle. Il est sans doute nécessaire qu'un évaluateur expérimenté s'en charge mais quelques échantillons peuvent être consultés (cf. 2.8 « Lectures complémentaires »). La durée peut être modifiée après expérimentation ou passation en situation réelle). Il arrive que des tests minutés soient utilisés, dans lesquels on demande aux candidats de répondre en un temps limite aux items. Dans ce cas aussi, une expérimentation doit avoir lieu ;
- le nombre d'items ou de questions. Il faut en avoir assez pour couvrir le contenu nécessaire et pouvoir donner une appréciation fiable des capacités du candidat. La longueur du test ou de l'examen est cependant limitée pour des raisons pratiques ;
- le nombre d'items par partie. Si le test ou l'examen a pour objectif de mesurer de façon fiable les différents aspects de la capacité langagière, il faut un nombre suffisant d'items par partie. On peut consulter des échantillons et calculer la fiabilité. (cf. annexe VII) ;
- le type d'items. Des items peuvent induire des réponses à choisir ou à fournir. Les items à choix de réponse sont les questions à choix multiple, les appariements ou les classements. Dans les items comportant des réponses à donner, celles-ci peuvent être courtes (exercices de phrases à compléter par un mot ou plus). Pour connaître les avantages et les inconvénients des différents types d'items, consulter ALTE (2005 :111-34) ;
- la longueur totale des textes et celle de chaque texte mesurée en nombre de mots. Des exemples peuvent donner une idée de la longueur communément admise (cf. 2.8 « Lectures complémentaires ») ;
- le format. Un examen « à items discrets » consiste en un examen comprenant des items indépendants les uns des autres. Dans un test conçu sur le principe des tâches, les items sont groupés et ont par exemple pour support un texte de compréhension orale ou écrite. Ces tests conçus à partir de tâches conviennent en général beaucoup plus à l'évaluation de type communicative car les stimuli utilisables sont plus longs et plus authentiques.(Pour plus d'informations sur les types d'items, voir ALTE 2005 :135-47) ;
- le nombre de points à donner à chaque item et à chaque tâche ou partie sachant que leur importance grandira avec le nombre de points qui leur sera attribué. On recommande en général d'attribuer un point par item. Il est parfois nécessaire de donner plus de poids à tel ou tel item. (cf. annexe VII) ;
- les caractéristiques des ECHELLES DE NOTATION. Va-t-on procéder par tâches, quelle sera l'éventail de l'échelle, cette échelle sera-t-elle analytique ou holistique ? (cf. 2.5 et 5.1.3 où il est question des échelles de notation).

L'étape de conception se termine une fois que seront prises les décisions concernant les objectifs du test ou de l'examen, les capacités langagières et les contenus à évaluer ainsi que les détails techniques de son utilisation. Il faut aussi penser à l'évaluation des tâches, à l'élaboration des échelles de notation des productions orales et écrites, (cf. 2.5), à la façon d'organiser la passation des tests ou des examens (cf. partie 4) et à la formation des correcteurs et des examinateurs (cf. 5.1.3). Toutes les parties prenantes devraient alors revoir ces propositions de façon détaillée afin de pouvoir en faire une estimation sérieuse.

Il faut également prendre en compte la communication avec les candidats et les parties prenantes sur les sujets suivants :

- le nombre d'heures requis si des cours de préparation au test ou à l'examen sont nécessaires ;
- la mise à disposition d'exemples de tests ou d'examen ;
- l'information à transmettre aux utilisateurs (toutes les parties prenantes concernées) avant et après le test ou l'examen.

Enfin, la prise en compte des attentes des partenaires :

- l'adéquation du test ou de l'examen avec le système en place en termes d'objectifs de programme et de pratique de classe ;
- l'adéquation du test ou de l'examen avec les attentes des parties prenantes.

Le chapitre 4 du CECR fournit un schéma de référence très utile qui met l'accent sur les caractéristiques de tout test ou examen en voie d'élaboration. Un diagramme en reprend l'essentiel. Cette approche est illustrée dans l'annexe III de ce Manuel. L'examen donné en exemple est destiné à des candidats de niveau B2, apprenant la langue en contexte professionnel. Il comprend quatre parties. On y trouve à la fois une vue d'ensemble du contenu de l'examen et une description générale de chaque partie.

2.4.2 Comment tenir compte à la fois des exigences propres au test ou à l'examen et des considérations d'ordre pratique

A cette étape de l'élaboration du test ou de l'examen, il faut mettre en rapport la structure proposée avec les contraintes d'ordre pratique. Le détail de ces contraintes est recueilli à l'étape de la planification, en même temps que les exigences

propres à l'examen (partie 2.3). Le concepteur doit concilier les exigences et les contraintes, et avoir l'accord du commanditaire. Pour ce faire, Bachman et Palmer (1996, chap. 2) proposent un cadre traduisant le concept d'utilité du test.

Selon eux, les qualités propres à ce concept sont:

- La validité : les interprétations des notes obtenues ou d'autres résultats sont significatives et appropriées.
- La fidélité : les résultats fournis sont constants et stables.
- L'authenticité : les tâches reflètent des situations langagières de la vie réelle dans les centres d'intérêt de l'utilisateur.
- L'interactivité : les tâches mettent en œuvre les mêmes processus et stratégies que celles mises en œuvre dans des tâches de la vie réelle.
- L'impact : l'effet du test ou de l'examen, que l'on espère positif, sur les personnes, les pratiques de classe et plus largement la société.
- L'application : on doit pouvoir élaborer, produire et organiser la passation du test ou de l'examen tel qu'il est planifié avec les ressources disponibles.

Il se peut que ces qualités se contredisent : ainsi plus une tâche est authentique moins elle est fidèle. C'est pour cette raison qu'il faut constamment rechercher un équilibre qui renforce l'utilité du test dans son ensemble.

2.4.3 Spécifications du test ou de l'examen

Le résultat de l'étape d'élaboration constitue un ensemble complet de spécifications. La première version de ces spécifications inclut des décisions concernant une grande partie des points abordés ci-dessus. La version finale des spécifications sera rédigée après l'étape d'expérimentation (cf. 2.5). Les spécifications sont d'autant plus importantes que l'enjeu du test ou de l'examen est grand. Elles sont l'outil même qui atteste de la qualité du test ou de l'examen et montre aux autres que les interprétations des résultats sont valides.

Mais les spécifications sont tout aussi importantes pour les tests ou les examens dont l'enjeu est moindre. Elles sont une garantie que les formes du test ou de l'examen ont les mêmes bases et qu'il tient rigoureusement compte du programme et du contexte d'évaluation.

La rédaction des spécifications peut varier en fonction des besoins de l'organisme certificateur et de la population concernée. Les modèles de spécifications élaborés (cf. 2.8 « Lectures complémentaires »,) peuvent servir de référence.

2.5 L'expérimentation

L'objectif de cette étape est de «tester sur le terrain» les premières versions des spécifications afin de faire les changements nécessaires en tenant compte des résultats de l'expérience et des propositions des parties concernées. Une fois les spécifications rédigées, on passe à la fabrication d'échantillons du matériel. Pour ce faire, on peut se référer à la partie 3 de ce Manuel. On peut collecter ce matériel de différentes façons :

- **faire un test pilote** (demander à quelques candidats de passer le test ou l'examen) et analyser les réponses données (cf. 3.4 et VII). ;
- consulter des collègues ;
- consulter d'autres parties prenantes.

Le test pilote doit être proposé à des candidats dont les caractéristiques (âge, sexe, …) sont les mêmes que celles des candidats au test ou à l'examen final. La passation du test pilote doit avoir lieu dans les mêmes conditions que celles de l'épreuve finale. Mais même si toutes les conditions ne sont pas remplies (par exemple, manque de temps pour faire passer tout le test, nombre insuffisant de candidats), la phase pilote sera quand même utile. Elle peut fournir des renseignements sur la durée à allouer à chaque tâche, sur la clarté des consignes accompagnant les tâches, sur la mise en page pour les réponses, etc. Pour la production orale, il est recommandé d'observer (par exemple en les enregistrant) les performances orales.

La consultation des collègues ou des parties concernées peut se faire de différentes façons. Soit en face à face s'il s'agit de petits groupes, soit sous forme de questionnaires ou de rapports d'enquête s'il s'agit de projets plus importants.
Les renseignements que fournit cette phase pilote peuvent également donner lieu à la conception de **graphiques** et d'échelles de notation assez détaillés (cf. 5.1.3 pour les éléments de ces échelles). Les performances des candidats sont les plus à même d'illustrer les niveaux de compétences et de fournir ces éléments. C'est à partir de ces derniers que seront rédigés les descripteurs de chaque niveau. Une fois élaborées, les échelles de niveau doivent passer par l'étape du test pilote

et une analyse à la fois qualitative et quantitative doit être faite sur la façon dont elles ont été utilisées par les examinateurs (cf. annexe VII). Il se peut que d'autres tests pilotes et des modifications soient nécessaires.

Il faudra peut-être mener d'autres recherches pour répondre aux questions qui se sont posées durant l'étape d'expérimentation. Les données du test pilote peuvent y répondre et des études spécifiques peuvent être entreprises. On peut par exemple se demander :

- si les types de tâches que nous voulons utiliser conviennent à la population qui va passer le test (par exemple des enfants) ;
- si les types de tâche correspondent au domaine ciblé (par exemple le tourisme ou le droit ;
- si les items et les tâches évaluent véritablement la compétence concernée ? Des techniques statistiques peuvent être utilisées pour décider à quel point les items et les tâches choisies évaluent les différents aspects de l'activité langagière (cf. annexe VII) ;
- si les examinateurs vont être capables d'interpréter et d'utiliser correctement les échelles de notation et les critères d'évaluation ;
- quand un test doit être révisé, s'il est nécessaire de faire une étude de comparabilité pour s'assurer que le nouveau test ou le nouvel examen fonctionnera de la même façon que le précédent ;
- si les items et les tâches font appel aux processus cognitifs prévus du candidat. On peut s'en assurer en mettant en place des protocoles verbaux au cours desquels les apprenants expriment ces processus quand ils accomplissent ces tâches.

La rédaction des spécifications peut donner lieu à plusieurs versions avant la version du test ou de l'examen final.

2.6. L'information des parties concernées

Les spécifications peuvent être à usage multiple : servir aux rédacteurs d'items et aux enseignants qui veulent préparer leurs élèves à l'examen et adapter leurs programmes. Cela implique l'élaboration de différentes versions en fonction des acteurs concernés. On peut par exemple rédiger à l'intention de ceux qui préparent à l'examen, une version simplifiée comprenant les éléments linguistiques (lexique, grammaire…), les thèmes, le format, etc. Une version beaucoup plus élaborée sera destinée aux rédacteurs d'items.

Outre ces spécifications, les parties concernées voudront voir des exemples d'épreuves (pour plus d'information sur la fabrication du matériel, voir partie 3). Ces exemples peuvent comporter, non seulement les épreuves sur papier mais aussi les enregistrements audio ou vidéo des épreuves de compréhension orale. Ce matériel peut servir en classe à la préparation à l'examen. Par la suite, des épreuves d'examens déjà passées pourront être utilisées.

Les réponses données dans les tâches de production orale et écrite doivent également faire partie des échantillons, à condition que le matériel ait été prétesté ou qu'il ait déjà été utilisé dans un test ou un examen final. On peut aussi donner des conseils aux candidats pour les aider à se préparer.

Quel que soit le matériel, il doit parvenir aux intéressés en temps utile, bien avant les dates de passation. Il en est de même pour les règlements, les responsabilités des uns et des autres, les calendriers.

2.7. Questions clés

- Qui a pris la décision d'organiser un test ou un examen et pour quel objectif et quel usage?
- Quel sera l'impact en termes d'enseignement et sur la société ?
- Quel type et quel niveau de performance langagière doivent être évalués ?
- Quel type de tâches est nécessaire pour y arriver ?
- Quelles sont les ressources pratiques disponibles ? (locaux, personnel…)
- Qui doit faire partie de l'équipe de rédaction des spécifications et d'élaboration des éléments des échantillons du test ou de l'examen ? (en termes d'expertise, d'influence, d'autorité, etc.)
- En quels termes le contenu, les détails techniques et de procédure du test ou de l'examen seront-ils décrits dans les spécifications ?
- Quel type de renseignements doit-on donner aux utilisateurs (une version publiable des spécifications) et comment la diffuser ?
- Comment le test ou l'examen peut-il être expérimenté ?
- Comment les parties prenantes peuvent-elles s'informer sur le test ou l'examen, ?

2.8 Lectures complémentaires

Ceux qui sont impliqués dans l'élaboration de tests ou d'examens et qui veulent comprendre les niveaux du CECR, trouveront de nombreux exemples dans le CECR lui-même. Voir le Conseil de l'Europe (2006a, b; 2005), Eurocentres / Fédération des Coopératives Migros (2004), University of Cambridge ESOL Examinations (2004), CIEP / Eurocentres (2005), Bolton, Glaboniat, Lorenz, Perlmann-Balme et Steiner (2008), Grego Bolli (2008), Conseil de l'Europe et CIEP (2009), CIEP (2009).

Des exemples d'intitulés de spécifications sont disponibles dans Bachman et Palmer (1996:335–34), Alderson, Clapham et Wall (1995:14–17) et Davidson et Lynch (2002:20–32).

Des modèles de grilles décrivant et comparant les tâches sont disponibles : voir les membres de ALTE (2005a, b; 2007a, b), Figueras, Kuijper, Tardieu, Nold and Takala (2005).

3 Assemblage du test ou de l'examen

3.1. Le processus d'assemblage

L'objectif de l'étape d'assemblage est de fournir les éléments selon les données des spécifications afin que test ou l'examen soit prêt dans les délais. Le processus d'assemblage comprend trois grandes étapes, représentées dans la figure 7.

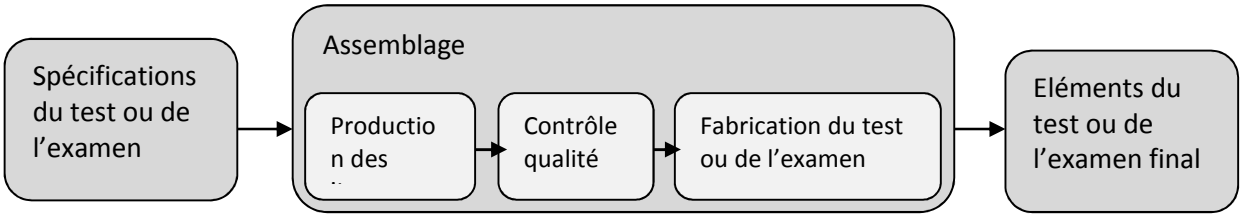


Figure 7 ***Les grandes étapes du processus d'assemblage du test ou de l'examen***

Dans ce schéma, la production des épreuves du test ou de l'examen et la fabrication du test ou de l'examen lui-même sont considérés comme deux étapes distinctes car cette distinction permet de clarifier les objectifs de chaque étape. Il est cependant possible d'en faire une seule étape. L'essentiel est le contrôle qualité qui peut entraîner, selon le résultat, des modifications.

3.2 Les premiers pas

Avant de produire les items, il faut :

- ✓ recruter et former des rédacteurs d'items ;
- ✓ assurer la gestion de cette production.

3.2.1 Le recrutement et la formation des rédacteurs d'items

Il se peut que les rédacteurs d'items soient ceux qui élaborent aussi les épreuves du test ou de l'examen. Dans ce cas, le problème du recrutement et de la formation ne se pose pas car ils sont familiarisés avec le test.

Dans le cas où il faut recruter des rédacteurs, le concepteur de test doit penser aux exigences professionnelles requises pour ce travail. La compétence dans la langue concernée et la connaissance du contexte d'évaluation font partie des critères. Il n'est pas nécessaire d'inclure parmi les critères une bonne connaissance des tests ou examens existants ou des principes de l'évaluation dans la mesure où une formation peut être mise en place (voir ALTE 2005) pour ce faire. La formation, le suivi et l'évaluation continus assureront le perfectionnement professionnel du rédacteur d'item.

Les professeurs de langue sont souvent les mieux placés pour être de bons rédacteurs dans la mesure où ils ont une très bonne compréhension des apprenants et de la langue à évaluer. Ce travail leur conviendra d'autant plus qu'ils auront préparé leurs élèves à la passation d'un test ou d'un examen similaire ou ont été impliqués comme correcteur ou examinateur d'épreuves orales. On peut leur demander de rédiger soit toutes les épreuves soit une partie selon les besoins de l'organisme certificateur.

3.2.2 La gestion des items produits

L'organisme certificateur doit mettre en place un dispositif permettant de collecter, stocker et traiter les items. Ce dispositif est d'autant plus indispensable que le nombre d'items et de tâches est important. Tous les items doivent être soumis au même traitement d'assurance qualité tel que la vérification et la passation du test pilote. Il est donc indispensable de connaître l'historique du traitement de chaque item et ce à tout moment. Cela est indispensable quand la production d'items et le nombre de personnes impliquées à chaque étape sont importants. Tout dispositif devrait au moins comprendre :

- ✓ un numéro d'identification pour chaque item ;
- ✓ une liste de contrôle comprenant les étapes achevées, les modifications et d'autres renseignements ;
- ✓ un moyen de s'assurer de la possibilité d'accès aux items et aux renseignements concernés et de l'impossibilité d'accès aux versions des étapes antérieures. La meilleure façon de s'en assurer est de les stocker au même endroit ou de les transférer par courriel après chaque étape de rédaction et de mise en forme.

3.3 La production des items

On demande aux rédacteurs de produire les items qui seront utilisés dans l'épreuve finale. Dans le Manuel, nous appelons cela « passation de la commande ». Les rédacteurs trouveront dans l'annexe IV les renseignements qui peuvent les aider à accomplir cette tâche. Ils doivent connaître le nombre et le type d'items ainsi que les dates de remise exigées.

Cette partie concerne essentiellement la description du matériel requis et les moyens de communiquer aux rédacteurs le travail exigé. Un compte à rebours à partir de la date de passation du test final permet de décider de la date de remise des éléments demandés.

3.3.1 L'évaluation de la demande

Pour fabriquer un test ou un examen, les organismes certificateurs doivent pouvoir faire leur choix parmi les items et les tâches produits. Il est difficile de connaître le nombre exact d'items ou de tâches requis dans la mesure où dans la constitution d'un test ou d'un examen il faut prendre en compte les types d'items, le thème, le niveau de langue (cf. 3.5). Par conséquent, il faut demander, lors de la commande, plus d'items que ceux qui seront utilisés dans le test, d'autant que l'on sait que certains seront rejetés à l'étape du contrôle qualité.

3.3.2 La commande

La commande peut répondre à la nécessité d'avoir des items pour la passation d'un test ou d'un examen ou à la constitution d'une banque d'items qui serviront ultérieurement à la constitution d'un test ou d'un examen. Dans les deux cas, il faut prévoir les délais de production.

Il faut se mettre d'accord sur un certain nombre de paramètres et formaliser cet accord, afin d'éviter tout malentendu. Il est nécessaire, quand beaucoup de rédacteurs d'horizons divers sont impliqués, d'établir une liste officielle des exigences. Les points ci-dessous, utiles quel que soit le cas de figure, doivent être clairement et formellement indiqués.

Précisions sur les éléments attendus

- ✓ Indiquer le nombre de textes, de tâches et d'items requis ;
- ✓ s'agissant des textes, indiquer si les items doivent être rédigés en même temps que le texte ou s'il faut attendre que le texte soit accepté pour le faire ;
- ✓ pour la production orale avec un déclencheur visuel, indiquer s'il faut fournir le déclencheur visuel et dans ce cas quels types de déclencheurs sont requis ;
- ✓ informer sur les problèmes de droits de reproduction d'images ou de textes et la façon de les traiter ;
- ✓ préciser qu'il faut donner la CLÉ et la répartition des points pour chaque item, y compris pour la réponse correcte ;
- ✓ pour les tâches de production écrite, s'assurer que les candidats vont pouvoir accomplir la tâche en tenant compte du lexique et de la capacité langagière de leur niveau, en prévoyant des réponses simples ;
- ✓ indiquer le format standard de la rédaction de la tâche.

Précisions sur la présentation attendue des items

- ✓ Le document électronique est ce qui convient le mieux car il peut être facilement stocké et le rédacteur peut travailler à partir d'un modèle qui assurera une présentation cohérente ;
- ✓ si un examen complet est requis, indiquer si les items doivent être numérotés en continu et si les parties se suivent ou si chaque partie ou chaque exercice doit être présenté sur une nouvelle feuille ;
- ✓ penser à la façon d'identifier le rédacteur d'items, la date et l'intitulé de l'examen.

(Toutes ces précisions peuvent être indiquées dans le guide du rédacteur d'items – voir ci-dessous)

Précisions sur les échéances

Il est important que les rédacteurs sachent quand leur production va être mise en forme et si on attend d'eux qu'ils y participent. Si les rédacteurs ne sont pas impliqués dans la suite du processus de production, il faut leur indiquer comment leur travail s'intègre au calendrier général de production afin qu'ils comprennent l'importance des échéances qu'on leur demande de respecter.

Précisions supplémentaires, telles que les conditions d'emploi

Il faut préciser aux rédacteurs le type de contrat de travail auquel ils seront soumis, soit parce que le travail demandé vient en supplément de ce qu'ils font dans leur établissement ou leur entreprise, soit parce qu'ils sont travailleurs indépendants. On peut ne rémunérer que le matériel accepté (ne pas payer le matériel rejeté) ou ne payer qu'une partie à la remise du matériel et régler le complément correspondant au matériel accepté. On peut aussi avoir des tarifs différents selon le type d'items ou donner une somme correspondant à une partie de l'examen ou à l'examen complet.

Les professeurs d'un établissement scolaire auxquels on aura demandé de rédiger des items devront disposer d'assez de temps dans le cadre de leur emploi du temps.

Les documents suivants sont à mettre à la disposition des rédacteurs :

- Des spécifications détaillées à l'intention des rédacteurs. Ces spécifications dont il faut souligner le caractère confidentiel, décrivent de façon plus détaillée que les spécifications destinées au grand public, les conditions de sélection et de présentation du matériel. Ces indications permettent de gagner du temps et d'éviter tout malentendu sur ce que des rédacteurs peuvent considérer comme étant acceptable.
- Des échantillons de matériel ou d'épreuves déjà passées.
  Il est également important que les rédacteurs aient des indications sur la population à laquelle le test est destiné : l'âge, le sexe, le contexte linguistique (L1, Niveau d'études) des candidats.

Enfin, selon la situation, d'autres documents et consignes peuvent être donnés :

- un formulaire d'acceptation de la commande signé par le rédacteur ;
- un accord écrit selon lequel l'organisme certificateur a les droits de copyright du matériel ;
- une liste ou un glossaire définissant l'étendue et le niveau du lexique et des structures à utiliser ;
- un livret d'informations sur l'organisme certificateur.

3.4 Le contrôle qualité

3.4.1 La vérification des nouveaux items

La qualité du matériel qui a été remis par les rédacteurs doit être vérifiée par des experts et les items doivent être expérimentés. Si des items doivent être changés, une nouvelle vérification doit avoir lieu.

L'idéal est que la vérification, qui est essentielle, ne soit pas menée par la personne qui a rédigé les items. Les items et les tâches peuvent par exemple être revus par des collègues. Dans le cas où un rédacteur travaille seul, le fait de laisser un certain temps entre la production et la révision, ainsi que la révision en une seule fois d'un ensemble peut renforcer l'objectivité.

La première démarche consiste à vérifier si le matériel est conforme aux spécifications et aux exigences formulées à la commande. Il faut bien sûr transmettre aux rédacteurs les conclusions de cette vérification afin qu'ils revoient leur travail et se perfectionnent. Celles-ci peuvent inclure des suggestions sur la façon de changer un item (cf. annexe V).

La commande peut ne concerner tout d'abord que des textes sans items, le rédacteur ne produisant les items qu'après acceptation du texte. Cette première vérification ainsi que la révision d'un petit nombre d'items peut se faire rapidement. Il faudra prévoir une réunion spécifique si les items à vérifier sont nombreux.

A l'étape de la mise en forme, chaque item et chaque tâche sont revus de façon plus détaillée et il est important que cette révision ne soit pas faite par la personne qui a produit le matériel. Les professeurs d'un établissement peuvent par exemple échanger leur production avec leurs collègues pour vérification.

Le nombre de personnes faisant partie du groupe chargé de la mise en forme est un élément important : plus de quatre ralentit le travail et moins de trois rend la diversité des points de vue insuffisante. Si plusieurs réunions sont envisagées, il est souhaitable de désigner un coordinateur qui organisera les réunions en termes de dates, de personnes et de travail à effectuer.

On peut gagner du temps dans les réunions en donnant le matériel à l'avance à chaque membre qui travaillera de la façon suivante :

- des items s'appuyant sur des textes doivent être lus avant le texte. On peut ainsi repérer ceux auxquels on peut répondre sans se référer au texte (c'est à dire par bon sens ou grâce à la culture générale) ;
- on répondra aux autres items sans regarder la réponse, comme si on passait le test. Cela permettra d'identifier les items pour lesquels plus d'une réponse correcte est possible, ceux qui sont mal formulés, les distracteurs improbables ou les items qui sont difficiles ;

- on vérifiera si la longueur ou la durée, le sujet, le style et le niveau de langue des textes de compréhension écrite et orale conviennent. Il est nécessaire de faire appel à un expert ou éventuellement à des référentiels pour la vérification du niveau de langue.

Si la vérification se fait en groupe, tout problème relevé dans le matériel sera discuté en détail par le groupe. Cela donne souvent lieu à de longues discussions sur le matériel et les rédacteurs doivent être, ce qui n'est pas toujours facile, capables d'accepter les critiques constructives et d'en formuler. Lorsqu'un rédacteur se sent obligé de justifier et d'expliquer certaines de ses propositions à des collègues expérimentés, c'est qu'elles ont des faiblesses.

Le groupe désigne un rapporteur qui recueillera de façon précise et détaillée toutes les décisions qui auront été prises et rendra clairement compte de toute modification. Il est essentiel qu'à la fin de la réunion tous soient d'accord et qu'il n'y ait aucun doute sur les modifications décidées.

C'est à l'organisme certificateur de prendre les décisions finales et de clore les discussions.

Les points à revoir de façon détaillée sont les suivants:

- l'attention donnée aux consignes et à la clé ;
- la surveillance des biais lors de la rédaction d'items en se référant à une liste de sujets à éviter et en prenant les précautions nécessaires (cf. annexe VII) ;
- il se peut que certaines propositions soient potentiellement intéressantes mais les modifications ne peuvent être faites pendant la réunion. On rendra les items à leur rédacteur qui apportera les changements nécessaires ou on les confiera à un rédacteur plus expérimenté pour qu'il apporte les corrections qui s'imposent ;
- pour des raisons de sécurité, après la réunion, on détruira toutes les copies d'exemplaires supplémentaires du matériel préparé et les copies de travail. L'organisme certificateur garde les exemplaires révisés du matériel accepté.
- Les rédacteurs sont en droit d'attendre de l'organisme certificateur une explication sur le matériel refusé, surtout s'ils n'ont pas participé à la révision ou étaient absents lors du traitement de leur propre matériel ;
- les réunions de vérification offrent aux nouveaux rédacteurs une très bonne occasion d'apprendre à travailler en groupe avec des rédacteurs plus expérimentés.

3.4.2 Pilotage/test pilote, pré-test et expérimentation

Il est nécessaire de tester le matériel élaboré car les réponses des candidats peuvent être inattendues.
Cela peut se faire soit sous la forme de la passation d'un test pilote, d'un prétest et d'une expérimentation ou en combinant les trois formes en fonction des objectifs et des moyens dont on dispose.

Le test pilote peut être passé de façon informelle par un nombre restreint de personnes qui peuvent par exemple être des collègues. Leurs réponses sont analysées et leurs remarques prises en compte (cf. annexe VI) en vue de modifications éventuelles.

Le prétest concerne avant tout les items qui donnent lieu à une évaluation objective (réception orale et écrite).
Les conditions de passation du prétest sont les mêmes que celles de l'épreuve finale : les candidats sont identiques à la population attendue et le nombre de réponses sera suffisamment important pour mener à bien des analyses statistiques. (cf. annexe VII). Ces analyses montrent comment les choix ont fonctionné, la difficulté d'un item, la moyenne des résultats, si le test correspondait au niveau des candidats, les erreurs, les biais éventuels (cf. annexe VII), leur adéquation générale au concept, etc. Il existe, pour les analyses statistiques, du matériel peu sophistiqué et de moindre coût qui donne des informations très utiles.

Les tâches qui donnent lieu à une évaluation subjective (production orale et écrite) peuvent également donner lieu à des analyses statistiques, mais des analyses qualitatives d'un nombre réduit de réponses peuvent être d'une plus grande utilité. On donne parfois le nom *d'expérimentation* à cette forme de prétest à petite échelle pour la distinguer de celle du prétest composé d'items de type objectif.

L'expérimentation permet de savoir si les tâches fonctionnent et indiquent de façon explicite la performance attendue.
A la différence du test pilote, l'organisation du prétest est la même que celle de l'épreuve finale dans la mesure où il est nécessaire d'avoir :

- un grand nombre de candidats (cf. annexe VII) ;
- des épreuves révisées et sécurisées ;

- des lieux de passation et du personnel ;
- des correcteurs.

La population passant le prétest doit réellement avoir des caractéristiques identiques à celle qui va passer l'épreuve finale (âge, sexe…). L'idéal est de faire appel à des apprenants se préparant à passer un examen.

Pour les motiver à participer et donner des réponses qui correspondent vraiment à leur compétence, on leur proposera un retour d'informations sur leur performance. Ces informations leur permettront ainsi qu'à leur professeur d'avoir une idée du niveau atteint et de prendre conscience des domaines dans lesquelles ils doivent s'améliorer avant la passation de l'épreuve finale.

Le principal inconvénient de ce dispositif est le risque que l'on fait courir à l'épreuve finale en termes de sécurité. C'est parfois la raison invoquée par certains organismes certificateurs pour ne pas faire de prétest.

Pour réduire les risques, on recommande de ne pas présenter les épreuves sous leur forme définitive. Il faut par ailleurs prévoir un laps de temps assez grand entre l'utilisation d'un item dans un pré-test et l'utilisation du même item dans l'épreuve finale. Dans le cas de prétests organisés ailleurs que dans l'organisme producteur du test, il faut donner au personnel qui va s'en occuper des consignes impératives de sécurité et faire signer des engagements de confidentialité.

Il n'est pas nécessaire que les épreuves du prétest soit exactement identiques aux épreuves de l'examen final dans la mesure où ce sont les items et non pas le test lui-même qui sont prétestés. Il faut quand même savoir que la motivation de ceux qui vont passer le prétest sera d'autant plus grande qu'ils sauront que le format du prétest est très proche de l'épreuve finale. Il est donc recommandé de proposer un format quasi identique.

Quoi qu'il en soit, les conditions d'organisation de la passation du prétest doivent être les mêmes que celles de l'épreuve finale. Pour que les interprétations des données ne soient pas faussées, il est indispensable que les candidats au prétest puissent se concentrer, ne trichent pas et que la durée du test soit la même pour tous.

Quand la qualité des données statistiques est de première importance (par exemple en cas de calibrage des items, cf. annexe VII), il faut faire passer le prétest par un nombre important de candidats. Le nombre requis dépend des analyses à effectuer. On peut malgré tout détecter des problèmes que certains items peuvent poser avec un nombre réduit de candidats (moins de 50). Avec des effectifs encore plus réduits, il vaut mieux faire des analyses qualitatives.

Il est également essentiel de faire appel à des candidats dont les caractéristiques sont aussi proches que possibles que celles des candidats au test final. Avec un échantillon plus petit et moins représentatif, des conclusions hasardeuses seront tirées des analyses et celles-ci devront être rééquilibrées par le jugement d'experts lors de la révision des items. Voir l'annexe VII pour plus de renseignements sur les analyses.

Si la passation d'un prétest a pour but de recueillir des renseignements de type qualitatif sur les items, il faut tenir compte des éléments suivants pour optimiser cette opération :

- pour des items dont l'évaluation/correction est objective, il est possible de recueillir les renseignements fournis par les candidats et les professeurs après la passation. On peut utiliser une liste de questions ou un questionnaire à cet effet (cf. annexe VI) ;
- dans le cas de tâches de production orale faisant intervenir un interlocuteur, l'information donnée par ce dernier peut être d'une grande utilité. L'organisme certificateur saura si l'étudiant a compris la tâche, si elle est adaptée à son expérience et à son âge et si les informations données ont été suffisantes pour lui permettre de la réaliser correctement (cf. annexe VI) ;
- dans le cas d'items et de tâches dont l'évaluation est subjective, les réponses des candidats montrent à quel point on leur a donné l'occasion de s'exprimer et de montrer l'étendue des structures syntaxiques et du lexique attendue au niveau du test ;
- on peut également recueillir des informations sur leur expérience en tant que candidat à un prétest ainsi que d'autres informations concernant la session elle-même.

3.4.3 La révision des items

Il faut prévoir une réunion de révision des items après les phases de pilotage et de prétest. Participent à cette réunion l'organisme certificateur, des rédacteurs expérimentés et, pour les items et les tâches de production, un examinateur expérimenté.

Le but de cette réunion est de garder, améliorer ou rejeter les items en fonction des données du pilotage et du prétest. La figure 8 montre à quel moment les items qui doivent être améliorés repassent le test pilote et le prétest

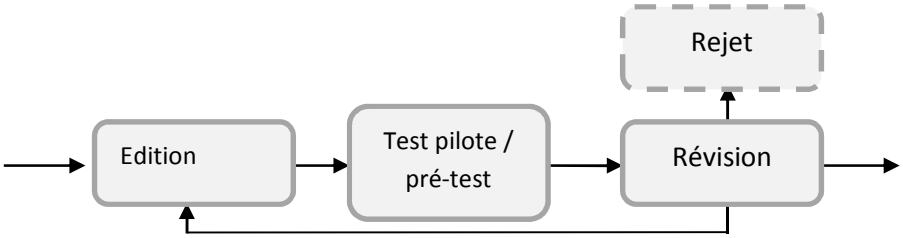


Figure 8 ***Amélioration des items selon la procédure de l'assurance qualité***

La révision du prétest traite des points suivants :

- Quels items et tâches sont prêts à être utilisés tels quels dans l'épreuve finale ?
- Quels items et tâches doivent être rejetés car ne convenant pas ?
- Quels items et tâches peuvent être réécrits et prétestés à nouveau avant de les inclure dans l'épreuve finale ?

La réunion de révision doit envisager de répondre aux questions suivantes :

- dans quelle mesure les réponses des candidats au prétest correspondaient à celles de la population cible ? L'adéquation permettra d'évaluer le degré de fiabilité des données des analyses ;
- dans quelle mesure les tâches et les thèmes étaient intéressants et étaient à la portée des candidats ? Les procédures ont-elles bien fonctionné ?
- en ce qui concerne les items et les tâches individuels, il est utile, pour évaluer ceux qui donnent lieu à une correction subjective, d'étudier un certain nombre de réponses de candidats. Dans le cas d'items à évaluation objective, on pourra déceler des problèmes grâce aux analyses statistiques, qu'une révision par des experts pourra confirmer et corriger. Prudence, en revanche, si les données servant aux analyses sont insuffisantes (par exemple avec un petit nombre de candidats ou des candidats qui ne conviennent pas). On peut également donner une certaine importance à l'appréciation qualitative des items et des tâches ;
- il faut avoir une approche cohérente et assurer un suivi des données concernant les tâches qui ont posé problème lors des analyses statistiques et qui se trouvent dans une banque d'items. On en verra l'utilité lors de la fabrication du test ou de l'examen. Voir l'annexe VII pour plus d'information sur les analyses statistiques.

3.5 La constitution du test ou de l'examen

Une fois le matériel disponible, les tests ou les examens peuvent être constitués. L'objectif de cette étape est de produire un format de test qui réponde aux normes de qualité et corresponde aux spécifications requises.

L'étape de fabrication doit prendre en compte un certain nombre d'éléments tels que le contenu du test ou de l'examen et la difficulté de l'item afin que de répondre aux exigences des spécifications.

Certaines caractéristiques du test ou de l'examen peuvent être fixées à partir des spécifications ou du format (par exemple le nombre et le type d'items/de tâches à inclure), d'autres peuvent rester plus souples (par exemple les thèmes, les accents différents, etc.). Des directives pourront être données pour arriver à un équilibre entre les caractéristiques suivantes:

- le niveau de difficulté. Il peut être décidé soit en faisant appel à un jugement subjectif soit en se référant à la difficulté moyenne des items du test et à l'étendue de difficulté couverte (cf. annexe VII) ;
- le contenu (thème) ;
- l'étendue (la représentativité des tâches par rapport au concept) ;
- la graduation (à savoir s'il y a une graduation de la difficulté dans le test).

Ces directives devraient concerner le test ou l'examen dans son ensemble, ainsi que ses différentes composantes, à fin de comparaison.

D'autres considérations pour certains types de tests ou d'examens sont à prendre en compte. Par exemple dans une épreuve de compréhension écrite comprenant plusieurs textes et items, il faut s'assurer que les thèmes ne sont pas répétés, que le

nombre de mots n'est pas trop élevé. De la même façon, dans une épreuve de compréhension orale, il faut assurer l'équilibre entre les voix féminines et masculines, les accents régionaux (si cela est pertinent).

3.6 Questions clés

- Comment le processus de production du matériel va-t-il être organisé ?
- Peut-on disposer d'une banque d'items ?
- Qui va rédiger les épreuves?
- Quelles doivent être les compétences professionnelles des rédacteurs d'items ?
- Quelle formation doit être donnée ?
- Qui va faire partie des réunions de vérification?
- Comment les réunions de vérification seront-elles dirigées ?
- Est-il possible de prétester ou d'expérimenter le matériel ?
- Quelles peuvent être les conséquences si le matériel ne peut être ni prétesté ni expérimenté et quelle solution peut-on trouver ?
- Quel type d'analyse doit être faîte des données sur les performances recueillies grâce au prétest ?
- Comment les analyses seront-elles analysées ? (par exemple en vue de l'élaboration de l'épreuve finale, pour la formation des rédacteurs d'items, etc.)
- Qui va participer à l'élaboration du test ou de l'examen ?
- Quelles sont les variables dont il faut tenir compte et quel poids doit-on leur donner ? (par exemple le niveau de difficulté, le contenu thématique, l'éventail du type d'items, etc.)
- Quel sera le rôle des analyses statistiques ? (par exemple en établissant une difficulté moyenne et l'étendue de difficulté).
- Quelle sera le poids des analyses statistiques par rapport aux informations venant d'autres sources dans la prise de décision ?
- Les éléments du test ou de l'examen une fois assemblés seront-ils contrôlés de façon indépendante ?
- Comment les présentations des différentes parties vont-elles s'inscrire dans la présentation générale du test ou de l'examen et comment cette présentation va-t-elle être reprise dans une série de tests ou d'examens?

3.7 Lectures complémentaires

Pour le manuel du rédacteur d'items, voir ALTE (2005)

Pour l'analyse des tâches, voir ALTE (2004a,b,c,d,e,f,g,h,i,j,k).

Des référentiels de certaines langues en relation avec le CECR sont disponibles : référentiels de descripteurs de niveaux, (Beacco et Porquier 2007,Beacco, Bouquet et Porquier 2004 ; Glaboniat, Müller, Rusch, Schmitz et Wertenschlag 2005 ; Instituto Cervantes 2007 ; Spinelli et Parizzi 2010 ; www.englishprofile.org). Niveau Seuil (Van Ek et Trim 1991), Niveau indépendant, (Van Ek et Trim 1990) et Maîtrise, (Van Ek et Trim 2001) sont des ouvrages antérieurs aux référentiels.

L'annexe VII comprend un complément d'informations sur la façon d'utiliser les données statistiques.

4. La délivrance des examens

4.1. Les objectifs de la délivrance des examens

L'objectif principal du processus de délivrance des examens est de recueillir des renseignements précis et fiable sur les compétences de chaque candidat.

Les plus grands défis auxquels la délivrance des examens doit faire face sont d'ordre logistique. Il ne s'agit pas à ce niveau d'améliorer la qualité du matériel comme précédemment. Les organismes certificateurs doivent s'assurer que :

- la performance du candidat dépend avant tout de ses compétences langagières et le moins possible de facteurs extérieurs tels que le bruit ou la triche ;
- les réponses et les corrections du candidat sont recueillies de façon efficace et sûre en vue de la correction et de la notation ;
- tout le matériel lié à l'examen soit livré au bon endroit et à temps.

Toutes ces tâches sont importantes, que l'examen soit organisé sur une grande échelle ou localement. Le moindre détail, tel que l'aménagement de la salle d'examen, a son importance.

Le recueil de plus amples informations sur le profil des candidats peut constituer un objectif supplémentaire. Ces informations sont d'autant plus utiles si l'organisme certificateur ne les connaît pas. Une bonne connaissance des candidats est un élément tangible de la validité (cf. annexes I et VII).

4.2. Le processus de délivrance des examens

La figure 9 montre le processus de délivrance de l'examen. Il se peut que, dans un contexte tel que la passation de l'examen dans une salle de classe, certaines étapes telles que l'inscription des candidats ou la délivrance du matériel ne posent pas de problème. Il n'en reste pas moins qu'il faut être attentif aux conditions de passation telles que la taille de la salle, le bruit environnant. Dans d'autres contextes, il faut résoudre des problèmes de logistique beaucoup plus importants.

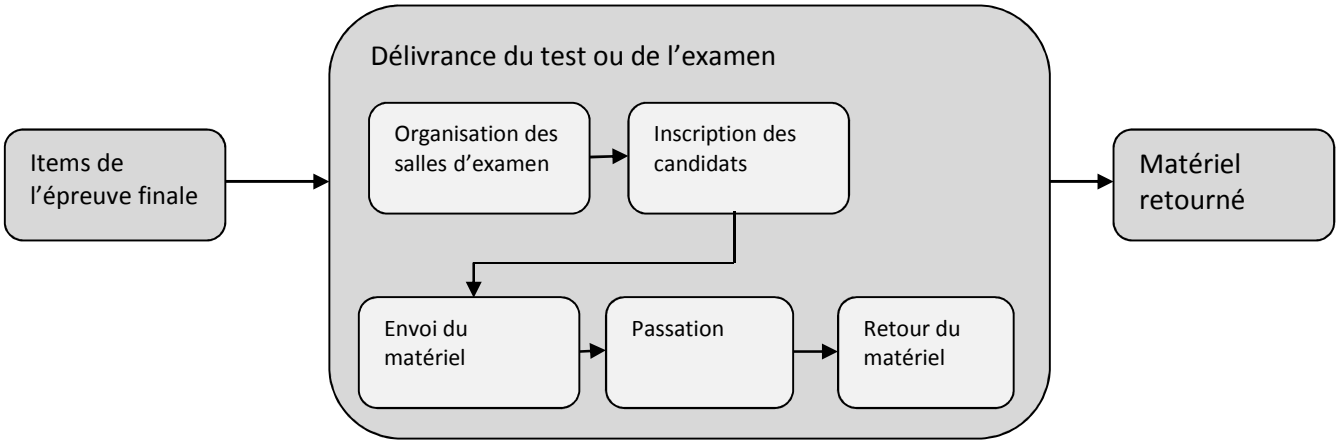


Figure 9 ***Processus de délivrance du test ou de l'examen***

4.2.1. Organisation des salles d'examens

Les salles où se déroulera l'examen doivent être inspectées avant la passation. Cette inspection peut être faite soit par l'organisme certificateur soit par une personne de l'école où l'examen est organisé et en laquelle l'organisme a confiance. Dans le cas où la passation a lieu dans d'autres centres, ceux-ci doivent avoir été agréés. Les critères d'agrément sont les suivants :

- Espace suffisant pour accueillir le nombre prévu de candidats
- Accès aux locaux
- Sécurité des conditions de stockage
- Adoption sans réserve des règles imposées par l'organisme certificateur
- Formation du personnel aux procédures de l'organisme certificateur

Quand c'est une tierce personne qui a trouvé les centres de passation, l'organisme certificateur doit mettre en place un dispositif d'inspections aléatoires afin de vérifier les conditions de passation faite en son nom.
Les critères servant aux inspections doivent toujours être les mêmes. Il est recommandé de vérifier les conditions matérielles avant chaque passation car les organisateurs de l'examen n'ont peut-être pas toujours été informés de travaux en cours dans le voisinage.

Les points à vérifier sont les suivants :

- Le bruit ambiant
- L'acoustique de la salle (particulièrement pour l'épreuve de compréhension orale)
- Les capacités d'accueil (permettant un espace entre les tables)
- La configuration de la salle (permettant aux surveillants de bien voir tous les candidats)
- L'accès à la salle
- La mise à disposition de toilettes et d'une salle d'accueil des candidats
- Des lieux de stockage du matériel avant et après la passation comportant la sécurité nécessaire

Les centres qui ne remplissent pas les conditions requises ou les organisations qui commettent des erreurs doivent être supprimés de la liste d'éventuels lieux de passation ou de collaborateurs.

4.2.2 L'inscription des candidats

Si l'examen a lieu à la fin d'un cours, il suffit alors d'avoir la liste des étudiants connus du professeur. Par contre, si l'organisme certificateur ne connaît pas les candidats ou si des candidats supplémentaires sont susceptibles de s'inscrire, il est alors indispensable de recueillir des informations sur eux. Un processus d'inscription fournissant les informations nécessaires pour la passation de l'examen et la remise des résultats doit être mis en place. Les candidats peuvent aussi demander à ce que les conditions d'examen soient adaptées à leurs capacités réduites :

- Personne sourde ou malentendante
- Personne aveugle ou mal voyante
- Personne dyslexique
- Personne à mobilité réduite

Il faut savoir évaluer les différents types de demandes avec précision pour prévoir l'assistance ou la compensation nécessaires. Il est donc recommandé de mettre en place des procédures pour les demandes les plus communes comprenant les preuves que le candidat doit fournir (par exemple la lettre d'un médecin), les dispositifs à mettre en place et la date de la demande.

Pour des besoins particuliers comme une mobilité très réduite supposant une aide pour que le candidat accède à sa place, cette aide devrait pouvoir se trouver sur place.

Il est parfois nécessaire de prendre d'autres mesures plus adaptées dans le cas par exemple de candidats ayant des difficultés à lire (dyslexiques ou mal voyants). Par contre, il faut faire attention à ne pas avantager certains candidats.

A ce stade, il est également possible de recueillir des informations sur le contexte des candidats. Des informations sur le profil des candidats peuvent permettre de tirer des conclusions importantes en termes de comparabilité des groupes qui se présentent à l'examen. Ces informations concernent :

- Le niveau des études
- La première langue apprise
- Le sexe
- L'âge
- L'expérience d'apprentissage de la langue cible.

Il est indispensable que les candidats sachent pourquoi ces informations sont demandées, de même qu'il faut que ces données soient gardées en lieu sûr et restent confidentielles afin d'assurer tous les droits à la vie privée des candidats. L'inscription est également l'occasion de fournir des informations aux candidats telles que les conditions d'inscription, les règles à respecter lors de la passation, les possibilités de faire appel et les moyens mis à leur disposition pour une assistance particulière. Il faut bien sûr donner aux candidats toutes les informations nécessaires en particulier celles sur les lieux, les jours et heures de passation. Pour une bonne diffusion, ces informations peuvent être imprimées et distribuées, disponibles sur internet ou par courrier électronique.

L'inscription peut être faite directement par l'organisme certificateur, les centres de passation ou des institutions indépendantes telles que le ministère de l'Education. Dans la mesure du possible, l'organisme certificateur doit s'assurer que les modalités d'inscription sont identiques pour tous les candidats.

4.2.3 L'envoi du matériel

Il est parfois nécessaire d'envoyer le matériel aux centres de passation. Le dispositif de transport doit être sécurisé et prévoir un suivi du matériel pour être sûr qu'il arrive à destination à temps.

Il est préférable d'envoyer le matériel bien avant la date prévue de l'examen pour éviter tout retard et avoir éventuellement le temps d'envoyer des pièces manquantes. Il faut de toute façon s'assurer qu'une fois sur place le matériel est en lieu sûr pour toute la durée de l'opération.

Les responsables de la passation doivent vérifier le contenu de l'envoi en comparant avec une liste du matériel. En cas de matériel manquant ou endommagé, les responsables suivent alors les procédures mises en place et demandent les pièces à ajouter ou remplacer.

4.2.4 La passation de l'examen

Les centres d'examens doivent prévoir un nombre suffisant de personnel : surveillants, correcteurs, autres. Ces personnes doivent connaître leurs responsabilités et quand beaucoup de monde est impliqué, un emploi du temps doit être établi. Les directives de passation doivent comprendre des instructions pour le contrôle des documents des candidats et l'admission des retardataires.

Avant le début de l'examen, il faut donner des instructions précises aux candidats sur la conduite à tenir pendant l'examen : précisions sur le matériel non autorisé, l'utilisation des portables, les conditions pour quitter la salle, le début et la fin de la passation. Il faut également avertir des conséquences en cas de bavardage ou de copiage.

Pendant la passation, il faut que les surveillants sachent comment réagir en cas de non-respect du règlement ou d'événements prévisibles ou non, par exemple s'ils voient un candidat tricher, s'il y a une panne d'électricité ou quelque autre événement provoquant un biais ou une injustice impliquant l'arrêt de la passation. En ce qui concerne la tricherie, il faut que les surveillants connaissent les moyens actuels tels que les enregistreurs digitaux, le MP3, les stylos qui scannent et les portables avec une caméra incluse.

Dans le cas d'événements non prévisibles, les surveillants doivent évaluer le degré de gravité et prendre les initiatives appropriées puis rédiger un rapport indiquant tous les détails de l'incident comme le nombre de candidats concernés, l'heure et une description de l'incident. On peut également mettre à disposition des surveillants un numéro de téléphone d'urgence.

4.2.5 Le retour du matériel

Le matériel servant à la passation est soit retourné à l'organisme certificateur, soit détruit. En cas de retour dès la fin de la session, les centres joignent à l'envoi les feuilles de présence et les plans de disposition des tables dans la salle. L'envoi du matériel doit être sécurisé, en utilisant le même mode de transport que pour l'envoi. La société qui s'occupe de l'envoi doit pouvoir assurer sa traçabilité en cas de retard ou de perte.

4.3 Questions clés

- Quelles sont les ressources disponibles pour la passation de l'examen ? (personnel administratif, surveillants, salles, lecteurs de CD, etc.)
- Comment former l'équipe ?
- Comment s'assurer de la conformité des salles et du fonctionnement des lecteurs CD avant le jour de l'examen ?
- Quelle est la fréquence des sessions ?
- Combien de candidats sont attendus ?
- Comment va se dérouler l'inscription des candidats et l'enregistrement de leur présence ?
- Combien de lieux de passation sont utilisés et s'il y a plus d'une salle, sont-elles regroupées ou dispersées?
- Comment acheminer le matériel dans les salles et le récupérer ?
- Quels sont les endroits sécurisés où le matériel peut être stocké ?
- Quel dysfonctionnement peut se produire et quelles sont les procédures et le règlement pour y répondre ?

4.4 Lecture complémentaire

Voir ALTE (2006b) pour une liste de contrôle d'auto évaluation pour la logistique et l'administration.

5 Correction, notation et délivrance des résultats

Le but de la correction est d'évaluer la performance de tous les candidats et d'assurer à chacun une correction juste et fiable. La notation, elle, vise à placer chaque candidat dans une catégorie significative de façon à ce que son score soit aisément compréhensible. Une catégorie significative peut, par exemple, être le niveau A2 ou C1 du CECR. Délivrer les résultats, signifie fournir au candidat et aux parties concernées les résultats au test ainsi que toute information nécessaire pour utiliser ces résultats correctement. Cela peut aller jusqu'à la décision d'engager ou non le candidat à l'issue d'un entretien de recrutement. La figure 10 propose une vue d'ensemble du processus général. L'évaluation de la performance du candidat peut cependant parfois avoir lieu en même temps que l'examen. C'est le cas pour la production orale qui est parfois évaluée ainsi, bien que la note puisse être ajustée par l'organisme certificateur avant la délivrance des résultats.

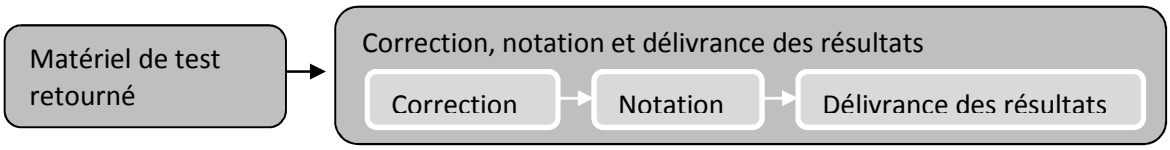


Figure 10. ***Le processus de correction, notation et délivrance des résultats***

Etapes préliminaires
Les étapes suivantes sont impératives avant d'entreprendre la correction et la notation :

- ✓ Définir l'approche choisie pour la correction
- ✓ Recruter les CORRECTEURS et les EVALUATEURS
- ✓ Former les correcteurs et les évaluateurs

5.1 La correction

Le terme *correction* couvre toutes les activités qui permettent d'attribuer une note aux réponses données à un test ou à un examen. On fait souvent une différence entre le correcteur et l'évaluateur, le premier étant moins qualifié que le second, qui a lui, bénéficié d'une formation professionnelle. Cette distinction est faite dans cet ouvrage. Cette section couvre la correction administrative (c'est-à-dire humaine) ainsi que les machines à corriger.

5.1.1 La correction humaine

Il n'est nul besoin que les CORRECTEURS soient des experts en évaluation par les tests –il suffit qu'ils aient un excellent niveau de compétence dans la langue évaluée. Pour mener à bien leur travail, les correcteurs ont cependant besoin de formation et de conseils ainsi que clés de réponses univoques. Si la correction est effectuée par un petit groupe de collègues, ils peuvent vérifier la qualité du travail des uns et des autres.

Le processus de correction doit être géré de façon à ce que les procédures respectent la planification prévue et que les résultats soient prêts à temps. La charge de travail de chacun des correcteurs ne doit pas être trop élevée, sous peine de mettre en péril la fiabilité ou l'exactitude des corrections.

Le recrutement et la formation des correcteurs

Dans sa forme la plus simple, l'acte de corriger implique que le correcteur associe la réponse du candidat à une question du test, à une ou plusieurs séries de réponses. Les questions à choix multiple (QCM) en sont le plus clair exemple, puisqu'aucune modification des choix donnés n'est possible. Lorsqu'il s'agit de ce type de correction, les correcteurs doivent simplement avoir une excellente connaissance du langage concerné, être attentifs aux détails et être prêts à accomplir des tâches répétitives. Aucune autre compétence particulière n'est requise. Dans ce cas, la formation consiste en une familiarisation avec les procédures à suivre. Avec une technologie appropriée, ce type de correction peut s'effectuer aussi bien, voire mieux, à l'aide d'une machine.

Dans le cas où la correction nécessite autre chose qu'un simple appariement entre questions et réponses, le correcteur peut avoir besoin de connaissances sur la langue, sur la langue des apprenants et sur la construction du test. Selon le degré de réussite, par exemple, aux QUESTIONS A CREDIT PARTIEL on peut leur attribuer une note choisie dans une série de note. Une

note peut par exemple être attribuée si le choix d'un verbe s'est révélé exact et une autre note si la forme correcte a été utilisée. Le correcteur doit avoir un niveau d'expertise adéquat afin de pouvoir reconnaître une réponse incorrecte.

Pour des questions de ce type, il peut être difficile de s'assurer que la clé est suffisamment exhaustive. C'est pourquoi il est utile que le correcteur puisse identifier et relever les différentes réponses rencontrées.

Lorsque les correcteurs sont recrutés de façon temporaire mais régulière, il est utile de les évaluer selon un certain nombre de paramètres tels que la justesse, la fiabilité et la rapidité de correction. Les correcteurs ne donnant pas satisfaction peuvent alors être soit remerciés, soit formés à nouveau. Un tel système peut faire partie de la formation, comme le montre la figure 11. Les correcteurs qui sont fréquemment appelés à corriger peuvent être dispensés de certaines sessions de formation. L'estimation de leur performance (cf. 5.1.3 Surveillance et contrôle de qualité) rendra plus facile la décision de renvoi à une session complète de formation, ou à une formation complémentaire, ou à un remerciement.

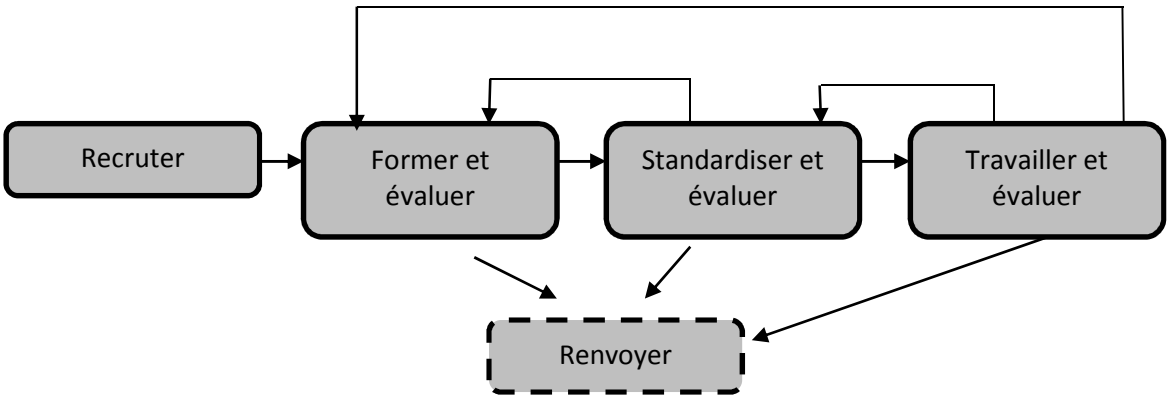


Figure 11 ***Recrutement, formation et évaluation des correcteurs et évaluateurs***

Conseils pour évaluer les réponses

Une clé de réponse formalisée est la meilleure façon d'enregistrer la réponse correcte et de la communiquer aux correcteurs. Les clés sont conçues en même temps que les questions et suivent les mêmes procédures de rédaction. La clé doit prendre en compte toutes les réponses acceptables de manière globale et être totalement univoque.

La figure 12 montre un exemple de question où l'on demande au candidat de compléter un blanc en utilisant le mot donné (« quelle ») . La clé donne quatre choix possibles pour un élément (1 point) et une possibilité pour le second (1 point). Le nombre total de points accordés à cette question est donc de 2.

Le travail des correcteurs ne peut être que plus efficace, juste et fiable si la clé est présentée de façon claire.

Le magasin fermera, que vous le vouliez ou non.

quelle

Le magasin va …… votre opinion.

Clé :

(certainement / sûrement / obligatoirement) fermer <u>quelle que soit</u>

Figure 12 ***Exemple d'exercice à trou***

D'autres réponses peuvent être correctes mais ne sont pas données par la clé. C'est pourquoi les correcteurs doivent relever les réponses qu'ils pensent être correctes. Ces réponses doivent être examinées et si elles sont réellement correctes, les

points seront attribués aux candidats. Si les corrections sont effectuées par un petit groupe de correcteurs, les problèmes peuvent être aisément résolus en discutant régulièrement avec un concepteur. Dans quelques cas, si la clé est réévaluée ou modifiée, tout ou partie des copies devront être recorrigées.

Gérer le processus de la correction

Les corrections s'effectuent généralement pendant une période fixée, les résultats devant être délivrés aux candidats à des dates précises. Pour estimer le temps nécessaire il suffit de mettre en relation le nombre de candidats et le nombre de correcteurs disponibles. Il est prudent de légèrement surestimer le temps nécessaire ou bien d'engager plus de correcteurs afin de s'assurer que tous les problèmes pourront être réglés.

Si on a à faire à un grand nombre de candidats et de correcteurs, on doit mettre en place un système de traçage des copies tout au long du processus. Un système simple consiste à noter le nombre de copies et le numéro du correcteur, ainsi que la date de remise des copies et la date de correction. L'organisme certificateur peut ainsi estimer le temps et le nombre de correcteurs requis pour un nombre donné de candidats.

Le système de traçage donne également des informations importantes sur la performance de chaque correcteur, comme par exemple le temps moyen dont ils ont besoin pour corriger une copie. Si on s'attache à vérifier le travail du correcteur on peut également compter le nombre moyen d'erreurs faites. Ces statistiques peuvent être obtenues en vérifiant, pour chaque correcteur, un échantillon représentatif de son travail.

5.1.2 La correction par une machine à corriger

Les machines à corriger les copies utilisent généralement une LECTURE OPTIQUE / une technologie de reconnaissance optique de la correction (ROC). La ROC est très utile lorsqu'il s'agit de corriger un nombre élevé de copies qui ne requièrent aucune évaluation humaine (c'est le cas des QCM, des questions de type VRAI/FAUX ou d'appariement). Les candidats peuvent alors noter leurs réponses sur des feuilles adaptées, comme le montre la figure 13. Ces feuilles sont ensuite scannées, de façon à enregistrer les données et à les transmettre à un ordinateur. La technologie ROC peut également être utilisée pour des questions qui requièrent une correction humaine. Le correcteur note les réponses sur la feuille ROC qui est ensuite scannée.

Les scanners accélèrent le processus de correction et réduisent les erreurs humaines mais ce processus n'est pas infaillible : le scanner peut se tromper en lisant une case cochée, ou peut lire par erreur une case non voulue. Pour éviter de telles erreurs, des contrôles de données doivent être effectuées en cherchant dans toutes les feuilles ROC des réponses contraires aux directives du test, par exemple plusieurs cases cochées alors qu'un seul choix est demandé. On devra alors corriger les feuilles ROC à la main.

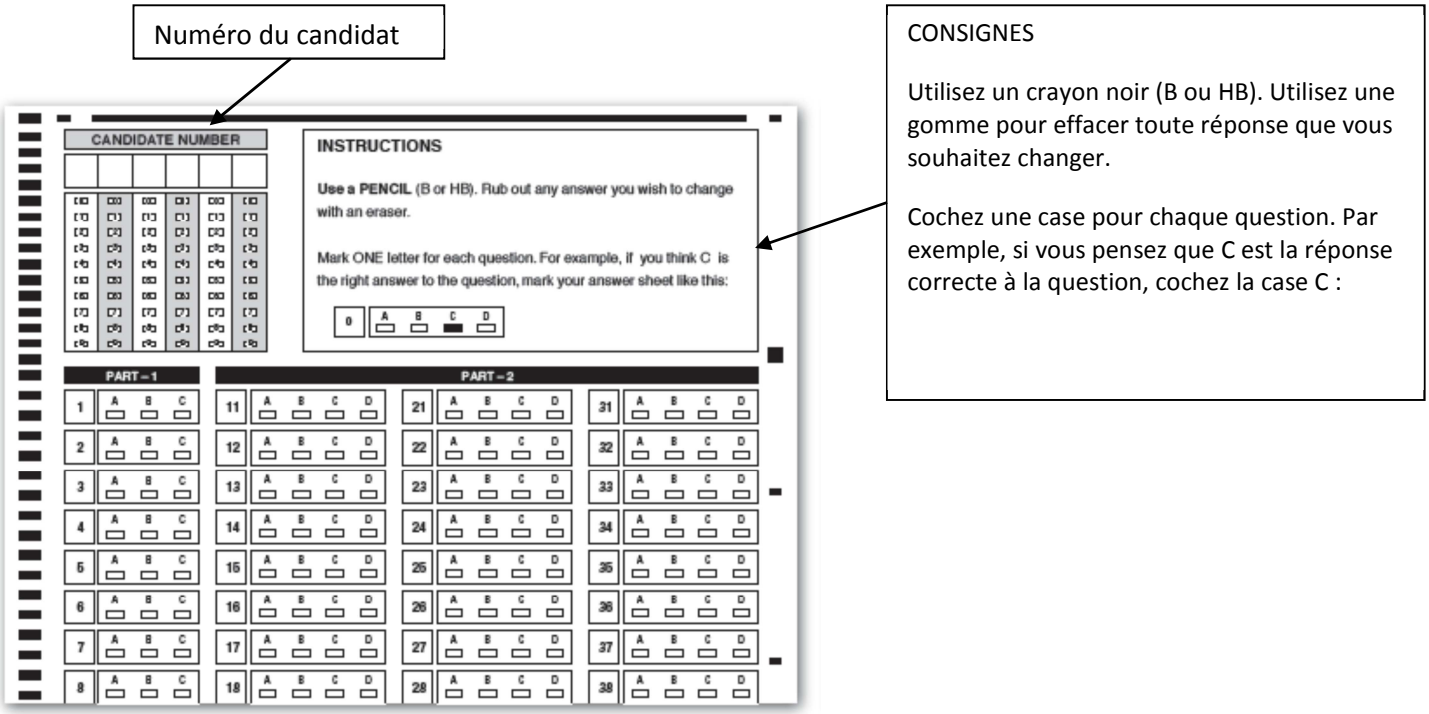


Figure 13 ***Partie d'une feuille ROC***

5.1.3. L'évaluation

On utilisera les termes *évaluation* et *évaluateur* lorsqu'un jugement d'expert intervient de façon bien plus importante que dans le type de correction décrit précédemment. Lorsque le jugement entre en jeu, c'est que le concepteur du test donne plus d'une seule « réponse correcte ». Il y a, dans ce cas de plus grandes possibilités de désaccord entre les jugements des évaluateurs que dans d'autres types de correction, laissant ainsi la place à un plus grand danger de divergence entre les évaluateurs ou dans le travail d'un évaluateur individuel. Pour assurer la justesse et la fiabilité, on devra combiner des sessions de formation, des conseils et des remarques correctives.

Beaucoup de ce qui a été dit de la correction humaine est également vrai dans l'évaluation : on doit gérer le processus afin d'utiliser les ressources de façon efficace, contrôler et surveiller afin d'assurer la justesse de l'évaluation. La fiabilité doit également être surveillée (cf. Section 1.3, Annexe VII).

Les échelles d'évaluation

La plupart des approches de la compétence évaluative sont liées à une échelle d'évaluation. Il s'agit d'une série de descripteurs des performances à différents niveaux, indiquant la note ou le classement que mérite chaque performance.

Les échelles d'évaluation limitent les variations inhérentes à la subjectivité des jugements humains. On prend généralement en compte les options suivantes :

- ✓ **Echelles holistiques ou analytiques** : on peut attribuer une note à une performance en utilisant une échelle qui décrit chaque niveau de performance à l'aide d'une série de caractéristiques. L'évaluateur choisit le niveau qui décrit le mieux les performances. De la même façon, des échelles peuvent être conçues pour toute une série de critères (par exemple effet communicatif, justesse, adéquation au contexte, etc.), et une note peut être attribuée à chacun de ces critères. Les deux approches peuvent relever du même concept de compétence langagière décrits en termes similaires – la différence réside dans le jugement que l'évaluateur est appelé à donner.

- ✓ **Echelles relatives ou absolues :** les termes utilisés dans les échelles peuvent être relatifs, liés à l'évaluation (par exemple « insuffisant », « adéquat », « bon ») ou peuvent tendre vers la définition du niveau de performance en termes positifs et précis. Pour interpréter la performance selon les échelles et les niveaux du CECR, cette dernière option est préférable, les échelles de descripteurs du CECR permettant de construire de telles échelles d'évaluation.

- ✓ **Echelles ou listes de contrôle** : une autre approche de l'évaluation à l'aide d'une échelle, qui peut être complémentaire, consiste à attribuer des notes à partir d'une liste de jugements oui/non si la performance correspond ou non à ce qui a été demandé.

- ✓ **Echelles généralistes ou sur tâches spécifiques** : Un examen peut utiliser soit une échelle dite généraliste ou un jeu d'échelles pour toutes les tâches, soit encore fournir des critères d'évaluation spécifiques à chaque tâche. Il est possible de combiner les deux. On peut, par exemple, fournir à la fois des critères spécifiques pour permettre l'évaluation (une liste des points qui doivent être traités), et des échelles généralistes.

- ✓ **Jugement comparatif ou absolu** : On peut définir une échelle à partir de performances modèles, de façon à ce que la tâche de l'évaluateur ne soit pas de donner le niveau indiscutable de la performance, mais d'indiquer simplement si cette performance est en-dessous, équivalente ou au-dessus d'une ou de plusieurs performances modèles. La note correspond alors à un classement sur une échelle. Pour le CECR, l'interprétation de ce classement dépend du jugement sur le niveau attribué aux modèles. Cette approche fonctionne à merveille si les modèles sont des tâches spécifiques.

Ces approches peuvent sembler grandement différentes, elles dépendent cependant toutes de principes sous-jacents semblables :

- ✓ Toute évaluation repose sur la compréhension que l'évaluateur a des niveaux.
- ✓ Les modèles sont essentiels pour définir et communiquer sur cette compréhension.
- ✓ Il est impératif que les tâches permettant de produire la performance évaluée soient liées aux échelles.

Il est classique de dire que les niveaux avaient une signification locale, correspondant au contexte d'un examen particulier et qu'il était donc difficile d'établir une comparaison avec les niveaux d'un autre examen pris dans un autre contexte. La création de cadres de compétences tels que le CECR a permis de comparer les niveaux de différents contextes. Cet état de fait a eu une incidence sur la façon dont les échelles d'évaluation sont articulées.

Lorsque le niveau était classiquement implicite et compris, les échelles étaient traduites en termes évaluatifs relatifs. Aujourd'hui, on a plus tendance à traduire les échelles en fonction du CECR et de son approche, qui est de décrire les niveaux de performance de façon identifiable, en termes positifs et précis. Les modèles (encore plus que le texte des descripteurs), restent essentiels pour définir et indiquer le niveau, et ils poussent les organismes certificateurs à être plus explicites sur ce que signifie atteindre un niveau.

Le CECR favorise la réflexion et le travail en termes de niveaux de compétence critériés. Deux éléments permettent de définir les niveaux : *ce que* les gens peuvent faire et *à quel degré* ils peuvent le faire. Dans un examen, le « ce que » est défini par les tâches spécifiées. « A quel degré » ces tâches sont réalisées, c'est ce que l'évaluateur doit juger.

C'est pourquoi l'approche classique de l'évaluation, qui consiste à appliquer des échelles d'évaluation, fonctionne relativement bien, à condition que les tâches soient bien choisies et que les jugements portent sur la réalisation des tâches. Les tâches servent alors grandement à définir les échelles, même si on s'y réfère de façon plus ou moins explicite dans la définition de ce que signifie une performance qui permet le « passage ».

Le CECR (p. 142) traite de certains aspects de l'évaluation subjective.

Le processus d'évaluation

Pour que le processus se déroule correctement les évaluateurs doivent avoir une compréhension identique des normes. Pour arriver à cette compréhension commune, il faut s'accorder sur des exemples de performance.

Dans le cas d'examens sur une petite échelle, un groupe d'évaluateurs peut arriver à un accord à la suite d'une discussion. Dans cette situation où les évaluateurs sont sur un pied d'égalité, la norme reconnue par tous risque de n'avoir qu'une portée locale et de ne pas être la même d'une session à l'autre. Dans le cas d'examens sur une grande échelle, la norme doit être stable et doit être significative. Pour y arriver, il faut s'appuyer sur la pratique d'examinateurs expérimentés qui, de par l'autorité qu'on leur reconnait, transmettent la norme aux nouveaux.

C'est ainsi qu'un petit groupe d'évaluateurs expérimentés va former le noyau qui assurera la continuité en termes de normes de la formation, du contrôle et de la correction des autres correcteurs.

Un tel système hiérarchique peut avoir différents niveaux comme le montre la figure 14. C'est une façon assez efficace d'assurer une formation en face à face ou le contrôle du travail des correcteurs. Mais les nouvelles technologies de l'information ainsi que l'évolution de la formation par internet réduisent les besoins d'une telle hiérarchie. Il faut aussi noter que la transmission précise de la norme est vraiment assurée grâce à des exemples de corrections établies de manière autoritaire pour chaque niveau.

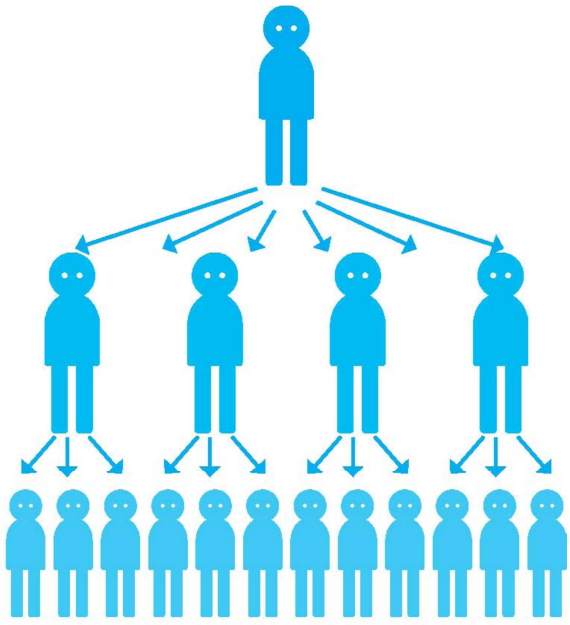


Figure 14 ***Maintien des normes grâce à un système de chef d'équipe***

La formation des évaluateurs

Le but de la formation est d'arriver à une correction constante et juste. On appelle standardisation le processus de formation des évaluateurs visant à appliquer la norme concernée. Si le CECR est pris comme référence pour fixer les normes, il faut alors que la formation commence par des exercices de familiarisation avec le CECR et utilise des échantillons de performance de production orale ou écrite se référant au CECR (Conseil de l'Europe 2009). Il peut aussi s'avérer nécessaire de former les évaluateurs en utilisant une échelle de classement avec laquelle ils sont familiarisés. La formation doit alors se faire par étapes en partant d'une discussion informelle pour arriver à une évaluation indépendante utilisant des échantillons en relation avec l'examen qui doit être corrigé :

- discussion guidée à partir d'un échantillon qui débouche sur la compréhension du niveau par les correcteurs ;
- correction indépendante d'un échantillon, suivie d'une comparaison avec la correction préétablie puis large discussion sur les raisons des éventuelles différences ;
- correction indépendante de plusieurs échantillons pour montrer combien les correcteurs sont proches de la correction préétablie.

L'idéal serait que les échantillons représentent des performances réalisées à partir des tâches de la session du test ou de l'examen en cours. Si ce n'est pas possible, on utilisera des tâches de sessions précédentes.

La surveillance et le contrôle qualité

L'idéal est qu'à l'issue de la formation, tous les correcteurs arrivent à une justesse et une constance telle qu'aucune correction ou retour d'information ne soit nécessaire. La phase de correction peut alors se dérouler sans problème. Mais il y a des cas où un contrôle s'impose pour identifier sans tarder les problèmes.

On peut identifier quatre types de problèmes, ce qu'on appelle aussi les « effets évaluateurs » :

1. **La sévérité ou le laxisme** : l'évaluateur sous-estime ou surestime le travail.
2. **L'utilisation de l'éventail de notes** : l'évaluateur utilise un éventail trop étroit, et de cette façon ne fait pas de distinction entre une performance faible et une bonne performance.
3. **L'effet de halo** : dans le cas d'attribution de plusieurs notes dans un examen, l'évaluateur se fait une impression du candidat à partir de la première note mise et l'applique aux notes qui suivent, indépendamment du niveau réel des performances.
4. **Manque de constance :** l'évaluateur n'applique pas la norme avec constance et ses résultats diffèrent de ceux des autres évaluateurs.

La gravité de ces problèmes dépend en partie des corrections qui peuvent y être apportées. Prenons l'exemple de la sévérité. Beaucoup d'évaluateurs ont une nette tendance à la sévérité, tenter de la remettre en question peut avoir pour effet de diminuer leur confiance en eux et par conséquent de les rendre moins constant. Il vaut donc mieux accepter une certaine systématisation dans la sévérité ou le laxisme si cela peut être corrigé par une procédure statistique. L'échelonnage ou le modèle de réponse à l'item sont deux options possibles (cf. annexe VII).

Un éventail trop étroit de notes ne peut être corrigé que partiellement à l'aide des statistiques. Le manque de constance ne peut être corrigé de façon statistique. Ces deux problèmes doivent donc être repérés et la solution apportée sera soit de demander à l'évaluateur de suivre une nouvelle formation soit de ne plus faire appel à cette personne.

Il faut donc mettre en place d'un système de contrôle. Il s'avère plus aisé pour la correction de la production écrite car les évaluateurs peuvent se transmettre la copie pendant la séance de correction. Le contrôle sur la production orale est par contre bien plus difficile, sauf si l'on dispose d'enregistrements. Dans ce cas, l'effort doit porter sur la formation et l'appréciation du travail de l'évaluateur avant la session de correction. Il est recommandé de s'aider de statistiques montrant la performance de l'évaluateur (voir l'annexe VII).

Les différentes approches du contrôle vont de la plus simple – par exemple, vérification ponctuelle informelle et nombreux retours d'information oraux aux évaluateurs, – au plus complexe – par exemple nouvelle correction partielle du travail d'un correcteur et création de statistiques d'indices de performance. Une méthode intéressante consiste à

inclure des copies déjà évaluées à celles attribuées à un évaluateur et de comparer les notes. En fait, pour que cette procédure soit fiable, il faut que les copie ne puissent pas être distinguées les unes des autres afin qu'il ne soit pas possible de les photocopier. Pratiquement, cette méthode ne peut s'appliquer qu'avec des copies issues de tests sur ordinateur ou avec des copies scannées, dans un système d'évaluation en ligne.

Une autre façon de diminuer la marge d'erreur et de comparer les évaluateurs entre eux (ce qui permet d'identifier des effets de l'évaluation et de les corriger statistiquement), est d'opérer une double correction ou une correction multiple partielle consistant à faire corriger un certain nombre de copies par plus d'un correcteur. En fonction de l'approche statistique utilisée, il faudra mettre en place une méthode pour combiner les informations et arriver à donner une note au candidat.

5.2 La notation

Tout le processus de conception, d'élaboration, de passation et de correction qui vient d'être décrit débouche sur l'évaluation de la performance de chaque candidat et la façon de la rapporter.

Dans certains contextes, un test ou un examen classe les candidats en les regroupant du niveau le plus haut au niveau le plus bas en fixant des limites de niveaux arbitraires - par exemple les 10% les plus hauts ont le niveau A, les 30% suivants ont le niveau B et ainsi de suite. *Cette approche, qui se réfère à une norme* qui peut être d'une certaine utilité sociale est peu satisfaisante dans la mesure où la performance est évaluée uniquement par rapport à celle des autres mais ne donne aucune indication sur ce qu'elle signifie en termes de niveau de compétence langagière.

L'alternative, qui est une approche plus significative, *se réfère à des critères*. La performance y est évaluée en tenant compte des critères ou des normes fixes et absolues. C'est en fait l'approche adoptée par les tests ou les examens qui délivrent des résultats en termes de niveaux du CECR.

Un examen peut être conçu sur plusieurs niveaux du CECR ou sur un seul. Dans ce dernier cas, les candidats qui sont du niveau sont considérés comme ayant « réussi » et les autres comme ayant « échoué ». Les degrés de réussite ou d'échec de la performance peuvent aussi être indiqués.

Le fait d'identifier la note qui correspond à la réussite dans un niveau s'appelle la détermination ou la DEFINITION DU SCORE DE CESURE. Cette décision suppose un jugement subjectif si possible fondé sur des faits probants.

Il y a différentes façons d'appliquer la définition des scores de césure dans les épreuves de production (écrites et orales) et de réception (écrite et orale) qui sont souvent corrigées de façon objective. Les épreuves de production sont relativement faciles à traiter. La réception écrite et orale pose plus de problème dans la mesure où il faut interpréter des processus mentaux qui ne sont observables qu'indirectement, ce qui rend donc la notion de niveau de compétence critériée difficile à cerner

Quand un test ou un examen comprend plusieurs sous épreuves de réception ou de production différentes, il faut fixer une norme pour chacune séparément, et ne pas s'occuper de l'ensemble (cf. 5.3 pour plus d'information sur cette question).

Le lecteur est appelé à se référer au Manuel *Relier les examens de langues au CECR* (Conseil de l'Europe 2009) qui traite en détail de la définition des scores de césure. Concernant l'organisation et la terminologie du Manuel, veuillez noter que :

- Le chapitre 6 sur *la définition des scores de césure* ne fait référence qu'aux tests et examens qui donnent lieu à une correction objective (c'est-à-dire à la réception écrite et orale).
- Les épreuves de production sont abordées au chapitre 6 sous le titre de *Formation à la standardisation et au calibrage*.
- Le chapitre 7 sur la VALIDATION est également important. Il y a deux approches pour définir les points de césure : soit centrée sur la tâche, soit centrée sur le candidat. L'activité centrée sur la tâche qui est décrite dans le chapitre 6 dépend d'un jugement d'experts sur les items du test ou de l'examen. En revanche, l'activité centrée sur le candidat suppose la collecte d'information sur celui-ci et est abordée dans le chapitre 7.

- Ce n'est pas pour autant que la définition du score de césure à partir de l'activité centrée sur la tâche est plus importante que celle centrée sur le candidat.
-

Pour être clair, la définition des scores de césure est une opération qui ne devrait être menée qu'une seule fois, quand le test ou l'examen est organisé pour la première fois, même si arriver à la norme désirée est un processus itératif. Avec le temps la notation devrait non plus concerner la définition des normes mais leur *maintien*. Cela suppose que le cycle de conception du test ou de l'examen présente des procédures adéquates. Ces questions sont largement abordées dans le document en supplément du Manuel (North et Jones 2009).

5.3 La délivrance des résultats

C'est à l'utilisateur de décider de la façon de publier les notes du candidat : soit en donnant un résultat global soit en donnant le profil du candidat avec la performance dans chaque composante du test ou de l'examen.

La première option est la plus communément pratiquée car la plupart des parties concernées (candidats, institutions) préfèrent une réponse simple. La seconde option donne plus d'informations qui peuvent être très utiles dans certains cas.

La troisième possibilité est de publier les deux résultats sachant que le CECR apprécie la publication de résultats indiquant le profil du candidat.

Dans le cas où un résultat simple est demandé, il faut mettre au point une méthode permettant de tenir compte des notes attribuées dans chaque activité langagière et pour cela décider du poids qui va leur être attribué, qui peut être le même pour toutes ou plus important pour certaines. Cela suppose quelques ajustements des scores bruts (cf. annexe VII).

Si des certificats sont délivrés, l'utilisateur doit prendre en compte les éléments suivants :

- les informations supplémentaires qui doivent être fournies pour illustrer les niveaux (par exemple les « descripteurs ») ;
- comment s'assurer que le document est l'original (par exemple empêcher toute falsification du document ou mettre en place un service de vérification) ;
- les précisions qui doivent être données sur l'interprétation des résultats.

5.4 Questions clés

- Quelle doit être la proportion de correction du test ou de l'examen qui n'est pas faite par la machine et à quelle fréquence?
- Quelle est la proportion qui concerne l'évaluation et à quelle fréquence ?
- Quel est le niveau d'expertise requis pour vos évaluateurs ?
- Comment vous assurer que la correction et l'évaluation sont justes et fiables ?
- Quelle est la meilleure façon de noter les candidats dans votre contexte ?
- Qui sont les destinataires des résultats et comment allez-vous les délivrer ?

5.5 Lectures complémentaires

Voir ALTE (2006c) pour la liste de contrôle d'auto évaluation pour la correction, la notation et les résultats

Kaftandjieva (2004) North et Jones (2009), Figueras et Noijons (2009) donnent tous des informations sur la définition des scores de césure.

6 Contrôle et révision

Il est important de vérifier le travail accompli pour l'élaboration et l'utilisation du test ou de l'examen. Répond-il aux normes en vigueur ou des changements sont-ils nécessaires ? L'objectif du contrôle est de vérifier si des aspects importants du test ou de l'examen sont acceptables alors que le test est utilisé ou juste après son utilisation. Si des modifications doivent être faites, il est souvent possible de les faire tout de suite. Des améliorations ne peuvent être que bénéfiques aux candidats en cours ou à venir.

La révision est une sorte de projet consistant à passer en revue différents aspects du test ou de l'examen. A cette occasion, on revient sur la conception du test ou de l'examen et on se pose des questions essentielles telles que « quelle est l'utilité du test ? » « quel en est objectif ? », « pour quelle population ? », « que cherche-t-on à tester ? ». Cela ressemble à la phase d'élaboration mais avec l'avantage d'avoir des données et d'avoir acquis l'expérience de l'utilisation. De par son étendue, la révision du test ou de l'examen ne peut pas faire partie du cycle normal de l'évaluation et ne peut pas être organisée à chaque session.

6.1 Le contrôle de routine

Le contrôle fait partie des opérations de routine dans la production et l'utilisation d'un test ou d'un examen. Les preuves nécessaires à la révision vont être utilisées pour s'assurer que tout ce qui concerne le test en cours est au point : les éléments sont conçus correctement, ils sont distribués à temps, les niveaux corrects sont attribués aux candidats, etc. Après cela, les mêmes preuves peuvent être utilisées pour estimer la performance du processus utilisé, tels que les processus de rédaction et de correction des items, le **processus de construction** du test ou de l'examen, le processus de correction, etc. Les preuves peuvent aussi servir pour l'ARGUMENT DE VALIDITE (cf. annexe I) et doivent aussi être pensées en ces termes.

Plusieurs exemples sur la façon de collecter des preuves pour le contrôle ont été présentés dans ce Manuel. Ainsi :

- Faire appel au jugement d'experts, expérimenter et prétester pour s'assurer de la qualité de rédaction des items (cf. 3.4).
- Utiliser les réponses des candidats pour savoir si les items fonctionnent correctement (Annexe VII).
- Utiliser des formulaires pour le retour d'information sur la passation (cf. annexe VI).
- Collecter et analyser des données sur la performance des correcteurs (Annexe VII).

Contrôler l'efficacité du travail est tout aussi important. Ce contrôle permet aux organismes certificateurs de mesurer le temps nécessaire à la préparation et de décider de le raccourcir ou de l'allonger.

6.2 Révision périodique du test ou de l'examen

Des révisions périodiques peuvent avoir lieu hors passation du test ou de l'examen. Cette révision peut être décidée après un certain temps d'utilisation ou après des modifications concernant par exemple la population cible ou des modifications du test lui-même ou du programme auquel il se réfère. La nécessité d'une révision peut aussi avoir été décidée à l'occasion d'un contrôle. Lors de la révision, le test dans son ensemble ainsi que la façon de le concevoir doivent être étudiés de près. Les preuves qui ont été collectées pendant l'utilisation du test ou de l'examen, telles que la performance des évaluateurs analysée à l'occasion du contrôle, peuvent être d'une grande utilité. Enfin l'organisme certificateur peut décider de la nécessité d'un complément de preuves que l'on collectera à l'occasion de la révision.

Au cours de la révision, des renseignements sont réunis sur le test ou l'examen. Ils permettent de décider des aspects à changer (par exemple le concept, le format, les règles de passation…). Il se peut que l'opération de révision ne débouche que sur peu ou pas de modifications.

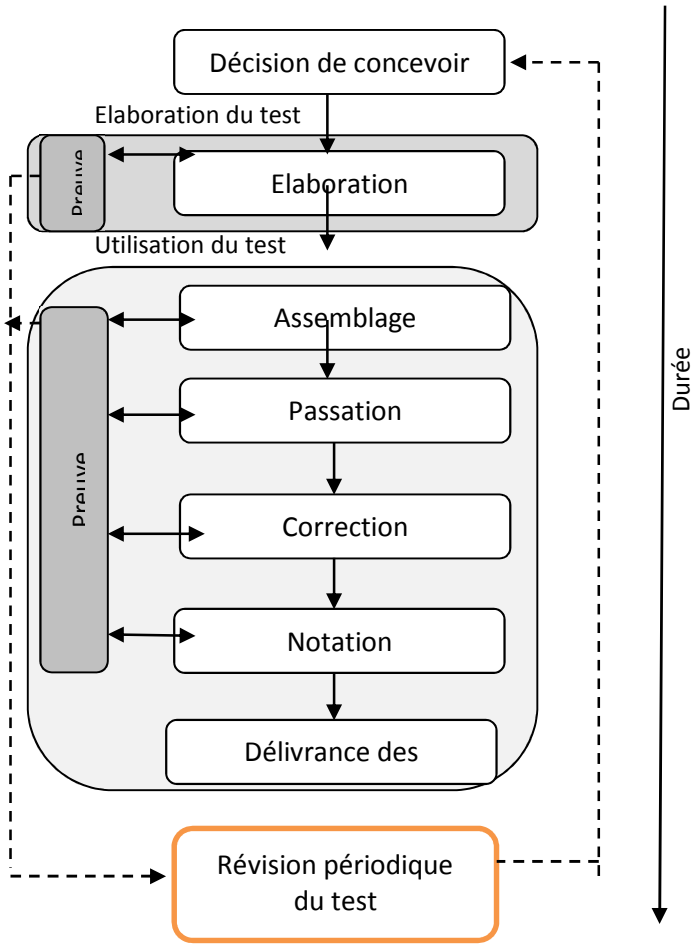


Figure 15 ***Cycle d'élaboration et de la révision d'un test ou d'un examen***

La figure 15 est une reproduction de la figure 5 (partie 1.5.1) à laquelle a été ajoutée la révision périodique. Elle montre l'apport de la révision aux premières étapes du diagramme : la décision de concevoir un test ou un examen. Le processus d'élaboration du test ou de l'examen fait partie de la révision.

Il ne faut pas oublier de prévenir les PARTIES CONCERNEES en cas de modifications (cf. 2.6)

6.3 A quoi servent le contrôle et la révision

Le contrôle et la révision font partie du travail de routine d'élaboration d'un test ou d'un examen. Ils montrent à l'organisme certificateur que tout fonctionne correctement et que des modifications ont été faites pour palier d'éventuels dysfonctionnements. La révision peut également aider à montrer aux autres, directeurs d'école ou partenaires accrédités, qu'ils peuvent avoir confiance dans le test ou l'examen.

De toute façon, analyser ce qui est fait pour savoir si cela convient constitue en quelque sorte un audit de l'argument de validité.

ALTE (2007) a établi une liste de 17 normes, appelées NORMES MINIMALES qui permettent aux organismes certificateurs de structurer leur argument de validité. Ces normes sont répertoriées dans les cinq domaines suivants :

- Conception du test ou de l'examen
- Passation et logistique
- Correction et classement
- Analyse de test
- Communication avec les parties prenantes

Ces normes peuvent être utilisées avec des listes plus spécifiques et détaillées, telles que les listes de contrôle d'analyse du contenu de ALTE (ALTE 2004a-k,2005,2006a-c).

Les organismes certificateurs peuvent utiliser d'autres outils pour concevoir et vérifier les arguments de validité. Jones, Smith et Talley (2006 :490-2) propose une liste de 31 points clés pour l'évaluation sur une échelle plus réduite. L'essentiel de leur liste est inspiré des normes pour l'évaluation éducationnelle et psychologique (AERA et al 1999).

6.4 Les questions clés

- Quelles données faut-il collecter pour un contrôle efficace ?
- Certaines de ces données sont-elles déjà collectées permettant de prendre des décisions en cours d'utilisation du test ou de l'examen ? Comment peuvent-elles servir à la fois au contrôle et à l'utilisation ?
- Ces données peuvent-elles être gardées et utilisées par la suite pour la révision ?
- Qui doit prendre part à la révision ?
- Quelles ressources sont nécessaires pour la révision ?
- Avec quelle fréquence la révision doit-elle se faire?
- Dans la liste des points clés, quels sont ceux qui peuvent servir à la vérification de l'argument de validité.

6.5 Lectures complémentaires

ALTE (2007) propose des rubriques auxquelles se référer pour évaluer un test ou un examen.

Voir ALTE (2002) pour consulter la liste d'auto-évaluation pour l'analyse statistique et la révision.

Fulcher et Davidson (2009) ont une façon intéressante d'illustrer la hiérarchie des preuves pour la révision d'un test ou d'un examen. Ils utilisent la métaphore de la construction pour savoir quelles parties d'un test ou d'un examen doivent être modifiées régulièrement et celles qui doivent être modifiées en permanence.

Les descriptions des différents aspects de la révision d'un test ou d'un examen sont proposés par Weir et Milanovic (2003).

Annexe I – Développer un argument de validité

Cette annexe aborde une approche de la VALIDATION incluant le développement d'un ARGUMENT DE VALIDITÉ. La démarche est plus détaillée que les grandes lignes présentées en 1.2.3. Elle montre qu'il ne s'agit pas d'une succession d'étapes discontinues, mais qu'elles sont toutes imbriquées et en corrélation.

Kane (2006), Kane, Crooks et Cohen (1999), Bachman (2005) et Bachman et Palmer (2010) font une description plus complète des argumentas de validité. La validation est en effet un processus continu, dans lequel on ajoute et on précise les preuves au fur et à mesure.

L'interprétation et l'usage des résultats au test constituent le point central de l'argument de validité, lorsque l'on définit la validité comme « le degré auquel la preuve et la théorie confirment les interprétations des scores en fonction des utilisations prévues des tests. » (AREA et al 1999).

Un argument de validité consiste en une série de propositions qui décrivent les raisons selon lesquelles les interprétations conseillées des résultats au test sont valides, et apportent la preuve pratique et la théorie qui les étayent. Cette annexe donne une vue d'ensemble sur la façon de procéder.

Lors de la présentation de l'argument aux différentes PARTIES PRENANTES, on commence par préciser clairement de quelle façon les résultats au test devraient être interprétés pour chaque utilisation. Un ARGUMENT D'UTILISATION DE L'EVALUATION (également appelé argument d'interprétation), explique cet état de fait. Ce que l'on appelle simplement un argument de validité vient légitimer l'argument d'utilisation à l'aide de la théorie et de preuves.

La figure 16 illustre de façon conceptuelle l'argument d'utilisation selon Bachman (2005). Il s'agit d'un raisonnement en quatre étapes (montrées chacune par une flèche), qui justifie l'utilisation du test. Chaque étape apporte la base conceptuelle de l'étape suivante. Des scores fiables par exemple (score universel), ne sont utiles que s'ils représentent la performance au test (score observé). Il ne s'agit *pas* d'une suite d'étapes à compléter obligatoirement dans l'ordre. Les preuves permettant d'étayer chaque étape peuvent provenir de différentes étapes du développement et de la production du test.

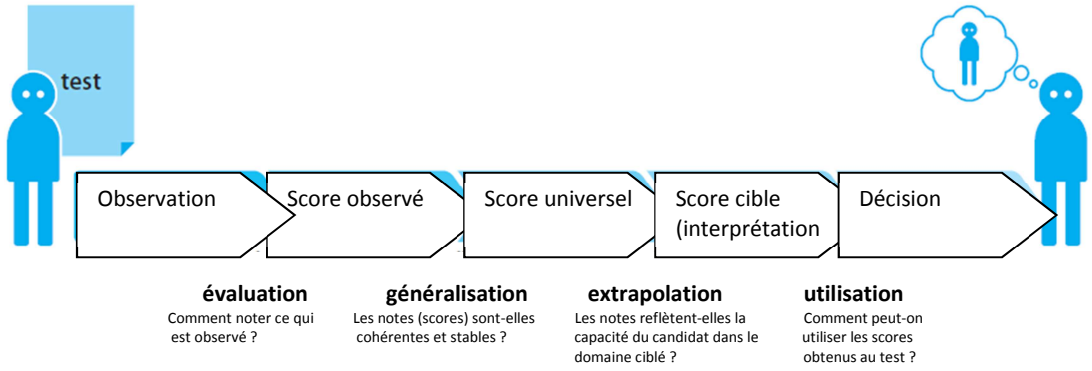


Figure 16 ***Chaîne du raisonnement dans un argument de validité (adapté de Kane, Crooks, Cohen 1999, Bachman 2005)***

La justification de l'argument de validité est d'étayer l'argument d'utilisation et consiste en preuves, théorie et propositions raisonnées. Les preuves qui étayent chaque étape sont réunies pendant le développement du test, sa construction et son utilisation.

De nombreuses preuves utilisées dans l'argument de validité proviendront du processus de routine de l'utilisation du test. Des exemples de ce type de preuves sont énumérés dans la section 6.1. Les preuves sont également réunies pour un objectif plus immédiat, qui est de contrôler le travail du correcteur, et servent naturellement lorsqu'on établit l'argument de validité, comme le montre la figure 17.

Grâce aux preuves, on peut améliorer l'argument de validité à chaque fois que l'on développe et utilise une nouvelle forme de test. La conception de l'argument de validité devrait débuter lors de la tout première étape du processus, lorsque l'on définit les objectifs du test. Une grande partie de l'argument de validité pour une forme donnée du test peut être utilisée pour la forme suivante.

Certains théoriciens (Bachman 2005, Mislevy et al 2003) soulignent que l'argument de validité devrait être présenté comme un *argument informel*, en opposition à un argument logique. Cela signifie que le raisonnement seul ne peut établir que

l'argument est juste ou erroné. Il peut en revanche sembler plus ou moins convaincant à quelqu'un qui l'examinerait. Le degré de crédibilité dépend de la théorie appliquée et des preuves disponibles pour l'étayer.

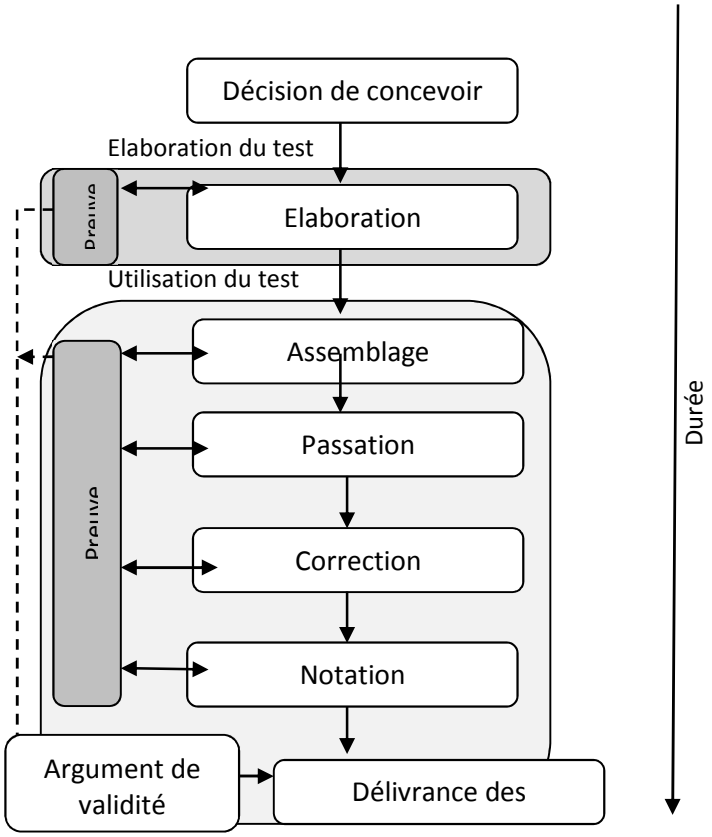


Figure 17 ***Le cycle du test, la révision périodique et l'argument de validité***

Il est possible que l'argument de validité paraisse moins convaincant à cause d'une nouvelle preuve ou d'une nouvelle théorie ou par une interprétation différente d'une preuve existante. Il est également possible que les fournisseurs de tests appuient, de façon non intentionnelle, leur interprétation favorite, sans être suffisamment critiques. Lorsqu'on a développé l'argument pour la première fois, on doit alors le remettre en question, même si cela doit modifier l'interprétation souhaitée des résultats. On peut par exemple examiner les différentes façons d'interpréter une preuve ou de vérifier que toutes les conclusions auxquelles on est arrivé sont bonnes. Les fournisseurs de tests pourraient alors revoir leur argument en y faisant les changements nécessaires et en présentant les raisons de l'interprétation des preuves.

Des exemples de preuves utilisables pour étayer un argument sont donnés dans ces annexes. Des exemples des différentes façons de comprendre les preuves sont également donnés. Ces exemples sont tirés des travaux de Kane (2004) et Bachman (2005). Ils correspondent au sommaire de ce Manuel : Développer des tests ou des examens ; assembler des tests ou des examens ; corriger, noter et délivrer les résultats. Les fournisseurs de tests peuvent commencer la conception de leur propre argument de validité à l'aide de ces preuves. Leur liste n'est cependant pas exhaustive.

Lectures complémentaires

ALTE (2005 : 19) propose un résumé utile des différents types de validité et décrit le contexte de la conception moderne de la validité.

AERA et al (1999) expose les grandes lignes du concept moderne de la validité et des standards qui soulignent certains aspects spécifiques du problème et peuvent ainsi aider à la conception d'un argument de validité.

Messik (1989) débat du concept unitaire de la validité ainsi que des considérations éthiques qui en découlent.

Haertel (1999) exemplifie la façon dont les preuves et l'argumentation sont reliées aux interprétations des scores.

Kane, Crooks et Cohen (1999) présentent de façon claire les premières étapes de l'argument de validité. Cela est traité de façon plus approfondie par Kane (2006).

Bachman (2005) examine la relation entre les arguments de validité et l'évaluation en langues. Il relie également entre eux les modèles de Bachman et de Palmer (1996) du modèle d'argument de validité. Le premier modèle considérait la notion d'utilité comme étant la plus grande qualité d'un test, réunissant la fiabilité, la validité, l'authenticité, l'interactivité et l'impact.

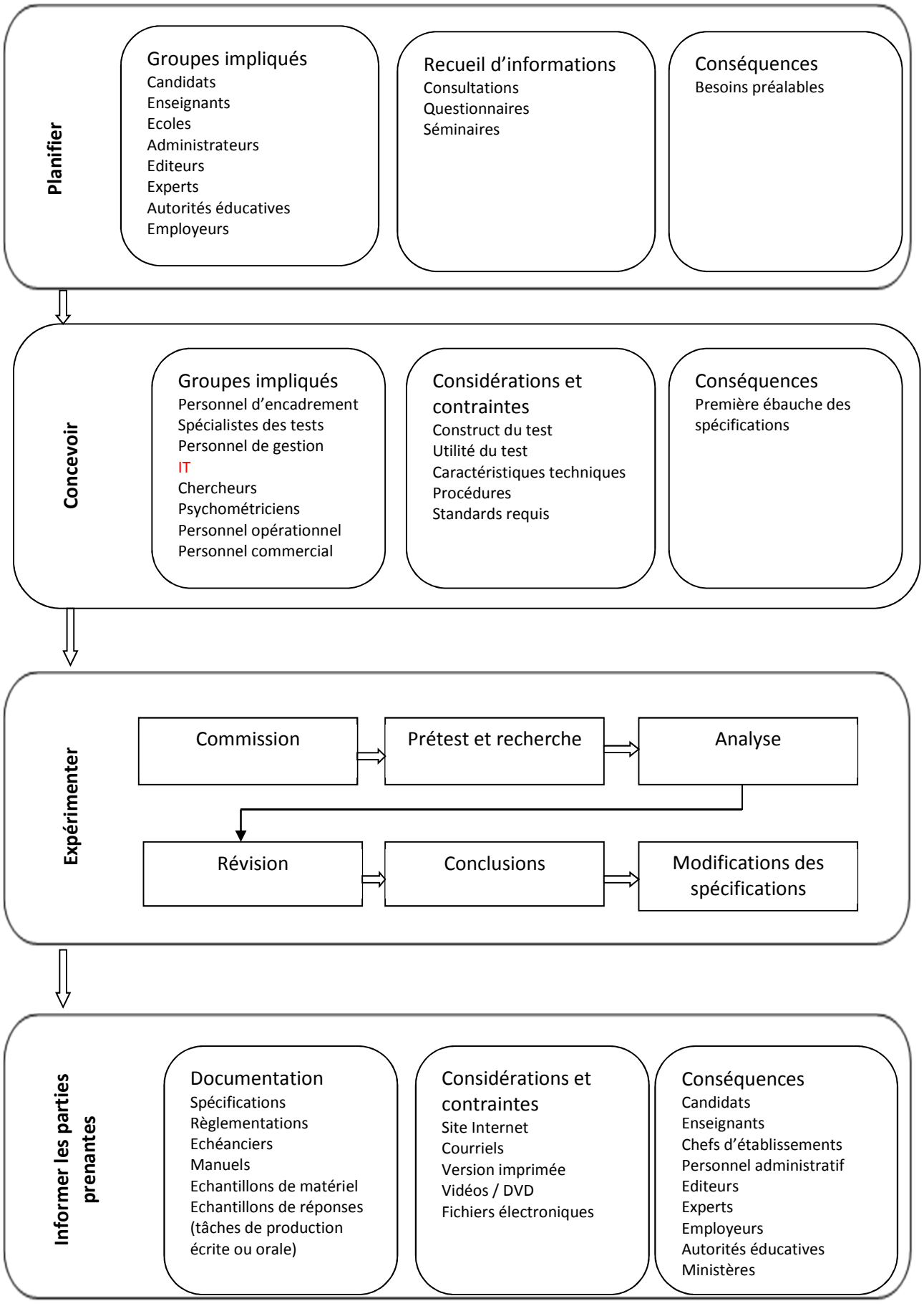
Bachman et Palmer (2010) expliquent comment les arguments de validité sont au cœur du développement du test et peuvent proposer un cadre pour ces tâches.

	Evaluation	Généralisation	Extrapolation	Utilisation
	Comment noter ce qui est observé ?	**Les notes (scores) sont-elles cohérentes et stables ?**	**Les notes reflètent-elles la capacité du candidat dans le domaine ciblé ?**	**Comment peut-on utiliser les scores obtenus au test ?**
Développement du test (section 2)				
Preuve en faveur		Un test format standard est défini par les spécifications – cela signifie que les différentes versions du test sont semblables (cf. section 2 et annexe III)	Un domaine d'utilisation est clairement défini dans le Guide du rédacteur d'items et dans les spécifications. Ce domaine peut également avoir été identifié par une analyse de besoins (cf. section 2.4). La preuve que les notes de réussite au test ont été convenablement établies viendra étayer l'interprétation des résultats de chaque candidat (cf. sections 2.0 et 5.2).	
Preuve contre			Certaines parties du CONSTRUIT n'apparaissent pas clairement dans les spécifications. Cela signifie que les résultats au test n'apporteront pas d'information pertinente sur ce que le candidat est capable de faire (cf. sections 1.1 et 2).	
Production du test (section 3)				
Preuve en faveur	Toutes les clés de correction sont correctes. On peut, pour le vérifier, utiliser des grammaires et des dictionnaires et faire appel à des experts.	Les items de chaque version du test sont représentatifs du construit. Cela ne signifie pas que toutes les parties du construit sont à chaque fois représentées mais que ces parties ont été sélectionnées d'une façon comparable (cf. sections 2, 3.5 et annexe VII). La façon de relier les versions entre elles est convenable (cf. annexe	La rédaction d'items et la construction du test ont été confiés à des experts (cf. section 3.2).	

		7). Si des analyses statistiques ont été utilisées, on a trouvé de faibles niveaux d'erreur et les MODELES statistiques convenaient aux données (cf. annexe VII).		
Preuve contre		Les versions du test n'ont pas été reliées entre elles. Les versions du test ne sont pas représentatives du même construit.	Certaines parties du construit n'apparaissent pas suffisamment dans le matériel de test. Cela signifie que les résultats n'apporteront pas d'information pertinente sur ce que le candidat est capable de faire.	
Passation du test (section 4)				
Preuve en faveur	Des procédures ont été appliquées pendant la passation du test. Cela permet de montrer que les résultats ne dépendront pas d'autres facteurs (tels que trop ou trop peu de temps) (cf. section 4.2).	Des procédures ont toujours été respectées et permettront de montrer que les versions du test sont comparables au fur et à mesure des passations (cf. section 4.2).		
Preuve contre	Une fraude non traitée signifiera que les scores ne sont pas représentatifs des capacités du candidat.	Une fraude non détectée signifiera que les scores de quelques candidats ne sont pas suffisamment représentatifs de leurs capacités langagières. Cela se répercutera probablement dans les différentes versions du test.	Les résultats peuvent avoir été affectés par des facteurs extérieurs. Cela peut provenir du fait que les procédures de passation n'ont pas été respectées. Cela se répercutera sur les résultats au test. Ces résultats n'impacteront probablement pas uniquement les capacités langagières (cf. section 4.1)	
Correction, notation et délivrance des résultats (section 5)				
Preuve en faveur	Des procédures ont été appliquées pendant la correction. Cela permet de montrer que le score ne dépend pas d'autres facteurs (tels que la prise en compte d'une clé incorrecte ou encore des erreurs de scan) (cf. section 5.0). La correction a été juste et fiable (cf. section 5.1 et annexe VII).	La preuve de la fiabilité du score (généralement une preuve statistique), montre que la version du test donne une mesure cohérente de la performance du candidat (cf. sections 1.3 et 5.1, et annexe VII). Si les données de quelques candidats ont été analysées, ces données sont représentatives de l'ensemble des candidats (cf. annexe VII). Les points de césure qui ont de bas niveaux d'erreur indiqueront que les candidats sont vraisemblablement placés du côté correct de	L'emploi de correcteurs experts signifie que les corrections reflètent vraisemblablement le domaine d'intérêt (cf. section 5.1). De même, l'utilisation d'échelles de notation bien conçues augmentera les chances que les performances soient corrigées en fonction du domaine ciblé (cf. sections 2.5 et 5.1.3).	Si l'on suit des règles pour prendre des décisions spécifiques basées sur les résultats au test, il est probable que le test est utilisé comme il avait été prévu de l'être et que les effets indésirables seront minimisés (cf. sections 1.2, 5.3 et annexe 1).

		l'échelle (cf. annexe VII).		
Preuve contre		Si des données d'un groupe non représentatif de l'ensemble des candidats a été utilisé à des fins d'analyses, l'analyse peut contenir des erreurs ou des biais (cf. annexe VII).		Si aucune procédure standard ni aucune règle n'ont été suivies pour prendre des décisions, le test peut être utilisé de façon inappropriée (cf. sections 1.5 et 5.3).

Annexe II – Le processus de développement du test ou de l'examen



Annexe III – Exemple du format de l'examen – examen d'anglais

Contenu et vue d'ensemble

Durée	Format		Nombre de questions	Objectif
Compréhension écrite **1 heure**	**Partie 1**	TACHE D'APPARIEMENT à partir d'un texte suivi divisé en 4 sections informatives ; environ 250 à 350 mots au total.	7	L'accent est mis sur le balayage du texte et sur la compréhension générale.
	Partie 2	TACHE D'APPARIEMENT à partir d'un texte unique (article, reportage, etc.) avec des phrases manquantes ; environ 450 à 550 mots.	5	Compréhension de la structure du texte.
	Partie 3	QCM à 4 choix à partir d'un texte unique ; environ 450 à 550 mots.	6	Compréhension générale et recherche d'informations spécifiques.
	Partie 4	Test de closure en QCM à 4 choix à partir d'un texte informatif où du lexique manque ; les trous sont présents dans le texte ; environ 200 à 300 mots.	15	Lexique et structure.
	Partie 5	Tâche de relecture corrective impliquant l'identification de mots supplémentaires non nécessaires dans un texte court ; environ 150 à 200 mots.	12	Compréhension de la structure des phrases et recherche d'erreurs.
Production écrite **45 minutes**	**Partie 1**	Message, note ou courriel. Le candidat doit produire un écrit communicatif basé sur un seul sujet (ainsi que la mise en page de son texte) ; 40 à 50 mots.	1 tâche obligatoire	Donner des instructions, expliquer un événement, demander des précisions, des informations, accepter des demandes.
	Partie 2	Courrier d'affaires, court rapport ou projet. Le candidat doit produire une lettre, un court rapport ou un projet à partir d'un sujet et d'une ou plusieurs idées données ; 120 à 140 mots.	1 tâche obligatoire	Courrier : Expliquer, s'excuser, rassurer, se plaindre. Rapport : décrire, résumer. Projet : décrire, résumer, recommander, persuader.
Compréhension orale **40 minutes**	**Partie 1**	Tâche de complétion à partir de trois monologues ou dialogues d'environ 1 minute chaque. Deux écoutes.	12	Prise de notes.
	Partie 2	Tâche d'appariement multiple à partir de deux fois cinq courts monologues.	10	Identification d'un sujet, d'un contexte, d'une fonction, etc.
	Partie 3	Tâche d'appariement multiple à partir d'un monologue, une interview ou une discussion d'environ 4 minutes. 2 écoutes.	8	Compréhension des points principaux et identification d'informations spécifiques.
Production orale **14 minutes**	**Partie 1**	Conversation entre un interlocuteur et chaque candidat (questions orales)	Plusieurs	Donner des informations personnelles. Parler des circonstances présentes, des expériences passées et des projets futurs, exprimer des opinions, faire des suppositions, etc.
	Partie 2	« Mini-présentation » par chaque candidat. Chaque candidat reçoit un choix de trois Thèmes liés au domaine commercial et dispose d'une minute pour préparer un exposé d'une minute.	Une présentation par candidat	Organiser son discours. Donner des informations, exprimer et justifier des opinions.
	Partie 3	Tâche collaborative. Les candidats engagent une discussion dans le domaine commercial. L'interlocuteur réplique en élargissant la discussion à des sujets voisins.	Plusieurs	Commencer une discussion et répondre, négocier, collaborer, échanger des informations, exprimer et justifier des opinions, approuver ou désapprouver, suggérer, faire des suppositions, comparer et exposer des différences, prendre des décisions.

Exemple pour la compréhension écrite

Description générale	
Format de l'examen	L'examen consiste en une série de textes dans le domaine commercial et de tâches à réaliser. Un texte peut être constitué de plusieurs sections courtes.
Durée	1 heure
Nombre de parties	5 parties Parties 1 à 3 : compréhension de lecture Parties 4 et 5 : compréhension de lexique, locutions, phrases et paragraphes.
Nombre de questions	45
Type de tâches	Appariement QCM 4 choix Test de closure en QCM 4 choix Relecture corrective
Type de textes	Textes informatifs, articles et reportages.
Longueur des textes	De 150 à 550 mots
Format des réponses	Les candidats indiquent leurs réponses en grisant une case ou en écrivant un mot sur une feuille de réponse lisible par une machine.
Note	Toutes les questions sont notées

Annexe IV – Conseils aux rédacteurs d'items

Conseils sur le choix des textes

La définition d'un « texte » dans ce Manuel se réfère à celle donnée dans le CECR (section 4.6). « On appelle « texte » toute séquence discursive orale ou écrite.

Les instructions aux rédacteurs d'items sur la façon de choisir les textes doivent prendre en compte les points suivants :

- Les meilleures sources (qualité des articles de journaux, brochures)
- Les sources le moins à même de fournir des textes acceptables (matériels spécialisés)
- Une information générale sur la façon d'éviter les biais (biais culturel, biais de sexe, d'âge, …)
- Une liste de motifs de rejet des textes, parmi lesquels :
    - Les textes font appel à trop de connaissances culturelles ou locales (sauf si c'est précisément ce qui est évalué).
    - Les thèmes ne conviennent pas au groupe de candidats ciblés. Par exemple : la guerre, la mort, la politique, les croyances religieuses, ou d'autres thèmes qui peuvent choquer ou bouleverser certains candidats.
    - Des thèmes qui ne conviennent pas à la classe d'âge des candidats.
    - Un niveau lexical ou conceptuel trop élevé ou trop faible.
    - Des erreurs ou des idiosyncrasies techniques ou stylistiques.
    - Une rédaction originale médiocre.
- Une liste des thèmes déjà utilisés de nombreuses fois et qu'il n'est donc plus nécessaire de proposer.

Les chapitres 4 et 7 du CECR situent les textes dans le contexte de l'utilisation de la langue. Les médias listés dans la section 4.6.2 (voix, téléphone, radio, etc.) ainsi que les genres et types de textes oraux et écrits de la section 4.6.3 sont extrêmement utiles en tant que vérifications et possibilités de diversifier les types d'items.

Conseils sur la présentation

On peut recommander aux rédacteurs d'items de prendre en compte les points suivants :

- A quel texte affecter un interligne double
- Quelles informations faire figurer dans l'en-tête
- Faut-il joindre une photocopie ou le texte original
- A quel point détailler la source (exemple : date de publication)

Conseils détaillés pour chaque tâche

Voici un exemple fictif de conseils donnés aux rédacteurs d'items pour un test de closure modifié, conçu pour évaluer des mots de type plus structurels que lexicaux :

- Rechercher un texte authentique d'environ 200 mots, comportant un titre court. L'accent est mis uniquement sur les structures. Le texte ne doit pas comporter trop de vocabulaire inconnu.
- Pour être sélectionné après le prétest, il doit y avoir au minimum 16 items et plus si possible. Le premier item servira d'exemple et portera le numéro « 0 » (zéro). Les items évalueront les prépositions, les pronoms, les modificateurs, les verbes auxiliaires, etc. Ils seront répartis au hasard dans le texte et on veillera à ce qu'une réponse fausse n'induise pas une erreur à la réponse suivante (interdépendance des items).
- On ne supprime généralement ni le premier mot de la phrase, ni une forme contractée, car dans ce cas, le candidat ne saura pas s'il compte pour un mot ou pour deux. On évite également de supprimer un mot si la phrase reste grammaticalement correcte sans lui (par exemple en supprimant le mot « tous » dans la phrase « Nous avons été informés que *tous* les trains avaient du retard »). On évite de même les items qui traitent de structures très inhabituelles ou idiosyncratiques.

L'intitulé courant à utiliser pour cette tâche doit également être donné au rédacteur d'items.

Les rédacteurs qui ont l'habitude des items à partir de textes trouvent souvent, de façon continue, de bons textes, à partir des sources recommandées. Lorsqu'on leur commande des items, ils travaillent à partir des textes les plus intéressants qu'ils ont déjà sélectionnés. Le rédacteur d'item doit pouvoir disposer d'un dictionnaire ou d'un thésaurus pour rédiger certains

types d'items (par exemple ceux de grammaire et de vocabulaire). Lorsqu'il rédige du matériel de compréhension orale, il doit écouter les passages de façon à rédiger les items à partir de l'enregistrement et non de sa transcription.

De nombreux rédacteurs d'items trouvent utile de tester les tâches conçues auprès d'un collègue ou d'un ami non impliqués dans l'évaluation en langues. Cela peut aider à repérer des fautes de frappe, des consignes peu claires, des clés erronées, des items pour lesquels la réponse est très difficile ou bien qui comportent plus d'une réponse correcte.

Les SPECIFICATIONS doivent proposer des listes de contrôle que le rédacteur d'item peut utiliser pour vérifier le texte, les items et la tâche dans son ensemble, avant de les soumettre. La liste de contrôle de la tâche de closure modifiée est donnée ci-dessous à titre d'exemple. Si le texte, les items et la tâche conviennent, le rédacteur doit pouvoir répondre « oui » à chacune des questions suivantes.

texte
Le thème du texte est-il accessible / culturellement acceptable, etc. ?
Le texte est-il débarrassé de tout contenu indélicat ?
Le texte est-il au bon niveau de difficulté ?
Le texte est-il approprié pour une tâche centrée sur les structures ?
Le texte est-il suffisamment long pour qu'on puisse rédiger 16 items ?
Le texte comporte-t-il un titre approprié ?
Items
Le nombre d'items demandés a-t-il été pris en compte ?
Les items sont-ils bien répartis dans le texte ?
A-t-on pris en compte une gamme suffisante de langage ?
A-t-on vérifié que tous les items mettent l'accent sur les structures ?
A-t-on vérifié que les items ne sont pas interdépendants ?
A-t-on ajouté un ou deux items supplémentaires ?
Est-ce que les items idiosyncratiques ont été évités ?
Sujet et clé
Est-ce que les intitulés ont été vérifiés ?
Est-ce qu'un exemple a été donné ?
Est-ce que toutes les clés ont été fournies sur une feuille à part ?

Avant de soumettre leur matériel, les rédacteurs d'items doivent vérifier qu'ils en ont bien gardé une copie. Si les originaux des textes ont été fournis, le rédacteur d'items doit en garder une photocopie sur laquelle il aura reporté les détails de la source originale.

Annexe V – Etude de cas – révision d'une tâche de niveau A2

Cette annexe montre les modifications apportées à une tâche lors de sa révision et explique les raisons de ces modifications.

Chaque nouvelle version comporte des commentaires. Les parties qui font débat apparaissent en rouge.

Version 1 – soumise par le rédacteur d'items pour révision (réunion 1)

Complétez la conversation entre deux amis.

Que dit Josh à son amie Marta ?

Pour les questions **1 à 5**, écrivez la lettre correcte **A à H** sur la feuille de réponses.

Exemple 0

Marta : Salut, Josh ! C'est chouette de te voir ! C'était comment, tes vacances ?

Josh : 0 _____.E

réponse 

Marta : Où es-tu allé cette année ?
Josh : **1 ……**
Marta : Il a fait beau ?
Josh : **2 ……**
Marta : Super ! Tu as pris des photos ?
Josh : **3 ……**
Marta : Vous étiez à l'hôtel ?
Josh : **4 ……**
Marta : Ça avait l'air super ! Tu y retourneras ?
Josh : **5 ……**
Marta : Ce serait probablement plus intéressant.

A	Il faisait assez chaud. Je suis allé nager.
B	Non, on était chez des amis de mon oncle.
C	Je pensais qu'elles étaient vraiment super !
D	Sans doute pas. Je voudrais aller ailleurs l'année prochaine.
E	C'était super, merci !
F	J'en ai pris des super bonnes !
G	On n'avait pas assez d'argent.
H	Je suis allé chez mon oncle en Islande

Clés : 1H, 2A, 3F, 4B, 5D

Nouvelle vérification de la version soumise pour révision (réunion 1)

Lors de la première réunion de révision, il a été demandé au rédacteur d'items de soumettre à nouveau la tâche après les modifications suivantes :

- Eviter la répétition du modèle question / réponse dans la conversation.
- Eviter la répétition de vocabulaire.
- Modifier les distracteurs **G** et **C** ainsi que le texte qui y correspond.
- Reformuler le lexique et les structures qui ne sont ni dans la liste de vocabulaire ni dans les spécifications grammaticales.

La première modification était nécessaire afin d'éviter que la tâche soit trop facile et centrée sur des réponses et des questions isolées. Dans la version originale, chaque blanc évaluait la réponse de Josh à une question posée par Marta. On a demandé au rédacteur de varier le modèle d'interaction (en transformant par exemple les choix A-H en questions) et de reformuler certaines parties (en ajoutant par exemple une proposition à la fin du choix F) de façon à obtenir un dialogue plus cohérent.

La deuxième modification était d'éviter que la même forme verbale apparaisse à la fois dans les questions et dans les réponses et rendent alors la tâche trop facile. Par exemple : « tu as pris des photos ? » et « J'en ai pris des super bonnes » ; « Où es-tu allé cette année ? » et « je suis allé chez mon oncle en Islande ». On a également demandé au rédacteur de varier le vocabulaire. « Super », par exemple, apparaît cinq fois.

La troisième modification était d'éviter que les distracteurs **C** et **G** soient des clés possibles. On a demandé au rédacteur de les reformuler, ainsi que le texte qui y correspond, de façon à ce qu'ils ne puissent pas être des réponses correctes pour l'item 3, et de s'assurer que **G** ne pouvait correspondre à l'item 4.

Le quatrième changement était lié au niveau de difficulté du contenu des tâches. On a par exemple demandé au rédacteur de reformuler « probablement », qui n'est pas dans la liste de vocabulaire, ainsi que « ce serait … », qui n'est pas dans la liste des fonctions pour cet examen.

Version 2 – tâches modifiées soumises à nouveau par le rédacteur

Modifications effectuées par le rédacteur d'items :

i. Les modèles d'interaction sont diversifiés ; « probablement » et « ce serait » ont été supprimés.
ii. Le choix C a été modifié de façon à ne plus correspondre à l'item 3.
iii. Le texte précédant et suivant le blanc de l'item 4 a été changé de façon à éliminer le choix G pour les items 3 et 4.
iv. Le verbe « prendre » a été supprimé du choix F.
v. Selon le rédacteur d'items, remplacer « allé » par « rendu visite » aurait donné un aspect non naturel au dialogue. Pour dissimuler ces deux occurrences du verbe « être », des segments de texte ont été rajoutés.

Complétez la conversation entre deux amis.
Que dit Josh à son amie Marta ?

Pour les questions **1 à 5**, écrivez la lettre correcte **A à H** sur la feuille de réponses.

Exemple 0

Marta : Salut, Josh ! C'est chouette de te voir ! C'était comment, tes vacances ?

Josh : 0 _____.E

réponse 

Marta : Où es-tu allé cette année ? A nouveau chez ton oncle ?
Josh : **1 ……**
Marta : Non, il fait trop froid pour moi là-bas.
Josh : **2 ……**
Marta : Super ! Tu as pris des photos ?
Josh : **3 ……**
Marta : Oui, s'il te plaît. Vous étiez à l'hôtel ?
Josh : **4 ……**

Marta : Tu as eu de la chance !
Josh : **5 ……**
Marta : Je ne savais pas. Il faudra que tu me racontes.

A	Non, pas vraiment, l'été il fait assez chaud. On peut même se baigner dans la mer.
B	Mon oncle a des amis là-bas. On a habité chez eux.
C	Non, mais est-ce que tu as passé de bonnes vacances ?
D	Oui, les hôtels sont très chers là-bas.
E	C'était super, merci !
F	Beaucoup. Je te les montrerai si tu veux.
G	On n'avait pas assez d'argent.
H	On a fait quelque chose d'autre. On est allé en Islande. Tu connais ?

Clés : 1H, 2A, 3F, 4B, 5D

Version 2 – La tâche réécrite, soumise à nouveau par le rédacteur, après la discussion de révision (réunion 2)

Complétez la conversation entre deux amis.
Que dit Josh à son amie Marta ?

Pour les questions **1 à 5**, écrivez la lettre correcte **A à H** sur la feuille de réponses.

Exemple 0

Marta : Salut, Josh ! C'est chouette de te voir ! C'était comment, tes vacances ?

Josh : 0 _____.E

réponse 

Marta :	Où es-tu allé cette année ? A nouveau chez ton oncle ?
Josh :	**1 ……**
Marta :	Non, il fait trop froid pour moi là-bas.
Josh :	**2 ……**
Marta :	Super ! Tu as pris des photos ?
Josh :	**3 ……**
Marta :	Oui, s'il te plaît. Vous habitiez à l'hôtel ?
Josh :	**4 ……**
Marta :	Tu as eu de la chance !
Josh :	**5 ……**
Marta :	Je ne savais pas. Il faudra que tu me racontes.

A	Non, pas vraiment, l'été il fait assez chaud. On peut même se baigner dans la mer.
B	Mon oncle a des amis là-bas. On a habité chez eux.
C	Non, mais est-ce que tu as passé de bonnes vacances ?
D	Oui, les hôtels sont très chers là-bas.
E	C'était super, merci !
F	Beaucoup. Je te les montrerai si tu veux.
G	On n'avait pas assez d'argent.
H	On a fait quelque chose d'autre. On est allé en Islande. Tu connais ?

Clés : 1H, 2A, 3F, 4B, 5D

Vérification de la version soumise à nouveau pour révision (réunion 2)

A la seconde réunion de révision, les modifications suivantes ont été effectuées :

- Suppression de « son amie » (deuxième ligne de l'intitulé du sujet).
- Modification de la deuxième réplique de Marta « Non, il fait trop froid pour moi là-bas » (entre les blancs 1 et 2).
- Modification du choix **A** « Non, l'été il fait assez chaud. On peut même se baigner dans la mer. ».
- Modification du choix **B** « Mon oncle a des amis là-bas. On a habité chez eux. ».

Le premier changement avait des raisons stylistiques : éviter la répétition du mot « ami », qui apparaît à la première ligne de la consigne.

Il y a deux raisons à la modification de la deuxième réplique de Marta. La première était d'éliminer le choix **B** pour l'item 2 (la clé est **A**). La seconde était de donner une indication sur le contenu du blanc. Dans la première version, la conversation peut changer de sujet après l'intervention de Marta, mais si cette intervention est une question, la réponse de Josh était presque évidente.

Le choix **A** a été changé pour être cohérent avec la deuxième réplique de Marta. « Non, pas vraiment » est devenu une réponse à la question de Marta. L'information sur l'été et les baignades dans la mer ont été gardées mais légèrement modifiées.

Il y a eu également deux raisons aux changements apportés au choix **B**, qui est la clé de l'item 4. La première a été de supprimer la référence à l'oncle de Josh, car cela redirigeait le choix plus vers le début de la conversation que vers le quatrième blanc. La référence à « eux » portait également à confusion. C'est pourquoi on a préféré faire la distinction entre la maison de l'oncle et la maison des amis de l'oncle. Les candidats qui n'ont pas fait cette distinction n'ont pas pu choisir le choix **B** pour l'item 4. Le choix B a donc été changé pour « Nous avons des amis là-bas ». La seconde raison de la modification du choix **B** a été d'éviter une redite lexicale : « Habiter », dans « Vous habitiez à l'hôtel ? » apparaît dans la question de Marta avant l'item 4 et il apparaît également dans le choix B. Cette dernière occurrence a été changée pour « Nous avons dormi chez eux ».

Version 3 – Version utilisable en prétest, incluant les changements effectués lors de la seconde réunion de révision

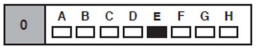
Complétez la conversation entre deux amis.
Que dit Josh à Marta ?

Pour les questions **1 à 5**, écrivez la lettre correcte **A à H** sur la feuille de réponses.

Exemple 0

Marta : Salut, Josh ! C'est chouette de te voir ! C'était comment, tes vacances ?

Josh : 0 _____.E

réponse 

Marta :	Où es-tu allé cette année ? A nouveau chez ton oncle ?
Josh :	**1 ……**
Marta :	Non, il ne fait trop froid là-bas ?
Josh :	**2 ……**
Marta :	Super ! Tu as pris des photos ?
Josh :	**3 ……**
Marta :	Oui, s'il te plaît. Vous habitiez à l'hôtel ?
Josh :	**4 ……**
Marta :	Tu as eu de la chance !
Josh :	**5 ……**
Marta :	Je ne savais pas. Il faudra que tu me racontes.

A	Non, pas vraiment, l'été il fait assez chaud. On peut même se baigner dans la mer.
B	Mon oncle a des amis là-bas et on a dormi chez eux.
C	Non, mais est-ce que tu as passé de bonnes vacances ?
D	Oui, les hôtels sont très chers là-bas.
E	C'était super, merci !
F	Beaucoup. Je te les montrerai si tu veux.
G	On n'avait pas assez d'argent.
H	On a fait quelque chose d'autre. On est allé en Islande. Tu connais ?

Clés : 1H, 2A, 3F, 4B, 5D

Révision de la version prétestée (réunion 3)

Aucune tâche n'a nécessité de modification lors de cette troisième réunion. Les statistiques ont indiqué que la tâche était au bon niveau de difficulté (cf. annexe VII pour la façon de comprendre ces statistiques). La cible moyenne du niveau de difficulté est de -2.09 pour KET et cette tâche a une difficulté moyenne de -2.31. Les items 1 à 5 sont d'un niveau de difficulté acceptable, compris entre -3.19 et -0.99.

	Difficulté de l'item (logits)
1	-2.72
2	-2.90
3	-2.86
4	-1.92
5	-1.13
Moyenne	-2.31

Eliminer les éventuelles doubles clés a également été abordé dans cette réunion. Le tableau ci-dessous montre la ventilation des réponses des candidats avec l'analyse statistique Classique. Exemple : pour l'item 2, 20 % du groupe faible a choisi **F** et pour l'item 4, 50 % du groupe faible a fait le choix **D**. Ces choix ont fait l'objet d'un nouvel examen pour voir s'ils pouvaient être des réponses possibles pour les items 2 et 4. Par exemple, **F** ne peut pas être la clé de l'item 2 parce que « les », dans « Je te les montrerai » ne se réfère à rien. **D** est exclu comme clé pour l'item 4 à cause de « Oui ».

Statistiques classiques					Autres statistiques					
N° de l'item	Rang de l'item	Proportion de réponses correctes	Indice de discrimi-nation	Point bisérial	Choix	Proportion totale	Bonnes réponses		Point bisérial	Clé
							Groupe faible	Groupe fort		
1	1-13	.73	.41	.40	A	.07	.15	.00	-.24	
					B	.07	.15	.00	-.27	
					C	.09	.11	.02	-.13	
					D	.01	.00	.00	-.04	
					E	.01	.00	.00	-.06	
					F	.01	.00	.00	.05	
					G	.02	.02	.00	-.04	
					H	.73	.57	.98	.40	*
					Autre	.00	.00	.00		
2	1-14	.76	.41	.36	A	.76	.57	.98	.36	*
					B	.04	.07	.00	-.08	
					C	.05	.07	.00	-.11	
					D	.03	.04	.00	-.06	
					E	.03	.04	.00	-.15	
					F	.07	.20	.02	-.30	
					G	.00	.00	.00		
					H	.02	.02	.00	-.04	
					Autre	.00	.00	.00		
3	1-15	.76	.41	.38	A	.02	.00	.00	.00	
					B	.03	.00	.00	-.02	
					C	.07	.15	.04	-.15	
					D	.03	.09	.02	-.16	
					E	.04	.11	.00	-.21	
					F	.76	.50	.91	38	*
					G	.04	.11	.02	-.18	
					H	.01	.02	.00	-.15	
					Autre	.01	.00	.00	.08	

4	1-16	.58	.56	45	A	.01	.00	.00	.01	
					B	.58	.28	.84	.45	*
					C	.02	.00	.04	.12	
					D	.29	**.50**	.07	-.37	
					E	.01	.02	.00	-.04	
					F	.00	.00	.00		
					G	.10	.20	.04	-.22	
					H	.00	.00	.00		
					Autre	.00	.00	.00		
5	1-17	.41	.60	.50	A	.02	.02	.00	-.01	
					B	.07	.09	.07	-.06	
					C	.10	.07	.07	-.05	
					D	.41	.13	.73	.50	
					E	.17	.35	.02	-.33	
					F	.06	.09	.07	-.04	
					G	.07	.11	.04	-.06	
					H	.09	.15	.00	-.22	
					Autre	.00	.00	.00		

Version 4 – version définitive (identique à la version 3)

Annexe VI - Recueil des données du prétest et de l'expérimentation

Cette annexe comprend des questions que l'on peut poser après le pré-test ou l'expérimentation (voir partie 3.4.2)

Retour d'information des surveillants- Tous les éléments

Commentaires attendus sur les points suivants :

1. Contenu : Etendue et types de questions/textes/tâches, etc. ;
2. Niveau : Difficulté, par exemple, linguistique/cognitive des différentes parties/tâches ;
3. Pré-test portant sur la compréhension orale seulement : clarté/vitesse du débit/accent des locuteurs,etc.
4. Candidature : âge des étudiants qui ont passé le pré-test ;
5. Autres commentaires

Retour d'information des candidats – test de compréhension écrite

1. Le temps imparti était-il suffisant pour faire le travail demandé ? (Combien de temps supplémentaire aurait été nécessaire ?)
2. Avez-vous eu des problèmes de compréhension de vocabulaire ? (Veuillez relever les mots/expressions qui ont posé problème)
3. Avez-vous suivi sans problème le fil des idées et l'argumentation des auteurs des textes ? (Facilement/Avec difficilement/Très difficilement)
4. Le thème proposé vous était-il familier ? (Très familier/Assez familier/Pas très familier/Pas du tout familier)
5. Quand pensez-vous passer le test réel (si vous avez l'intention de le passer)
6. Avez-vous d'autres commentaires à faire ?

Retour d'information des correcteurs – test de production écrite

Les composantes de la tâche : L'essentiel

1. La tâche a-t-elle été comprise ?
2. Le rôle du rédacteur a-t-il été clairement identifié ?
3. Le lecteur cible a-t-il été clairement identifié ?
4. Existe-t-il des biais culturels ? La tâche favorise-elle un candidat en fonction de son âge, sa formation…
5. Est-il nécessaire de reformuler la question ? Dans ce cas, quelles sont vos propositions ?

Les composantes de la tâche: Langue

6. La question a-t-elle été comprise par des candidats de niveau B2 ?
7. La formulation n'a-t-elle pas créé des malentendus ?
8. Les candidats ont-ils choisi le registre de langue adéquat ?
9. Est-il nécessaire de reformuler la question ? Que suggérez-vous ?

La production des candidats (output) : Contenu

10. Le type de tâche a-t-il été interprété de façon adéquate ?
11. Certains éléments du contenu ont-ils été mal compris/omis ? Donnez des exemples.
12. La longueur demandée était-elle adéquate ?

La production des candidats : Etendue/ton

13. Des éléments de la question ont-ils été omis dans la production ? Préciser.
14. Quel est le registre de langue utilisé par le candidat (formel/informel)

La production des candidats : Niveau

15. La question a-t-elle donné assez d'opportunité à un candidat C1 de montrer ses capacités ?

La grille de notation

16. Comment peut-on améliorer la grille de notation ? Que proposez-vous ?

Impression générale

17. Donnez votre impression générale sur la question.

Annexe VII – Utilisation des analyses statistiques dans le cycle d'élaboration de tests

Le recueil et l'analyse des données d'un test qui suppose une planification et des ressources a pour effet d'augmenter la qualité d'un test et l'interprétation des résultats. Il est de toute façon indispensable d'enregistrer des informations sur les candidats qui ont passé le test, leur note et le niveau qu'ils ont obtenu. Des analyses statistiques simples peuvent être utilisées à cet effet. Se référer exemple à Carr (2008).

Des données plus précises sur la performance d'un candidat peuvent montrer le bon fonctionnement des items et sur quels éléments l'effort de vérification doit porter. Il est possible de mener à bien la plupart des analyses décrites ci-dessous avec des logiciels d'un usage très simple. Ces outils peuvent être utilisés avec un nombre restreint de candidats (par ex 50).

Le recueil de données supplémentaires comprend :

- ✓ des données sur le niveau de la tâche :il s'agit de la note obtenue par un candidat pour chaque tâche et non pas simplement la note globale,
- ✓ des données de réponse à l'item : il s'agit de la réponse d'un candidat à chaque item du test,
- ✓ des données démographiques sur les candidats :âge, genre, 1ère langue utilisée, etc.

Les données

La plupart des logiciels d'analyse classique utilisent des données qui se présentent plus ou moins sous la forme de la figure 18. On peut saisir les données avec n'importe quelle application de processeur de mots, mais il faut alors :

- ✓ utiliser une police à espacement fixe tel que Courrier,
- ✓ ne pas utiliser de tabulations,
- ✓ sauvegarder le dossier au format d'un texte simple (txt)

Figure 18*: Exemple typique de présentation des données de réponse à l'item.*

Dans la figure 18 :

- ✓ chaque rangée comprend les réponses d'une seule personne,
- ✓ la première colonne indique l'identité de la personne (elle peut comporter des données démographiques),
- ✓ chaque colonne comprend les réponses à un item du test.

L'exemple ci-dessus porte sur un test d'items à choix multiple dans lequel les options (a-h) choisies par chaque personne sont saisies.

Il faudra fournir au logiciel des informations supplémentaires, comme par exemple l l'option correcte pour chaque item.

La théorie classique des tests

La théorie classique des tests est utilisée :

- ✓ pour l'analyse des données de prétest qui permettent ensuite de sélectionner et de vérifier les tâches utilisées dans les tests réels,
- ✓ pour l'analyse des données issues des tests réels.

Le résultat est un ensemble d'analyses statistiques sur la performance des items et de la totalité du test. En particulier :

Analyses décrivant la performance de chaque item :

- ✓ facilité et difficulté d'un item pour un groupe de candidats,
- ✓ indice de discrimination des items entre un candidat fort et un candidat faible,
- ✓ bon fonctionnement de la clé et de chaque distracteur.

Synthèse des analyses sur la totalité du test ou par partie, comprenant :

- ✓ le nombre de candidats,
- ✓ la déviation moyenne ou l'écart type des résultats,
- ✓ l'indice de fidélité.

Nous proposons ci-dessous quelques indications considérées comme des valeurs acceptables pour certaines de ces analyses. Elles ne doivent pas être considérées comme des règles absolues car en pratique les valeurs généralement observées dépendent du contexte. Les analyses statistiques classiques ont plus de poids quand elles comportent:

- ✓ un nombre plus important d'items dans un test,
- ✓ plus de candidats se présentant au test,
- ✓ un éventail plus large de compétences dans le groupe qui passe le test.

Et réciproquement, elles ont moins de poids quand elles portent sur peu d'items ou de candidats ou un éventail peu large de compétences.

La figure 19 montre des exemples d'analyses statistiques d'items à utilisant des logiciels d'analyses d'items MicroCat (voir les logiciels pour les analyses statistiques ci-dessous). Il s'agit là des analyses de trois items.

Statistiques classiques					Autres statistiques					
N° de l'item	Rang de l'item	Proportion de réponses correctes	Indice de discrimi-nation	Point bisérial	Choix	Proportion totale	Bonnes réponses		Point bisérial	Clé
							Groupe faible	Groupe fort		
1	1-1	.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		
2	1-2	.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		
3	1-3	.93	.19	.39	A	.93	.81	.00	.39	*
					B	.07	.18	.00	-.39	
					Autre	.01	.00	.00	-.03	

Figure 19 ***Exemple de statistiques classiques (Analyse d'items MicroCAT)***

Facilité

L'indice de facilité est la proportion de réponses correctes (proportion de réponses correctes dans la figure 9). Il montre la facilité de l'item en question pour ce groupe de candidats. La valeur se situe entre 0 et 1, un chiffre élevé correspondant à un item facile. La figure 19 montre que l'item 1 est le plus difficile et l'item 3 le plus facile.

L'indice de facilité est la première donnée statistique à consulter, car si le chiffre est trop élevé ou trop bas (par exemple non inclus dans l'éventail 0.25 -0.80%), cela signifie que l'estimation des autres données statistiques n'est pas correcte et que les informations sur ce groupe de candidats ne sont pas fiables.

S'il représente la population du test réel, on en conclura que l'item est tout simplement trop facile ou trop difficile. Si nous ne sommes pas sûrs du niveau des candidats, il se peut alors que l'item soit bon mais que le groupe ne soit pas au bon niveau. La conclusion à tirer est qu'il faut toujours faire passer le prétest à des candidats qui ont en gros le même niveau que celui des candidats qui passeront le test réel.

Discrimination

Les bons items doivent pouvoir distinguer un candidat faible d'un candidat fort. La théorie classique des tests propose deux indices : l'indice de DISCRIMINATION et le point bisérial de CORRELATION (Disc.Index et Point Biser dans la figure 19).

L'indice de discrimination est une simple donnée statistique : c'est la différence entre la proportion de réponses correctes obtenues par les candidats ayant les meilleurs résultats et celle obtenue par les candidats ayant les moins bons résultats (en général le tiers supérieur et inférieur des candidats). Les données de la figure 19 figurent dans les colonnes « bas (low) et haut (high) ».Pour l'item 1, la différence entre le groupe fort et faible est de 0.66-0.13. C'est la valeur de l'indice de discrimination (dans le cadre de l'erreur due à l'arrondissement).

Un item très discriminant a un indice de discrimination proche de +1, indiquant que les candidats les plus forts répondent correctement à l'item alors que les plus faibles se trompent.

Si l'indice de facilité est très élevé ou très bas, les groupes faibles et forts auront de bons résultats (ou des résultats mauvais). L'indice sous- estimera alors la discrimination. L'Item 3 en est une illustration : 1.00-0.81 = 0.19, c'est-à-dire une valeur basse.

Le point bisérial suppose un calcul plus complexe que l'indice de discrimination et est plus robuste que l'indice de facilité. Il s'agit d'une corrélation entre les résultats des candidats à un item (1 ou 0) et à la totalité du test.

On considère qu'en général les items qui ont une corrélation de point bisérial supérieure à 0.30 sont acceptables. Un point bisérial négatif signifie que les candidats forts sont susceptibles de ne pas répondre correctement à l'item. Dans ce cas, soit un des distracteurs est la réponse correcte, soit la clé est fausse.

Analyse des distracteurs

Les distracteurs sont les options qui ne sont pas correctes dans un item à choix multiple. On s'attend à ce que des candidats faibles choisissent un distracteur alors que les forts choisiront la clé (l'option correcte est indiquée par +).

L'analyse des distracteurs montre la proportion des candidats qui ont choisi chaque distracteur (prop..total dans la figure 19). L'item 1 a un indice de facilité bas pour l'option correcte B (0.38). Le distracteur B attire plus de réponses (indice de facilité=0.49). Le distracteur A n'attire aucun candidat, ce qui signifie que ce n'est pas un bon distracteur. Cependant, l'item fonctionne bien dans l'ensemble, avec un bon indice de discrimination. Il n'y a donc aucune raison de le changer. En fait, il est difficile de trouver trois distracteurs qui fonctionnent bien.

Les analyses de la figure 19 montrent aussi la proportion haute et basse dans les choix et le point bisérial pour chaque option. Le point bisérial d'un bon item sera positif pour la clé et négatif pour chaque distracteur.

Fidélité des résultats

Il y a plusieurs façons d'évaluer la fidélité et différentes formules existent pour ce faire. Chaque méthode a ses avantages. La méthode de bissection (split-half) consiste à diviser le test en deux parties égales et à comparer le résultat du candidat dans

ces deux parties. Si on utilise cette méthode, il est important que les deux parties soient aussi équivalentes que possible: équivalence du construit dans toute son étendue , équivalence de difficulté, etc.

D'autres méthodes consistent à mesurer la consistance interne du test. Elles fonctionnent bien à condition que le type d'items et le contenu soient similaires. Par contre, si les items sont hétérogènes, la fidélité sera sous-estimée.

Pour les analyses classiques:

Nombre de candidats : 50 à 80 (Jones, Smith et Talley 2006:495)

Pour plus ample information : Verhelst (2004a,b); Bachman (2004)

L'analyse de Rasch

L'analyse de Rasch est la forme la plus simple et pratique de la THEORIE DE REPONSE A L'ITEM ou TRI. Cette analyse permet d'une part de mieux comprendre ce qu'est la difficulté de l'item que par l'analyse classique et a d'autre part des applications supplémentaires telles que les façons de relier les tests entre eux.

Avec l'analyse de Rasch :

- la différence exacte de difficulté entre deux items est claire car les items sont placés sur une ECHELLE D'INTERVALLE mesurée en logits (appelée aussi échelle « logit »),
- la différence entre les items, les candidats, les résultats au test et les points de césure peut être interprétée de la même façon dans la mesure où toutes ces données sont sur une même échelle,
- la difficulté de l'item peut être interprétée indépendamment des capacités du candidat (alors qu'avec l'analyse classique, selon le niveau du groupe de candidats, un item peut paraître plus facile ou plus difficile).

L'analyse de Rasch est donc très utile pour contrôler et maintenir des standards d'une session à l'autre. Cependant, si on veut utiliser Rasch dans ce but, les items des différents tests doivent être reliés entre eux. Par exemple, deux tests peuvent être reliés entre eux de différentes façons :

- les mêmes items sont utilisés dans les deux tests,
- un groupe D'ITEMS ANCRES est utilisé dans les deux tests,
- quelques items ou tous les items sont CALIBRES avant d'être utilisés dans les tests réels (voir 3.4.2 le prétest),
- certains candidats passent les deux tests.

Quand les données des deux tests sont analysées, le lien créé permet d'avoir un seul cadre de référence pour tous les items, tous les candidats, etc. et des valeurs de difficulté calibrées sont attribuées aux items. D'autres tests peuvent être ajoutés au cadre de référence en utilisant la même procédure.

Les standards peuvent être contrôlés en comparant la position respective des éléments importants :

- Les items sont-ils de la même difficulté dans tous les tests ?
- Les candidats ont-ils les mêmes capacités ?
- Les points de césure (mesurés en logits) coïncident-ils avec les SCORES BRUTS (eux aussi mesurés en logits) dans tous les tests ?

Les standards peuvent être maintenus si les points de césure sont chaque fois décidés en tenant compte des mêmes valeurs de difficulté.

Il est certes plus facile de maintenir des standards et la qualité d'un test s'il est élaboré avec des items calibrés. Toute la difficulté d'un test peut être décrite par sa difficulté moyenne et son ETENDUE. La difficulté d'un test peut être contrôlée en sélectionnant un groupe d'items qui correspond à l'étendue cible et la moyenne cible.

Quand vous commencez à calibrer des items, les valeurs de difficulté ne signifient pas grand-chose. Mais avec le temps, on finit par bien se représenter les capacités réelles d'un candidat en regardant les points sur l'échelle de capacités. Une autre possibilité est aussi d'émettre un jugement subjectif sur des items (Je pense qu'un apprenant B1 aurait 60% de chance de répondre correctement à cet item) afin de donner un sens aux difficultés des items. C'est ainsi qu'on se familiarise avec les chiffres de l'échelle de capacités et qu'ils prennent un sens.

Nombre de candidats : 50 à 80 (Jones, Smith et Talley 2006:495)

Pour plus ample information : Verhelst (2004d); Fox (2007)

Analyses statistiques pour la notation et le classement

La correction humaine

Il est important de s'assurer de la qualité du travail des correcteurs. Si le travail n'est pas bien fait , il faut alors exiger qu'ils suivent une nouvelle formation- (voir partie 5.1).S'il y a peu de candidats au test, il est toujours possible de vérifier la note donnée à chaque item à chaque candidat. Par contre, quand les candidats sont plus nombreux, un échantillon (peut-être 10%) des épreuves corrigées par un correcteur peut être relevé et un taux établi. Un taux d'erreur est le nombre d'erreurs commises par un correcteur divisé par le nombre d'items corrigés. Si cet échantillon est représentatif de tout son travail, le taux d'erreur sera sans doute le même pour la totalité de sa correction.

Il est préférable que l'échantillon soit recueilli de façon aléatoire si l'on veut qu'il soit représentatif. Pour être sûr que la sélection de l'échantillon est faite de façon aléatoire, il faut savoir comment le correcteur travaille. Le choix aléatoire ne signifie pas les 10% de n'importe quelles épreuves corrigées dans la mesure où cet échantillon peut comprendre uniquement les dernières corrections effectuées par le correcteur, et qui ont choisies parce qu'elles étaient plus accessibles. Dans ce cas, le taux d'erreur sous-estime tout le temps passé par le correcteur avant le recueil de l'échantillon en oubliant qu'il a ensuite amélioré ses performances.

L'évaluation

La performance des évaluateurs peut être évaluée statistiquement de façon très simple en calculant la moyenne de leurs évaluations et L'ECART TYPE(une mesure de la dispersion de leurs évaluations, de la plus basse à la plus haute). Les évaluateurs peuvent être comparés les uns aux autres et une recherche peut être faite sur les évaluations d'un correcteur dont la mesure diffèrerait de celles des autres. Cela suppose que les épreuves du test soient distribuées de façon aléatoire. Si ce n'est pas le cas, un évaluateur peut très bien évaluer des candidats qui sont d'habitude meilleurs ou moins bons que la moyenne. Dans ce cas la moyenne risque d'être plus élevée ou moins élevée que les autres évaluateurs, mais cela ne remet pas en question la compétence de l'évaluateur.

Si certaines tâches peuvent être évaluées par deux évaluateurs, la fidélité de ces notes peut être évaluée. On peut le faire par exemple avec Excel en utilisant la fonction de corrélation de Pearson. Les données peuvent être présentées de la façon suivante:

	Evaluateur 1	Evaluateur 2
Candidat 1	5	4
Candidat 2	3	4
Candidat 3	4	5
…	…	…

Le coefficient de corrélation sera entre -1 et 1. Dans la plupart des cas, un nombre inférieur à 0,8 est suspect et demande vérification. Car il suppose que l'évaluateur n'a pas évalué de façon cohérente.

Une estimation de fidélité comme celle produite par l'Alpha de MicroCat (se référer aux logiciels pour les analyses statistiques mentionnés ci-dessous) peut être calculée pour tout le groupe d'évaluateurs. Les données peuvent être présentées comme dans la figure 18, avec quelques modifications, chaque rangée pouvant indiquer les performances d'un candidat à une tâche; et les colonnes les notes des évaluateurs.

Mesure établie avec un modèle de Rasch multi facettes

Une manière plus sophistiquée de porter un jugement sur la performance des évaluateurs est d'utiliser la mesure établie avec un modèle de Rasch multi facettes : (many-facet rasch measurement - MFRM).C'est une variante de l'analyse de

Rasch. La MFRM peut être menée en utilisant le logiciel Facets (Linacre 2009). L'analyse mesure, comme avec l'analyse de Rasch, la difficulté des tâches et les capacités des candidats, mais elle peut aussi évaluer la sévérité ou le laxisme des évaluateurs. De plus, les notes attribuées sont plus précises dans la mesure où les effets dus à la sévérité ou au laxisme sont supprimés.

Quand on utilise la mesure établie avec un modèle de Rasch multi facettes, il est très important de s'assurer que les données comprennent les liens entre les évaluateurs, les candidats, les tâches et les autres facettes mesurées. Il est par exemple nécessaire que des candidats accomplissent plus qu'une tâche afin d'établir un lien entre les tâches. Si les données sont groupés sans lien entre elles, la mesure établie avec le modèle de Rasch multi facettes ne pourra pas fournir des estimations pour tous les éléments.

Pour la mesure établie avec un modèle de Rasch multi facettes :

Nombre minimum de performances : 30 pour chaque tâche devant être évaluée (Linacre 2009)

Nombre minimum d'évaluations par évaluateur : 30 (Linacre 2009)

Pour plus ample information : Eckes (2009).

Validation du construit

Vérification de la structure du test

L'analyse factorielle ou les modèles d'équations structurelles permettent de vérifier si les items appliquent le construit prévu. La structure du test doit refléter le modèle d'usage de la langue qui a été choisi. (voir partie 1.1). L'analyse factorielle est très utile lors des étapes d'élaboration du test, car elle permet de vérifier que le test ou les spécifications fonctionnent comme prévu.

Pour les analyses factorielles:

Nombre minimum de candidats : 200 (Jones, Smith and Talley 2006:495)

Pour plus ample information : Verhelst (2004c)

La détection des biais d' items

On détecte des biais d'items quand des items favorisent ou défavorisent certains groupes de candidats de capacités équivalentes. Par exemple, un item peut être plus facile pour une candidate que pour un candidat alors qu'ils ont les mêmes capacités. Cela crée une injustice dans la mesure où le but du test n'est pas de mesurer des différences dans le domaine du genre mais dans celui des capacités langagières (voir partie 1.4).

Il faut cependant être prudent quand il s'agit de faire le diagnostic des biais dans la mesure où toutes les différences ne sont pas injustifiées. Des différences entre la langue 1 de deux groupes d'apprenants de même compétence peuvent les amener à trouver qu'un item de la langue cible est plus difficile pour un groupe que pour un autre. Comme il s'agit de mesurer les performances langagières , il faut considérer que cela fait partie de la nature de la performance dans la langue cible et ne pas le voir comme un problème de mesure de cette performance.

Une façon de minimiser ce biais est d'utiliser la méthodologie du Fonctionnement différentiel des items (FDI) (Differential Item Functioning DIF) pour détecter d'éventuels biais et effectuer les vérifications ultérieurement. Cela suppose que l'on compare les réponses des groupes de candidats de capacités identiques. Si par exemple le test est destiné aux adultes d'âges différents, on peut comparer les performances des plus jeunes et des plus vieux ayant sensiblement les mêmes capacités. Les analyses du type de l4TRI (théorie de réponse à l'item) conviennent tout à fait.

Pour le Fonctionnement différentiel des items (Differential Item Functioning - DIF) avec l'analyse de Rasch

Nombre minimum de candidats : 500 dont au minimum 100 par groupe(Jones, Smith and Talley 2006:495)

Pour plus ample information : Camilli et Shepard (1994); Clauser et Mazor (1998)Verhelst (2004c)

La vérification de l'échantillon de candidats

Toute analyse ou recherche utilisant des données de test, doit être menée de telle sorte que ces données soient représentatives du groupe cible de candidats (la population). On peut recueillir des informations sur les candidats de façon régulière et vérifier si les analyses sont menées sur un échantillon totalement représentatif de candidats.

On peut recueillir des données sur les candidats à chaque passation d'un test (voir partie 4). Ces données peuvent être comparées en utilisant tout simplement des pourcentages, par exemple pour comparer le nombre de femmes et d'hommes dans deux échantillons différents.

Une analyse plus sophistiquée permettra d'établir si les différences entre deux échantillons sont dues au hasard. On peut utiliser un test Khi-carré de cette façon. Les résultats d'une analyse doivent ensuite être vérifiés sur le plan qualitatif pour vérifier si les différences entrainent des différences significatives de performance du candidat.

Les outils pour des analyses statistiques

Un certain nombre de logiciels sont disponibles à des fins d'analyses. Il est possible de mener très facilement à bien des mesures de calcul en utilisant Microsoft Excel, ou un autre programme de feuilles de calcul. Une liste des fournisseurs spécialisés est donnée ci-dessous dans l'ordre alphabétique. Ils peuvent fournir des logiciels pour différents types d'analyses. Des versions de démonstration sont parfois disponibles.

Assessment Systems	http://www.assess.com/softwarebooks.php
Curtin University of Technology	http://lertap.curtin.edu.au/index.htm
RUMM Laboratory	http://www.rummlab.com.au/
Winsteps	http://www.winsteps.com/index.htm

D'autres outils gratuits sont disponibles à des fins spécifiques :

William Bonk, University of Colorado http://psych.colorado.edu/~bonk/

Del Siegle, University of Connecticut: http://www.gifted.uconn.edu/siegle/research/Instrument%20Reliability%20and%20Validity/Reliability/reliabilitycalculator2.xls

Annexe VIII – 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.

Approche de type actionnel

Façon de considérer l'usager et l'apprenant d'une langue comme des acteurs sociaux ayant à accomplir des tâches dans des circonstances et un environnement donnés à l'intérieur d'un domaine d'action particulier (définition du CERL)

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.

Argumentaire pour les utilisations de l'examen

La partie de l'argument de validité qui explique comment les résultats doivent être interprétés pour un usage spécifique.

Argument interprétatif

Voir » Argumentaire pour les utilisations de l'examen »

Argument de validité

Ensemble de propositions et de preuves qui ont pour but de soutenir la validité des interprétations des résultats du test.

Authenticité

Degré de ressemblance des tâches avec celles de la vie quotidienne. Par exemple, la prise de notes dans un test mesurant la compétence dans le domaine éducationnel plutôt que la simple écoute d'un document. Voir aussi Utilité d'un test.

Banque d'items

Gestion des items qui permet de stocker des informations afin de pouvoir élaborer des tests aux contenu et difficultés connus.

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 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 de questions.

Classement

Conversion des notes obtenues en niveaux.

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.

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..

Consigne

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

Construit

Capacité hypothétique ou trait mental qui ne peut pas être observé ou mesuré, comme par exemple dans l'évaluation, la capacité de compréhension orale.

Correcteur

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

Corrélation

Relation entre deux ou plusieurs mesures, en tenant compte du fait qu'elles peuvent varier de la même façon. Si, par exemple, les résultats de candidats sont les mêmes dans des tests différents, il existe une corrélation positive entre les deux ensembles de résultats.

Déclencheur

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

Définition des points de césure

Processus de définition des points de césure dans un test (par exemple la limite entre l'échec/le succès et par conséquent de la définition des résultats du test).

Descripteur

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

Ecart type

L'écart type est la mesure de la dispersion des résultats à un test (ou la distribution d'autres données). Si la distribution des résultats est normale, 68% d'entre eux sont compris dans la 1 ET de la moyenne et 95% dans la 2 ET. Plus l'écart type est élevé et plus il est éloigné de la majorité des données.

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. Voir l'annexe VII pour plus de renseignements.

Domaine d'usage de la langue

Vastes domaines de la vie sociale, telle que l'éducation ou la vie personnelle que l'on peut définir pour choisir le contenu et l'accent à mettre dans les activités langagières dans les examens.

Double notation

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

Echelle

Ensemble de nombres ou de catégories destinés à mesurer quelque chose. On distingue quatre sortes d'échelles: échelle nominale, ordinale, d'intervalle et de rapport.

Echelle 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.

Echelle de mesure

Une échelle de mesure est une échelle composée de nombres qui mesurent la différence entre les candidats, les items, les points de césure, etc. sur le construit du test. On élabore une échelle de mesure en appliquant des techniques statistiques à des réponses des candidats à des items. (cf. annexe VII). L'échelle de mesure fournit bien plus d'informations que des résultats bruts dans la mesure où elle ne montre pas seulement quels candidats sont meilleurs que tels autres mais aussi quel est de combien ils sont meilleurs. On utilise parfois les termes d'échelles nominales et ordinales pour désigner des échelles de mesure mais ces définitions n'ont pas été retenues dans ce Manuel

Echelle d'intervalle

Echelle de mesure dans laquelle la distance entre deux unités adjacentes de mesure est la même, mais dans laquelle il n'y a pas de points zéro absolus.

Elaboration 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 d'items.

Elaborateur de test

Personne impliquée dans l'élaboration d'un test nouveau

Enjeux

Degré d'importance que peut avoir les résultats d'un test sur l'avenir d'un candidat. On parle généralement de test à fort ou à faible enjeu, un test à fort enjeu ayant un impact plus grand.

Erreur standard de mesure

Dans la théorie de la note vraie, l'erreur standard de mesure (ES) indique l'imprécision de la mesure . Si, par exemple, l'erreur de mesure est 2, un candidat ayant obtenu une note 15 aura une note entre 13 et 17 (avec 68% de certitude). Une erreur plus petite aura pour conséquence une note plus précise.

Etendue

L'étendue est une mesure simple de la dispersion : c'est la différence entre le nombre le plus élevé et le plus bas dans un groupe.

Expérimentation

Etape de l'élaboration des tâches d'un test servant à vérifier que le test fonctionne de la façon attendue. Souvent utilisée dans le cas de tâches à notation subjective telles que la composition ou l'essai et administrée à une population limitée.

Examen réel (en grandeur nature)

Un test prêt à être utilisé et qui, pour cette raison, doit être stocké en toute sécurité.

Evaluateur.

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.

Faisabilité

Degré d'élaboration d'un test répondant à des exigences d'ordre pratique. Voir aussi Utilité d'un test.

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, transcrites sur une échelle de 0 à 1. Egalement exprimé sous forme de pourcentage. Aussi considéré comme la proportion correcte, l'indice de facilité ou la valeur-p.

Input

Composantes de la tâche fournies au candidat afin qu'il puisse produire une réponse adéquate. Par exemple, dans un test de compréhension orale, il peut s'agir d'un test enregistré et des items auxquels il doit répondre par écrit.

Interactivité

Degré auquel des items et des tâches font appel à des processus et des stratégies cognitifs s'approchant de ceux de la vie quotidienne. Voir aussi Utilité du test.

Item

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.

Modèle de crédit partiel

Un item dont la réponse n'est ni totalement vraie ni totalement fausse. Par exemple, les notes attribuées à un item peuvent être 0,1,2,3 selon le degré d'exactitude de la réponse.

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 ancre

Item inclus dans un ou plusieurs tests. Les caractéristiques de ces items ancres sont connues. Ils forment une partie de la nouvelle version d'un test. L'objectif est de fournir des informations sur le test et les candidats qui l'ont passé afin, par exemple, de calibrer un nouveau test sur l'échelle de mesure.

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.

Item dichotomique

Item qui est noté vrai ou faux. Les items sous forme de questions à choix multiple(QCM), vrai/faux, questions à réponses courtes (QRC) sont des items dichotomiques.

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.

Logit

Le logit est l'unité de mesure utilisée dans les analyses du modèle de Rasch (TRI) et le modèle multi facet de Rasch (MFRM).

Mise en relation

La mise en relation est une procédure qui « traduit » les résultats d'un test pour qu'ils puissent être compris en relation avec les résultats d'un autre test. Cette procédure permet de compenser les différences de difficulté d'un test ou de capacité des candidats.

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.

Modèle concordant

Quand un modèle (comme le modèle de Rasch) est utilisé pour des analyses statistiques, il est important de voir jusqu'à quel point les données et le modèle sont en concordance. Un modèle représente un ce que des données devraient être dans l'idéal et on ne peut donc s'attendre à une concordance parfaite. Par contre un degré élevé de discordance signifie que les conclusions tirées des données sont fausses.

Mesure établie avec un modèle de Rasch multi facettes

La mesure établie avec un modèle de Rasch multi facettes est un prolongement du modèle de base de Rasch. La difficulté de l'item ainsi que les capacités du candidat sont répartis en facettes , ce qui permet d'utiliser les données relatives à ces facettes pour expliquer les résultats données aux candidats. Par exemple, la sévérité d'un évaluateur peut permettre d'expliquer les résultats de candidats à la production écrite. Dans ce cas, on explique que les résultats sont dus aux capacités du candidat, à la difficulté de la tâche et à la sévérité de l'évaluateur. Il est alors possible de supprimer le facteur sévérité de l'évaluateur des résultats réels attribués aux candidats.

Modèle d'utilisation de la langue

Description des capacités langagières et des compétences nécessaires à l'utilisation de la langue et de la relation entre elles. Un modèle est la composante de base de la conception.

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 notes.

Niveau

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.

Notation

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.

Notation objective

Items qui peuvent être notés en appliquant un barème sans l'apport de point de vue ou de jugement subjectif d'expert.

Notation subjective

Items où le point de vue ou le jugement subjectif d'expert intervient dans la notation.

Notation standardisée (mécanique)

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

Parties prenantes / parties concernées

Personnes ou organisations parties prenantes du test. Par exemple, les candidats, les institutions scolaires, les parents, les employeurs, le gouvernement, les salariés du fournisseur de test.

Pilotage

Expérimentation du matériel sur une petite échelle en demandant par exemple aux collègues de répondre aux items et de faire des commentaires.

Pondération; syn.: **coefficient**

Action d'assigner un nombre plus grand 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.

Prétest ; 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.

Question

Terme parfois utilisé pour désigner une tâche ou un item.

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 de production écrite.

Révision; syn.: **contrôle**

Une étape au cours du cycle d'élaboration du test pendant laquelle les élaborateurs de tests évaluent le travail demandé aux rédacteurs d'items, et décident de garder ou rejeter les items produits selon qu'ils répondent ou non aux spécifications du test.

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 la capacité d'un candidat à 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.

Score brut

Résultat du test qui n'a pas donné lieu à des analyses statistiques supposant des transformations, des pondérations ou des reclassements.

Spécification

Description des caractéristiques d'un examen indiquant ce qui est testé, de quelle façon, ainsi que le nombre et la longueur des épreuves, les types d'items utilisés, etc.

Surveillant

Personne qui est responsable de la bonne passation de l'examen dans une salle d'examen.

Tâche

Ce qu'un candidat doit faire pour accomplir une partie du test et qui suppose plus de complexité qu'une réponse à un seul item discret. Le terme concerne en général des performances de production orale ou écrite ou un ensemble d'items liés entre eux comme par exemple un texte accompagné de questions à choix multiple auxquelles on peut répondre en suivant une seule consigne.

Tâche d'appariement :

Type de tâche consistant à comparer des éléments de deux lites distinctes. Un type de test d'appariement consiste à choisir la phrase correcte pour compléter chacune des phrases incomplètes. Un autre exemple est celui qui est utilisé dans les tests de compréhension écrite et qui consiste à choisir dans une liste des vacances ou un livre convenant à une personne aux caractéristiques correspondantes.

Théorie de l'item-réponse TIR

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.

Trait

Caractéristiques physiques ou psychiques d'une personne (comme les capacités langagières) ou l'échelle de mesure qui permet de les décrire. Voir aussi construit.

Utilité d'un test

Le concept d'utilité (Bachman et Palmer 1996) renvoie à l'idée qu'un test est d'autant plus utile que la relation entre la validité, la fidélité, l'authenticité, l'interactivité, l'impact et la faisabilité est optimale.

Validation

Le processus qui consiste à établir la validité des interprétations des résultats proposés par le fournisseur de test.

Validité

Degré auquel les interprétations des résultats d'un test permettent de tirer des conclusions appropriées, significatives et utiles, en relation avec l'objet du test.

Remerciements

Ce Manuel est une version actualisée d'une version publiée par le Conseil de l'Europe en 2002 intitulée « Passation et élaboration de tests et d'examens de langue ». Ce document était lui-même une version actualisée du « Guide pour les examinateurs » conçu par ALTE pour le Conseil de l'Europe en 1996.

Le Conseil de l'Europe tient à remercier pour sa contribution :

L'association des centres évaluateurs en Europe (ALTE)

L'équipe responsable de l'édition de cette nouvelle version:

David Corkill Neil Jones Martin Nuttall

Michael Corrigan Michael Milanovic Nick Saville

Les membres du groupe à objectifs spécifiques (ALTE/CECRL) ainsi que leurs collègues ayant proposé des documents et participé à la relecture des textes :

Elena Archbold-Bacalis Sharon Ashton Andrew Balch Hugh Bateman Lyan Bekkers Nick Beresford-Knox Cris Betts Margherita Bianchi Inmaculada Borrego Jasminka Buljan Culej Cecilie Carlsen Lucy Chambers Denise Clarke María Cuquejo Emyr Davies Desislava Dimitrova Angela ffrench Colin Finnerty Anne Gallagher Jon-Simon Gartzia Annie Giannakopoulou Begona Gonzalez Rei Giuliana Grego Bolli Milena Grigorova Ines Haelbig Berit Halvorsen Marita Harmala	Martina Hulešová Nuria Jornet Marion Kavallieros Gabriele Kecker Kevin Kempe Wassilios Klein Mara Kokina Zsofia Korody Henk Kuijper Gad Lim Juvana Llorian Karen Lund Lucia Luyten Hugh Moss Tatiana Nesterova Desmond Nicholson Gitte Østergaard Nielsen Irene Papalouca Szilvia Papp Francesca Parizzi Jose Ramón Parrondo Jose Pascoal Roberto Perez Elorza Michaela Perlmann-Balme Tatiana Perova	Sibylle Plassmann Laura Puigdomenech Meilute Ramoniene Lýdia Ríhová Shelagh Rixon Martin Robinson Lorenzo Rocca Shalini Roppe Dittany Rose Angeliki Salamoura Lisbeth Salomonsen Georgio Silfer Gabriela Snaidaufová Ioana Sonea Annika Spolin Stefanie Steiner Michaela Stoffers Gunlog Sundberg Lynda Taylor Julia Todorinova Rønnaug Katharina Totland Gerald Tucker Piet van Avermaet Mart van der Zanden Juliet Wilson Beate Zeidler Ron Zeronis

Les relecteurs du Conseil de l'Europe :

Neus Figueras Johanna Panthier

Brian North Sauli Takala

L'équipe chargée de la publication :

Rachel Rudge Gary White

L'Association des organismes certificateurs en Europe (ALTE), en tant qu'Organisation internationale non-gouvernementale (INGO) ayant un statut consultatif au sein du Conseil de l'Europe, a contribué aux ressources composant la boîte à outils, y incluant le Portfolio européen des langues (PEL) d'EAQUALS/ALTE ainsi que les grilles d'analyses de contenus du CECR pour la production orale et écrite.

En accord avec la Division des politiques linguistiques du Conseil de l'Europe, ALTE tient à ce que les utilisateurs de la boîte à outils se servent efficacement du Cadre dans leur propre contexte et afin de satisfaire leurs propres objectifs.

Produit par:
Association of language testers in Europe
1 Hills Road,
Cambridge CB1 2EU
Royaume Uni
www.alte.org

Au nom du :
Conseil de l'Europe