Techniques d'évaluation de la production orale

James HILL
Nice

L'enseignement des langues vise à l'acquisition de quatre aptitudes fondamentales (ou skills) :

Compréhension orale
Compréhension écrite
Production orale
Production écrite

Je veux limiter mon intervention ce matin à une seule de ces quatre aptitudes fondamentales, la production orale, car c'est dans ce domaine que les techniques mises au point à Michigan au cours des années 50 et décrites dans le livre de Robert Lado, « Language Testing », publié en 1961, semblent les moins satisfaisantes.

Les questions à Choix Multiples utilisées pour des tests de compréhension écrite ou orale donnent de bons résultats, car il s'agit d'évaluer des aptitudes passives. Il parait dangereux d'appliquer les mêmes techniques à l'évaluation de la production écrite ou orale, notamment les « Partial Production Techniques », techniques indirectes où on juge de la capacité de distinguer auditivement entre les phonèmes ou les éléments prosodiques : les résultats de ce test sont pris comme témoin de la prononciation du candidat. On justifie cette extrapolation en disant que la prononciation correcte d'un phonème dépend de son audition ; celui qui n'entend pas une opposition vocalique ne saura pas la reproduire. C'est peut-être une condition nécessaire, mais pas suffisante. En effet, la correspondance entre ce que l'étudiant croit prononcer et ce qu'il croit entendre, entre le phonème qu'il vise et le son qu'il émet, est loin d'être parfaite. Il y a quelques années, j'ai fait une comparaison, pour un très petit groupe, où les étudiants devaient d'abord reconnaître quel phonème ou quelle accentuation ils pensaient utiliser dans chaque mot, et ensuite lire les mêmes mots. La performance globale dans les deux tests n'était pas très différente : 63 % de réponses correctes pour la catégorie des phonèmes, 59 % pour la production.

Les rangs respectifs pour les deux tests avaient aussi une certaine corrélation statistique. Néanmoins l'étudiant avec la meilleure note pour la catégorisation des phonèmes se trouvait avant-dernier à la production. En outre, l'examen détaillé des résultats montre qu'il n'y avait correspondance entre ce que l'étudiant pensait prononcer et la réalité que dans 60 % des cas.

Ces tests indirects peuvent peut-être fournir une indication globale de capacité quand les effectifs, l'éloignement des candidats, la nécessité d'une correction rapide et plus ou moins mécanisée, rendent l'examen direct de production orale impossible ; mais les résultats subissent des distorsions selon l'expérience antérieure des candidats. Ainsi celui qui a suivi un enseignement systématique de phonétique obtiendra un score supérieur à sa performance réelle, et celui qui a

l'habitude d'un accent autre que celui de l'enregistrement des magnétophones, utilisés non seulement pour diffuser des exercices de compréhension orale mais pour enregistrer les réponses des candidats, rend possible un véritable contrôle de la production orale dans des conditions semblables à celles des examens de langue écrite.

Il y a deux façons d'utiliser le magnétophone : on peut conserver la situation traditionnelle de l'interview où un interrogateur établit le contact personnel, pose les questions qui susciteront l'échantillon de la production orale du candidat, et cet échantillon sera évalué ultérieurement par le truchement de l'enregistrement. Ou si l'on dispose d'un laboratoire de langues équipé de magnétophones d'enregistrement pour chaque élève, l'interrogateur peut être remplacé par des instructions et des questions enregistrées et diffusées simultanément dans chaque cabine. La première technique a été adoptée par la « Modern Language Association », en Grande-Bretagne, dans la réforme des examens oraux de langues vivantes au G.C.E. O-Levels (Le Général Certificate of Education : O-Levels correspond approximativement dans le système français au B.E.P.C, au niveau de la troisième. Pour les détails de l'expérimentation précédant la réforme, voir H.S. OTTER A Functional Language Examination (O.U.P. 1968). La deuxième au Laboratoire de Langues, a été adoptée, entre autres, par « M.L.A. Coopérative Foreign Language Tests » et, plus récemment, par l'Association of Recognized English Language Schools dans son « Examination in Spoken English » organisé deux fois par an depuis 1967 (détails disponibles : Arles Oral Examinations, 15 Holland Park Gardens London W14). Ce dernier est très intéressant à plusieurs points de vue, notamment par sa volonté d'évaluer la capacité des candidats à utiliser l'anglais oral pour la communication dans des situations de la vie quotidienne. Nous y reviendrons. Les avantages de l'utilisation de la bande magnétique pour remplacer la « conversation » ou l'interrogation traditionnelle sont évidents au stade de l'évaluation : l'enregistrement sur bande étant un document permanent, on peut, comme pour une copie de thème ou de rédaction, établir, après l'écoute rapide d'un échantillon représentatif des bandes ou pour les cas litigieux, une double correction (c'est le cas de l'examen ARELS) ; le correcteur peut réécouter un passage douteux, revenir sur une bande pour vérifier qu'il applique toujours le barème convenu ; le correcteur ne peut juger que les données linguistiques, la personnalité sympathique ou antipathique du candidat lui est inconnue sauf par son expression linguistique (la voix évidemment peut être plus ou moins sympathique, mais son impact est bien moindre que le face à face de l'interview. Le caractère différé du processus de notation permet de faire des compensations entre les correcteurs, et de vérifier que le même style d'interview a été maintenu pour tous les candidats. L'enregistrement permet aussi à l'interrogateur de ne plus se partager entre sa tâche d'évaluation et son rôle de stimulateur de la conversation. On sera amené à pratiquer des interrogations assez structurées, sinon le correcteur sera confronté à une série de bandes trop dissemblables. Il ne saura pas les raisons pour lesquelles l'interrogateur a dirigé la conversation dans un sens ou dans un autre.

Aux tests enregistrés, on objecte souvent des arguments techniques : les appareils ne sont pas infaillibles, les manipulations successives comportent des risques d'effacement, la qualité acoustique peut nuire à l'appréciation de certains détails. La fiabilité des magnétophones actuels et l'expérience de leurs utilisateurs permette de réduire ces risques au minimum. Par contre, on ne peut nier que certaines réponses peuvent ne pas être comprises à leur pleine valeur si elles ne sont pas accompagnées du geste de la main ou du sourire. Certains candidats (et certains examinateurs... !) peuvent être inhibés par la vue d'un magnétophone qui tourne ou d'un microphone posé sur la table. C'est une réaction qui tend à disparaître chez les jeunes, en Europe Occidentale du moins ; ils manient les appareils électroniques dès le berceau. Les tests enregistrés simultanément en laboratoire présentent les mêmes avantages à la correction, et permettent en outre de donner une épreuve rigoureusement pareille pour tous, ce qui facilite considérablement l'établissement d'un barème objectif. L'enregistrement des questions permet également de varier les voix et les registres dans les parties consacrées à la compréhension orale et d'atténuer ainsi les handicaps des candidats confrontés à un examinateur qui utilise une variante de l'anglais, régionale ou stylistique, dont il n'a pas l'habitude.

Par contre, certains candidats peuvent être inhibés par l'impersonnalité de la cabine de laboratoire et du micro. Utiliser la langue orale implique communication avec un autre être humain, et ce sont justement les candidats qui sont le plus à l'aise dans la langue étrangère, pour qui elle est véritablement un outil de communication qui risquent d'en souffrir le plus. Pour d'autres, plus timides, le micro sera peut-être moins traumatisant que l'interrogateur en chair et en os.

Il ne faut pas négliger non plus une autre conséquence du remplacement de l'interrogateur humain par des questions et des exercices enregistrés : le poids de la compréhension orale sera beaucoup plus grand. Un interrogateur s'aperçoit que sa question a été mal comprise, la reformule, et permet au candidat de répondre et de marquer des points, même si le correcteur le pénalise pour l'erreur de compréhension. Par contre, la bande enregistrée défile sans possibilité de repêchage.

Techniquement, on peut laisser au candidat la liberté de revenir sur une question, mais alors les uns profiteront beaucoup de cette liberté, et produiront des réponses soignées, mais fort éloignées d'une utilisation spontanée de la langue ; les autres, probablement ceux qui sont le plus à l'aise, répondront du premier coup avec moins de précision. Le correcteur ne pourra pas toujours distinguer les deux cas. L'uniformité pour tous les candidats, l'avantage principal de l'épreuve en laboratoire, est perdue.

La tradition des tests américains veut qu'on distingue entre l'utilisation active et passive de la langue « productive and réceptive skills ». On cherche à dissocier les différents éléments à tester. C'est utile pour les tests « diagnostiques », mais quand on veut apprécier la capacité globale, le degré d'aptitude à manier la langue (et non les connaissances des règles ou du vocabulaire) tout morcellement nous éloigne d'un emploi naturel où compréhension et production sont nécessairement liées. Si

l'on ne prononce que ce qu'on entend, l'expérience montre aussi que nos élèves entendent le plus souvent comme ils prononcent ! Inversement, un test de compréhension est toujours aussi un test de production. Nous ne pouvons atteindre la compréhension du candidat qu'à travers ce qu'il en dit ou ce qu'il écrit (même une croix dans une case de Q.C.M suppose la compréhension par la lecture des choix proposés (sauf peut-être au niveau élémentaire où l'on fait choisir une image selon les instructions données c'est la compréhension des instructions qui est testée. Voix Lado, 1961 p.p. 210-14. Le A.R.E.L.S Oral Examination est intitulé « Examination in Spokes English and Comprehension ». Les différentes sections mêlent délibérément compréhension et traduction orale. Par exemple, dans la section I (Utilisation de la langue dans les échanges quotidiens) :
(a) réaction immédiate à une remarque :
« Excuse me you've dropped something », on teste à la fois la compréhension et la capacité de produire une réponse appropriée. Il en est de même dans la 2ème partie où la situation est expliquée sur la bande et le candidat doit faire une remarque appropriée :
« You're in a friend's flat and want to telephone. What do you say? » Dans la section III, un dialogue d'une vingtaine de répliques est suivi d'une dizaine de questions. Les réponses enregistrées sont notées pour la compréhension du dialogue et de la question posée. L'examen A.R.E.L.S est extrêmement intéressant à plusieurs points de vue. Malgré le format mécanisé en laboratoire, il a le souci constant de contextualiser les questions, les rendre aussi proches que possible de l'utilisation réelle d'une langue. Par exemple, pour introduire le texte à lire, la bande explique au candidat qu'il doit aider un vieillard qui voit mal et lui lire un prospectus. Un autre exercice de lecture est présenté comme une 'conversation téléphonique où le candidat répond au correspondant déjà enregistré sur la bande (l'enregistrement sert aussi de contexte à la lecture et lui impose les intonations voulues). Ce sont des détails d'apparence triviale mais qui ont de l'importance pour créer les conditions psychologiques favorables.
La situation impersonnelle du laboratoire n'est pas naturelle, mais la situation de l'interview traditionnelle non plus. On a beau la baptiser « conversation », les rôles inégaux de l'interrogateur et du candidat conditionnent dans une grande mesure la performance de ce dernier. Le mutisme, les réponses monosyllabiques, la pauvreté du contenu de certains candidats dans ces interviews rappellent les expériences de W. Labov avec les enfants des ghettos noirs. Le candidat est « dans une situation asymétrique où tout ce qu'il dit peut littéralement servir à le condamner » W. Labov : "The Logic of Non-Standard English", reprinted in P. Giglioli, Language and Social Contact, (Penguin Education, 1972, p 185). En ce qui concerne l'étape d'évaluation de l'échantillon de langue fourni par le candidat, je serai beaucoup plus bref. Comme nous avons constaté ci-dessus, l'évaluation de cet échantillon enregistré est tout à fait analogue à la correction d'une épreuve écrite de traduction ou de rédaction. Nous retrouvons les mêmes problèmes de fiabilité et d'objectivité ; les écarts de notes pour un même échantillon entre correcteurs différents ou pour le

même correcteur à des moments différents ont été prouvés souvent (voir par exemple Henri Pieron : Examens et Docimologie, P.U.F 1969). Pourtant les difficultés sont moins aiguës en langues étrangères que pour les rédactions en langue naturelle.

Dans les parties très structurées d'un test, là où le candidat répond à une question, effectue une manipulation grammaticale, nous pouvons établir un barème de correction très objectif. C'est dans cette partie que la M.L.A.E.P, rapporte des corrélations de 0,94 entre les deux correcteurs. Mais pour les parties où le candidat improvise, où il a l'initiative de ce qu'il dit, et c'est ici qu'il montre vraiment qu'il sait utiliser la langue, les critères objectifs sont très difficiles à établir. Les chercheurs dans différentes disciplines ont utilisé des paramètres chiffrables dans leur investigation de tel ou tel aspect de la langue : la vitesse du débit (V.R.R Roy : "Rate of Output -a factor of Oral Proficiency", Canadian Modern LANG. Review, 1969, 26-1, p.14), le nombre et la durée des pauses, le nombre de mots, la proportion de phrases complexes, d'adverbes (V.B. Bernstein: Class Codes and Control, ch. 5 et 6, Paladin 1971, reprises d'articles publiés en 1962 dans Language and Speech; Voir aussi D. Lawton: Social Class, Language and Education ch. 6 Rutledge 1968) le nombre de mots inhabituels, en dehors d'une liste établie pour chaque tâche en fonction de la performance moyenne (voir International Educational Assessment: Study of English as F.L).

Mais ces paramètres sont rarement utilisables. Ces techniques, valables pour l'analyse d'une expérience, ne conviennent pas pour des tests qui doivent être répétés régulièrement, car il y a un risque de répercussion néfaste sur l'enseignement. Par exemple, les candidats qui sauront que la vitesse jouera un rôle dans l'attribution de la note, se forceront dans cette direction et la validité du critère sera faussée.

Mais l'expérience de ces dix dernières années a montré qu'à défaut de techniques d'évaluation réellement objectives on peut améliorer la fiabilité des jugements impressionnistes en utilisant une gamme restreinte de notes (de 0 à 4 ou 5 semble être le mieux) pour chacun des éléments d'appréciation : prononciation, syntaxe, vocabulaire, aisance (fluency), et en définissant la signification de chaque note. D.P. Harris

(Testing English as a Second Language, Mc Graw Hill, 1969, pp 84-85) recommande l'échelle suivante :

Prononciation :
- 5- peu de traces d'accent étranger,
- 4- intelligible, mais accent étranger,
- 3- la prononciation oblige à une écoute intensive et provoque parfois des erreurs de compréhension,
- 2- très difficile à comprendre ; il faut souvent faire répéter,
- 1- les problèmes de prononciation sont tels que la conversation est inintelligible.

On constate que le critère essentiel est l'intelligibilité. C'est une notion difficile à définir ou chiffrer mais sur laquelle les juges, surtout des natifs, peuvent tomber

d'accord après des séances d'entraînement où l'on écoute et note quelques enregistrements, pour discuter des divergences de jugement jusqu'à ce qu'on arrive à une notation régulière. Cet entraînement des correcteurs est la condition essentielle d'une évaluation fiable. Il permet également de trouver une solution opératoire au problème de la norme et des écarts acceptables. Pour la prononciation, cette norme sera double, on acceptera aussi bien des variantes britanniques qu'américaines, voire un mélange des deux.
Nous sommes encore assez loin d'une évaluation objective ; mais cette note attribuée par plusieurs juges pour une impression globale se révèle plus efficace qu'une notation plus objective de points précis prévus d'avance. Une des raisons en est sans doute l'irrégularité de la performance des étudiants de langues étrangères, en anglais du moins. On remarque cette irrégularité dans la prononciation d'un phonème, non seulement dans les mots difficiles, où interviennent les problèmes de distributions, mais aussi dans les réalisations différentes d'un même mot. Quand on juge l'aptitude à prononcer tel phonème (sa compétence) par sa performance dans un ou deux mots dans un texte de lecture, on néglige les autres occurrences de ce phonème, tandis que dans une appréciation globale le juge tient compte intuitivement de l'ensemble des réalisations (aussi bien A.R.E.L.S que M.L.A.E.P rapportent des meilleures corrélations entre correcteurs pour les récits ou la conversation que pour la notation de problèmes précis de structures). Les mêmes problèmes se posent sur le plan syntaxique ou lexical. Cet appel à l'intuition des juges peut paraître dangereux, mais à condition de prendre les précautions évoquées ici (entraînement des correcteurs, utilisation d'une échelle de valeurs bien définie, double correction des cas limites...) nous arriverons à une évaluation d'une fiabilité suffisante, et d'une validité supérieure à un test d'apparence plus objective, car elle se base sur un échantillon de l'utilisation normale de la langue.