Un test, ça se teste. Analyse d'un test de niveau

Nicole Décuré

Nicole Décuré
Les Après-midi de LAIRDIL 1, 1994, pp. 43-54.

Un test, ça se teste
Analyse d'un test de niveau

Buts de l'étude

Au cours de la discussion qui a suivi la conférence, Mike Nicholls a affirmé que les résultats des tests écrits de Cambridge n'étaient pas différents, de façon significative, de ceux d'un test oral. "The results of the oral correlate remarkably well with the results of the written test." Pourquoi alors faire un test écrit, argumente-t-il, puisque nos cours sont, dans la plupart des cas, essentiellement axés sur l'oral? Mais pourquoi faire un test oral, pourrait-on rétorquer? Car tout test d'expression orale prend beaucoup de temps et il est donc difficilement compatible avec de grands nombres.

Il faut d'abord savoir à quoi sert un test. Sanctionne-t-il un travail ou est-il un test d'évaluation d'un niveau de compétences avant de commencer un cours ou une formation? Le contexte est bien différent. Dans le premier cas, l'apprenant-e peut se sentir frustré-e, trompé-e d'avoir travaillé l'oral et de n'être évalué-e que sur l'écrit. Mais dans le deuxième cas, procéder à un test oral devient une question de temps, donc de moyens.

Cette remarque, non développée, de Mike Nicholls sur la corrélation de l'écrit et de l'oral m'a amenée à examiner le test décrit ci-dessous, utilisé par mes collègues et moi-même depuis une vingtaine d'années pour classer les étudiant-es de second cycle scientifique et médical de l'Université Paul Sabatier à Toulouse en groupes de niveau en début d'année. Il est bien loin de remplir toutes les conditions énumérées dans la conférence et cependant il nous donne satisfaction car les classements ainsi opérés s'avèrent fiables à quelques exceptions près. Mais, alors qu'il a été construit de façon empirique, il paraissait intéressant d'examiner son fonctionnement, ses avantages et ses limitations. Cette étude a permis de soulever quelques questions qui méritent d'être examinées dans un travail de recherche ultérieur.

Contexte

Le test s'adresse à des étudiant-es qui vont suivre 75 heures de cours à raison de trois heures par semaine, subdivisées en une séance axée sur l'écrit (grammaire, compréhension et expression écrites) et une partie axée sur l'oral (compréhension et expression orales). En fait, les deux parties accordent une place prépondérante à la communication orale.

Bien que le test ait été raccourci au fil des ans, sa conception est restée la même ainsi que son contenu, ce qui présente l'avantage de donner des résultats consistants d'une année à l'autre. Il teste la compréhension écrite, les connaissances grammaticales, syntaxiques, lexicales et la compréhension orale. L'expression orale a été volontairement écartée car elle prend trop de temps et ce test doit être administré à beaucoup d'étudiant-es à la fois (700 à l'heure actuelle) en un temps très court. Beaucoup d'exercices sont sous forme de QCM qui permettent une correction rapide et quelques-uns exigent des réponses plus ouvertes. Le test se fait en temps limité (les plus faibles ne finissent pas) et la partie de compréhension orale se déroule de façon identique à toutes les séances.

Ce test est insatisfaisant sur le plan intellectuel car il ne teste qu'une petite partie des connaissances et compétences des étudiant-es, notamment en n'évaluant pas la communication orale. Mais il est satisfaisant sur le plan pratique car le barème établi nous permet de classer les étudiant-es en quatre niveaux (de débutant à avancé) avec très peu d'erreurs d'appréciation. Les erreurs de classement sont dues essentiellement à des erreurs d'addition (!), au copiage, à la mauvaise forme le jour du test ou, au contraire, à un hasard heureux. Il y a peu de changements de niveau par la suite, 1 à 2%, essentiellement les cas limites. Ces changements se font surtout vers le bas, à l'exception du niveau 1 (faux-débutant): ils/elles font souvent un mauvais test parce qu'ils/elles n'ont pas fait d'anglais depuis longtemps et il suffit de quelques semaines pour réactiver leurs connaissances. Depuis que les niveaux sont équivalents pour l'obtention d'une u.v. libre (ce qui n'était pas le cas au début où seul le niveau 2 donnait l'u.v.) il y a moins de tricherie car il n'est pas dans l'intérêt des étudiant-es d'être admis-es dans un niveau trop fort. On peut aussi constater avec plaisir que très peu font délibérément un test plus faible afin de se trouver dans un groupe qui sera facile. C'est une u.v. libre, donc choisie, et la grande majorité a envie de progresser et non simplement d'obtenir un diplôme. De plus, le niveau atteint étant mentionné sur le diplôme, il vaut mieux, pour son CV, avoir un bon niveau. Enfin, bien que le test privilégie l'écrit, dans la grande majorité des cas, les étudiant-es sont à l'aise dans la partie orale du cours, mais en général, plus faibles, surtout au niveau 4.

Matériau

Sur les 700 tests d'octobre 1993, ont été gardés, pour faciliter les calculs:

- *niveau 2*: 200 tests (17 éliminés: tests écrits seulement, changements de niveau, non-inscrits, erreurs d'addition);
- *niveau 3:* 200 tests (les premiers dans l'ordre alphabétique);

- *niveau 4:* 150 tests (moins d'étudiant-es, les résultats ont été reportés sur 200);
- Les résultats du niveau 1 n'ont pas été pris en compte car le nombre de copies était trop faible.

Méthodologie

Toutes les notes ont été entrées dans une base de données: pour chaque étudiant-e les notes de chaque test, les notes totales de l'écrit, de l'oral et l'addition.

Il a été effectué

- des graphiques comparatifs à l'intérieur d'un même niveau: entre divers exercices, entre écrit et oral en tenant compte des totaux d'exercices (ramenés à 10);
- des graphiques comparatifs entre niveaux (sur la même base de nombre);
- un calcul de l'écart entre écrit et oral.

Description des exercices

Compréhension et expression écrites *(total des points: 72)*

	Note	**Objet**	**Type**	**Tâche**
Ex. 1	10	CE: reconnaître les verbes dans des titres de journaux.	lecture, fonction des mots	recopier un mot
Ex. 2	10	Vocabulaire: mots outils, expressions de temps (synonymes).	QCM	cocher
Ex. 3	6	Questions: formulation.	écrire des phrases	écrire
Ex. 4	9	Adverbes: liste, mettre des adverbes dans des phrases.	exercice lacunaire	choix dans liste
Ex. 5	5	Structure de la phrase, formes verbales.	éléments dans le désordre	flèches
Ex. 6	10	Emploi des temps dans des phrases séparées.	QCM	cocher
Ex. 7	12	Accord des temps dans un texte suivi.	transformation de l'infinitif	écrire
Ex. 8	10	Trouver des erreurs variées dans des phrases et corriger.	lecture, correction	écrire

Compréhension orale *(total des points: 42)*

	Note	**Objet**	**Type**	**Tâche**
Ex. 9	10	Discrimination des sons.	reconnaissance de sons	écrire des chiffres
Ex. 10	10	Compréhension de mots et de leurs définitions.	QCM oral	cocher
Ex.11	10	Reconnaître un mot dans une phrase.	QCM oral	cocher
Ex. 12	12	Compréhension de phrases (stéréotypes culturels) à l'aide de dessins.	compréhension globale	écrire des chiffres

Deux questions principales ont été examinées:

- *Peut-on raccourcir le test?*

L'augmentation des effectifs rend la correction de plus en plus fastidieuse. Le test, dans sa forme actuelle, n'est pas informatisable. Certaines questions, aussi fermées soient-elles, admettent un grand nombre de réponses, d'où la nécessité d'une correction "intelligente". L'informatisation d'ailleurs, en ne laissant que le seule possibilité de QCM, rend le test moins riche car il n'évalue que les connaissances passives et non actives, la réception et non la production.

Analyser les résultats du test permettrait de voir si certains exercices sont redondants, donnant les mêmes résultats à l'intérieur d'un même niveau. Par exemple, nous avions l'impression subjective que les deux premiers exercices, bien que portant sur des compétences très différentes, donnaient les mêmes notes sur les copies.

Deux ou trois exercices peuvent-ils être suffisamment révélateurs, malgré la notion, couramment admise, que plus un test est long, plus il est fiable (Hughes 36)?

Y a-t-il des exercices qui donnent des résultats plus significativement différents d'un niveau à l'autre? Il nous semblait que c'était le cas de certains exercices, en particulier ceux portant sur la formulation de questions et l'accord des temps.

- *Y a-t-il concordance entre la partie écrite et la partie orale?*

La partie écrite peut se faire individuellement mais la partie orale nécessite un magnétophone, donc une personne présente qui veille à ce que le test se déroule de la même façon pour tout le monde. Lorsque des étudiant-es doivent passer le test en retard, ils/elles ne font que la partie écrite. Ces résultats aussi donnent satisfaction. La compréhension orale est-elle alors redondante? La supprimer permettrait de gagner du temps et de se débarrasser d'exercices dont la correction est difficile.

Il était également tentant d'essayer de voir si les niveaux de départ se retrouvent à l'arrivée. En fait, trop de paramètres entrent en ligne de compte qui sont difficilement

mesurables: le travail fourni par l'étudiant-e, les professeur-e-s différent-e-s, les notations différentes.

L'étude est loin d'être terminée mais nous présentons ici les questions qui se sont posées en cours d'étude et quelques conclusions.

Remarques sur les exercices

Comme il est dit plus haut, les exercices ne sont pas que des QCM et laissent donc place au choix individuel, à la production (même limitée) de langage, aux compétences actives plutôt que passives. Il n'y a pas de hasard possible, comme dans un QCM. L'exercice donnerait peut-être des résultats très différents si des choix étaient proposés, les étudiant-es sachant plus facilement reconnaître une phrase juste que la produire. Cette étude sera menée ultérieurement.

Tout test présente des difficultés qui n'ont rien à voir avec les compétences langagières:

- *Les difficultés inhérentes aux types d'exercices proposés:* l'exercice 3, très artificiel, nécessite des exemples en français pour que l'on comprenne bien ce qui est demandé; la tâche des exercices oraux nécessite aussi des explications et démonstrations.
- *Une évaluation du temps nécessaire et une juste répartition*: si les exercices de la page 1 donnent de meilleurs résultats que ceux de la page 3, ce n'est pas qu'ils sont plus faciles (l'exercice 1 est assez difficile) mais les étudiant-es y consacrent sans doute plus de temps, car c'est le début de la séance et ils/elles évaluent mal le temps global nécessaire. S'ils étaient en fin de test, les résultats seraient peut-être modifiés. C'est également une étude à mener.
- *Un temps d'acclimatation au test*, notamment pour le test oral. Certain-e-s n'ont pas entendu d'anglais depuis longtemps.
- Pour les tests oraux, des éléments tels que la mémoire interfèrent, notamment dans les tests 10 et 11 où quasiment personne ne pense à noter le premier mot, ce qui aiderait grandement dans l'accomplissement de la tâche.
- Pour le test 9, une bonne oreille permet de pallier les déficiences de connaissances. Différencier un mot d'un autre n'implique pas qu'on sache ce qu'ils veulent dire et encore moins les employer. C'est aussi un test qui requiert d'avoir bien compris la technique (deux chiffres sous un mot entraînent les deux mêmes chiffres sous l'homophone) et exige des réflexes rapides.

- Enfin, les références culturelles peuvent faire défaut à certain-e-s: c'est le cas du test 12 qui porte sur des stéréotypes de nationalité. On peut aussi ajouter un paramètre pour cet exercice. Le dessin d'origine était moins clair. Depuis que le dessin a été refait, les résultats sont meilleurs. Ce n'est pas la compréhension qui s'est améliorée, mais la lisibilité du dessin.

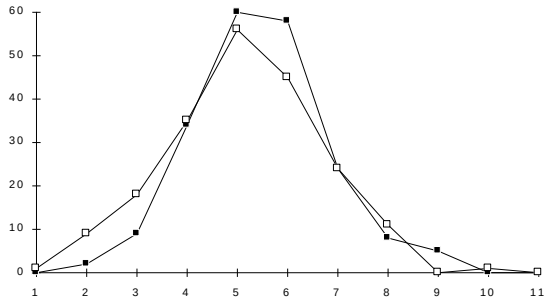
Quelques résultats

Essayons de répondre aux questions de départ.

- *Peut-on raccourcir le test?*

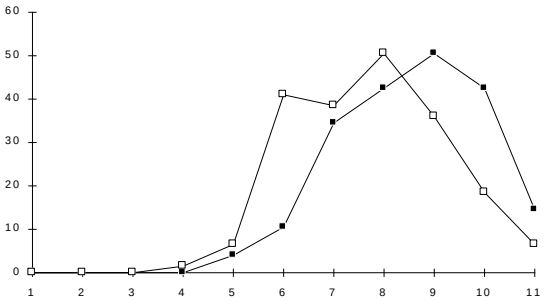
L'impression de départ, selon laquelle les exercices 1 et 2 donnent des résultats quasiment identiques est confirmée.

Niveau 2



Exercices 1 et 2

Niveau 4



Exercices 1 et 2

Comparaison exercices 1 et 2

À la lecture des moyennes de notes de chaque exercice, on s'aperçoit que le test 9, déjà mentionné, n'est pas "rentable", car c'est le test pour lequel il y a le moins de différence d'une copie à l'autre. Comme c'est aussi le plus long et le plus difficile à corriger, nous avons pu le supprimer cette année sans regret.

Moyenne des notes

	niv. 2	**niv. 3**	**niv. 4**
Exercice 1	4.5	5.7	7.6
Exercice 2	4.2	5.3	6.7
Exercice 3	1.9	3.7	4.3
Exercice 4	2.8	4.3	5.7
Exercice 5	2.6	3.5	4.3
Exercice 6	3.7	4.8	6.7
Exercice 7	3.1	5.1	7.4
Exercice 8	0.4	1.2	2.9
Total écrit	23	33	45.7
Total sur 20	***6,4***	***9,2***	***12,7***
Exercice 9	3.6	4.3	5.7
Exercice 10	2.8	3.7	4.9
Exercice 11	4.5	5.2	6.8
Exercice 12	5.9	7.1	8.6
Total oral	18.3	20.3	26
Total sur 20	***8,7***	***9,6***	***12,4***
Total général	39.8	57.8	71.7

- *Y a-t-il concordance entre la partie écrite et la partie orale?*

Les notes ramenées sur 20 viennent confirmer ce que Mike Nicholls affirmait: la corrélation entre les résultats d'écrit et ceux d'oral, sauf pour les niveaux faibles. La supériorité de l'oral, dans ce cas, peut s'expliquer par la nature même des tests qui, comme dit plus haut, requièrent une bonne oreille et une bonne mémoire plus qu'une réelle compréhension.

- Lorsque le test est administré individuellement, seule la partie écrite est donnée et un barème a été établi. En fait, pour affiner l'analyse, si l'on regarde uniquement les notes de la partie écrite, de façon individuelle et non plus globale, on s'aperçoit que la note d'oral est loin d'être négligeable car elle corrige la note d'écrit, surtout dans le niveau 3.
- En effet, au niveau 4, le nombre d'étudiant-es ayant obtenu moins de 37 points (la barre fixée par extrapolation) est infime: 10 sur 150, dont 5 ont une note globale à la limite des niveaux 3 et 4.
- Au niveau 2, 43 étudiant-es sur 200, soit un peu moins d'un quart, ont une note d'écrit très faible. Parmi eux/elles, une dizaine est à la limite du niveau 1 et six ont une note d'oral nettement supérieure. Pour le reste, la compréhension orale a compensé.

- Pour un tiers du niveau 3, la partie écrite est soit au-dessous, soit au-dessus des barres fixées.

Doit-on en conclure que pour économiser du temps on pourrait, comme dans les oraux de rattrapage, ne pas faire de test oral pour ceux et celles qui ont une note élevée à l'écrit? Le test oral permettrait d'affiner la notation pour les niveaux les plus bas.

Conclusion

Le test décrit ci-dessus est loin de remplir les critères de validité et de fiabilité qui semblent nécessaires à une évaluation objective et fine. Mais, *in fine*, tout test ne vaut que par l'habitude qu'ont les enseignant-e-s concerné-e-s de l'utiliser car ils/elles l'ajustent à leurs besoins, leur connaissance du terrain, leur enseignement.

Bibliographie

ALDERSON, J. Charles & Brian NORTH. *Language Testing in the 1990's: The Communication Legacy*. Londres: MEP/British Council, 1991.

ALLEN, J.P.B. & Alan DAVIES, eds. *Testing and Experimental Methods*. Oxford: Oxford University Press, 1977.

BAKER, David. *Language Testing: A Critical Survey and Practical Guide*. Londres: Edward Arnold, 1989.

CARROLL, Brendan J. *Testing Communicative Performance*. *An Interim Study*. Oxford: Pergamon Press, 1980.

COHEN, Andrew D. *Testing Language Ability in the Classroom*. Rowley, Mass.: Newbury House, 1980.

DAVIES, Alan. *Principles of Language Testing*. Oxford: Blackwell, 1990.

DAVIES, Susan & Richard WEST. *English Language Examinations*. Londres: Longman, 1989.

FULCHER, Glenn. Tests of Oral Performance: The Need for Data-based Criteria . *ELT Journal* 41 : 4, 1987, pp. 287-291.

HARRIS, David P. *Testing English as a Second Language*. New York: McGraw Hill, 1969.

HARRISON, Andrew. *A Language Testing Handbook*. Londres: MEP, 1983.

HEATON, J.B. *Classroom Testing*. Londres: Longman, 1990.

HORNER, David. Testing: Introduction and Review . *TESOL News* 10 : 1, 1990. (Publié également dans *The Best of TESOL France News* 1 : 1, 1994, pp. 109-114.)

HUGHES, Arthur. *Testing for Language Teachers*. Cambridge University Press, 1989. (C*ontient 3 pages de bibliographie*)

Issues in Language Testing. ELT Document 111, British Council, 1981.

KNIGHT, Ben. Assessing Speaking Skills: A Workshop for Teacher Development. *ELT Journal* 46 : 3, 1992, pp. 294-302.

Language Testing . Londres: Edward Arnold. (Revue bi-annuelle.)

RHEA-DICKINS, Pauline & Kevin GERMAINE. *Evaluation*. Oxford: Oxford University Press, 1992.

SKEHAN, Peter. Communicative Language Testing . *TESOL News* 10 : 1, 1990. (Publié également dans *The Best of TESOL France News* 1 : 1, 1994, pp. 115-127.)

The Testing of Oral Proficiency. *System* 20 : 3 (special issue), 1992.

VALETTE, Rebecca M. *Modern Language Testing: A Handbook*. New York: Harcourt, Brace & World, 1967.

WEIR, Cyril J. *Communicative Language Testing*. New York: Prentice-Hall, 1990.