# L'ESTIMATION A PRIORI DES NIVEAUX DE DIFFICULTE DE QUESTIONS PAR DES EXPERTS DU CONTENU EST-ELLE VALIDE ? L'EXEMPLE DES TESTS CERTIFICATIFS CHEZ LES FONCTIONNAIRES FEDERAUX EN BELGIQUE.

**Sébastien Remy*, Julie Camerman****

* *Institut de Formation de l'Administration fédérale (IFA), Bruxelles, Belgique, sebastien.remy@ofoifa.fgov.be*

** *Institut de Formation de l'Administration fédérale (IFA), Bruxelles, Belgique, julie.camerman@ofoifa.fgov.be*

---

***Mots-clés :*** *indice de difficulté, niveau d'études, évaluation des acquis, tests certificatifs*.

***Résumé.*** *La présente étude a pour objectif de tester la validité de l'estimation a priori des niveaux de difficulté des questions d'un test, par des experts du contenu. Nous proposons à cet effet, une double méthode qui consiste, d'une part, à comparer les résultats à trois niveaux de difficultés de questions (facile, moyen, difficile) au sein d'une même population et, d'autre part, à comparer deux populations, avec des niveaux d'études différents, pour les mêmes questions. Les résultats montrent, d'une part, que les scores les plus élevés sont obtenus, dans l'ordre décroissant, pour les questions estimées comme les plus faciles, moyennes, puis difficiles. Et, d'autre part, les agents de niveau d'études le plus bas obtiennent des scores qui deviennent significativement différents de ceux des agents de niveau d'études le plus haut, lorsque les questions estimées comme difficiles sont supprimées. Ainsi, la technique qui consiste a faire appel à des experts du contenu pour l'estimation a priori des niveaux de difficulté semble donc valide.*

---

## 1. Introduction

### *1.1 Contexte organisationnel*

L'Institut de Formation de l'Administration fédérale est un organisme fédéral belge dont la mission principale est la formation des fonctionnaires fédéraux. Il couvre 90 organisations clientes, soit environ 80 000 participants potentiels.

Parmi l'offre de l'IFA, des formations certifiées sont organisées. Il s'agit de formations liées à la carrière. Elles se terminent par un test qui vérifie si les agents ont atteint les objectifs d'apprentissage de manière suffisante. Le seuil de réussite est fixé à 60%. Si le test est réussi, l'agent bénéficie alors d'une prime de développement des compétences ou d'un changement d'échelle de traitement et ce, en fonction de sa situation et de son niveau.

Chaque année, environ 80 tests sont organisés dans l'une ou plusieurs des trois langues nationales (français, néerlandais et allemand), ce qui représente environ 12 000 participants/an. Des formations spécifiques sont disponibles pour tous les niveaux d'agent (A, B, C et D)[1]. D'autres

1. Ces niveaux sont relatifs aux niveaux d'études :
   A : porteurs d'un diplôme universitaire ou équivalent.

formations sont, quant à elles, jumelées pour deux niveaux d'agent contigus (A-B, B-C ou C-D), c'est-à-dire qu'ils suivent la même formation.

### *1.2 Problématique*

Ces formations certifiées jumelées ont nécessité que les tests soient différents entre les deux niveaux d'agents. En effet, d'une part, les niveaux d'études des agents sont différents. D'autre part, la réussite au test donne droit à des primes différentes en fonction du niveau de l'agent.

Les tests se déroulent entre 3 semaines et 3 mois après le dernier jour de formation. Ils sont construits sur base d'une table de spécification qui indique la répartition des questions en fonction des thèmes vus au cours, du niveau taxonomique et de difficulté des questions. Cette table de spécification permet également de s'assurer de la fidélité du test d'une session à l'autre. Son objectif principal est donc d'assurer une comparabilité des résultats et une égalité de traitement des agents. Les tests sont construits selon la répartition des niveaux de difficulté des questions suivante : 25% faciles, 50% moyennes et 25% difficiles.

Dans le cadre des formations certifiées jumelées, les questions difficiles du niveau d'agent le plus haut sont supprimées, ou remplacées par des questions faciles, pour le niveau d'agent le plus bas. Ainsi, ces tests comportent à la fois une partie commune aux deux niveaux d'agent et une partie spécifique à chaque niveau.

La mesure la plus valide pour estimer le niveau de difficulté d'une question se fait en principe a posteriori et sur base des résultats effectifs d'un test, à l'aide du calcul de divers indices psychométriques et/ou édumétriques (Leclercq, 2003). Parmi ces indices, nous trouvons notamment ceux de discrimination ou de difficulté d'une question. Il est alors nécessaire de réaliser un étalonnage du test (Mucchielli, 2008). Cependant, cela n'est pas toujours aisé (Laveault & Grégoire, 2002). Par exemple, organiser une phase de pré-test demande beaucoup de temps et de ressources. Il faut notamment un échantillon (autre que les futurs participants) disponible pour le pré-test, avec le risque que les questions soient diffusées avant le test.

De ce fait, une estimation a priori, basée sur l'avis d'experts du contenu (formateurs/rédacteurs des questions), sera le plus souvent réalisée. C'est l'option que nous avons choisie pour les formations certifiées. Ainsi, la présente étude a pour objectif de tester la validité de cette estimation a priori à travers une double méthode. Cette méthode consiste, d'une part, à comparer les résultats à trois niveaux de difficultés de questions (facile, moyen, difficile) au sein d'une même population et, d'autre part, à comparer deux populations de niveaux d'études différents pour les mêmes questions.

### *1.3 Hypothèses*

#### *1.3.1 Au sein d'une même population*

Nous posons l'hypothèse (H1) que les résultats aux questions estimées a priori de niveau difficile sont inférieurs à ceux obtenus aux questions de niveau moyen et de niveau facile. De même, nous posons l'hypothèse (H2) que les résultats aux questions estimées a priori de niveau moyen sont inférieurs à ceux obtenus aux questions de niveau facile. En résumé, nous supposons que les résultats aux questions, au sein d'une même population, sont de cet ordre : difficiles < moyennes < faciles.

---

B : porteurs d'un diplôme d'enseignement supérieur de type court (bachelor) ou un diplôme ou certificat de candidature.

C : porteurs d'un diplôme d'enseignement secondaire supérieur.

D : aucune exigence de diplôme n'est requise.

*1.3.2 Au sein de deux populations différentes*

Nous posons l'hypothèse (H3) que les résultats aux questions identiques seront supérieurs pour les agents de niveau d'études le plus haut par rapport aux résultats des agents de niveau d'études le plus bas.

## 2. Méthodologie

### *2.1 Matériel*

Les tests certificatifs dont il est question dans cette étude portent sur des matières bureautiques (suite Office 2007 : Excel, Word, Powerpoint et Outlook). Les formations, jumelées, concernées par ces tests, sont accessibles aux agents des niveaux B et C, francophones et néerlandophones, de la fonction publique fédérale belge. Les agents de niveau B sont des experts administratif, financier, technique et ICT. Les agents de niveau C sont des assistants administratif et technique.

Les tests sont constitués de 25 questions pour les agents de niveau B réparties comme suit : 7 faciles, 13 moyennes et 5 difficiles. Les cinq questions estimées comme difficiles sont supprimées pour les agents de niveau C.

### *2.2 Echantillon*

L'échantillon est constitué de 1052 fonctionnaires fédéraux, répartis sur 8 tests en bureautique passés d'octobre 2011 à mars 2012. Les caractéristiques de cet échantillon sont présentées dans le tableau ci-dessous *(Tableau 1)*.

| | | |
|---|---|---|
| **Total** | 1052 fonctionnaires fédéraux | |
| **Age moyen** | 43,04 ± 10,06 (21 à 64 ans) | |
| **Sexe** | 43,5% hommes | 56,5% femmes |
| **Rôles linguistiques** | 49,7% francophones | 50,3% néerlandophones |
| **Niveaux des agents** | 27,6% niveau B (experts administratif, financier, technique et ICT) | 72,4% niveau C (assistants administratif et technique) |

**Tableau 1 :** Description de l'échantillon

### *2.3 Analyses*

Les analyses ont été réalisées à l'aide de t-tests pairés de Student, de tests du chi-carré de Pearson et d'analyses de variances (ANOVA) avec test post-hoc de Bonferroni. Pour ces analyses, les groupes d'agents de niveaux B et C ont été répartis proportionnellement selon le sexe et les tranches d'âge.

## 3. Résultats

*3.1.1 Au sein d'une même population*

Pour les agents de niveaux B et C confondus, les comparaisons des scores (moyennes en %) par niveau de difficulté des questions montrent *(Tableau 2)* :

- Les questions estimées comme faciles sont mieux réussies que celles estimées comme moyennes (respectivement, 88.18 ± 16.11 et 77.56 ± 16.85 ; t (1051) = 22.091 ; p < .001) ;
- Les questions estimées comme moyennes sont mieux réussies que celles estimées comme difficiles (respectivement, 78.80 ± 14.44 et 74.12 ± 20.51 ; t (289) = 3.562 ; p < .001) ;
- Les questions estimées comme faciles sont mieux réussies que celles estimées comme difficiles (respectivement, 92.06 ± 13.17 et 74.72 ± 20.51 ; t (289) = 14.968 ; p < .001).

| | **Estimées faciles (F) 92.06 ± 13.17** | **Estimées moyennes (M) 77.56 ± 16.85** | **Estimées difficiles (D) 74.12 ± 20.51** |
|---|---|---|---|
| **Estimées faciles (F) 88.18 ± 16.11** | | **H1** t (1051) = 22.091 ; p < .001** | |
| **Estimées moyennes (M) 78.80 ± 14.44** | | | **H2** t (289) = 3.562 ; p < .001** |
| **Estimées difficiles (D) 74.72 ± 20.51** | **H1** t (289) = 14.968 ; p < .001** | | |

**Tableau 2 :** Scores (moyennes en %) par niveau de difficulté des questions

*3.1.2 Au sein de deux populations différentes*

La comparaison des taux de réussite en fonction du niveau des agents montre *(Tableau 3)* que les agents de niveau C ont des taux de réussite significativement plus bas que ceux des agents de niveau B (respectivement, 91.60% et 95.17% ; $\chi^2$ = 3,903 ; p < .05). Cependant, aucune différence significative n'est observée au niveau des scores (moyennes en %) obtenus par les agents de niveau B et C (respectivement, 81.60 ± 12.14 et 80.40 ± 15.67 ; F (1,1050) = 1.378 ; p = .241).

Les scores ont ensuite été comparés en ne prenant en compte que les questions communes, c'est-à-dire après avoir ôté les questions estimées comme difficiles pour les agents de niveau B. Les comparaisons montrent alors *(Tableau 3)* que les agents de niveau C obtiennent des scores significativement plus bas que les agents de niveau B (respectivement, 80.40 ± 15.67 et 83.16 ± 12.05 ; F (1,1050) = 7.333 ; p < .01).

| | | **Agents niveau B** | **Agents niveau C** |
|---|---|---|---|
| **Taux de réussite** | $\chi^2$ = 3,903 ; p < .05* | 95,17% | 91,60% |
| **Scores (moyennes en %) F + M + D** | F (1,1050) = 1.378 ; p = .241 | 81.60 ± 12.14 | 80.40 ± 15.67 |

| | | | |
|---|---|---|---|
| **Scores (moyennes en %)**<br>**F + M[2]** | **H3**<br>F (1,1050) =<br>7.333 ; p < .01* | 83.16<br>± 12.05 | 80.40<br>± 15.67 |

**Tableau 3 :** Taux de réussite et scores (moyennes en %) en fonction du niveau des agents

## 4. Discussion

La validité de l'estimation a priori des niveaux de difficulté des questions, par des experts du contenu, semble donc démontrée. En effet, d'une part, les scores les plus élevés sont obtenus, dans l'ordre décroissant, pour les questions estimées comme les plus faciles, moyennes, puis difficiles. Et, d'autre part, les agents de niveau C obtiennent des scores qui deviennent significativement différents de ceux des agents de niveau B lorsque les questions estimées comme difficiles sont supprimées.

Ainsi, la technique qui consiste a faire appel à des experts du contenu pour l'estimation a priori des niveaux de difficulté des questions semble valide. Ces résultats représentent donc une piste intéressante pour les tests qui s'adressent à des populations de niveaux d'études différents en créant des tests de niveaux de difficulté qui leur soient adaptés. Ils s'avèrent également intéressants pour des petits groupes où des techniques de TRI ne sont pas applicables (Pini, 2006) ou lorsqu'on ne peut pas faire d'étalonnage.

Toutefois, cette étude porte sur des tests liés à un contenu spécifique (bureautique) et des populations spécifiques (agents de niveaux B et C). Il faudrait donc répliquer cette étude pour d'autres contenus et auprès d'agents d'autres niveaux, afin de vérifier la validité externe de ces résultats (Meyer, 2005).

D'autres critères pourraient également être pris en compte. Par exemple, Raîche (2004) oppose le postulat déterministe qui suppose que si une personne réussit l'item le plus difficile, elle réussira alors tous les items plus faciles, à celui probabiliste qui considère qu'une personne ne réussira pas forcément les plus faciles si elle a réussi les plus difficiles. En effet, une personne peut échouer à une question facile pour diverses raisons qui ne sont pas inhérentes au niveau de difficulté de cette question (par exemple, parce qu'il a mal lu la consigne).

Par ailleurs, ces résultats peuvent également indiquer que cette estimation est valide uniquement pour un niveau d'agent. En effet, les questions ont été rédigées pour les agents du niveau le plus haut. Or, ces dernières ont été dans l'ensemble moins bien réussies par les agents du niveau le plus bas.

## 5. Références

Laveault, D. & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Bruxelles : Editions De Boek université.

Leclercq, D. (Eds) (2003). *Diagnostic cognitif et métacognitif au seuil de l'université. Le projet MOHICAN mené par les 9 universités de la Communauté Française Wallonie Bruxelles.* Les Editions de l'Université de Liège.

Meyer, T. (2005). Validité externe et méthode expérimentale, *Questions de communication*, 7, 209-222.

Mucchielli, R. (2008). *Les méthodes actives dans la pédagogie des adultes*. Issy-les-Moulineaux : ESF Editeur.

2. Uniquement les questions communes faciles (F) et moyennes (M), sans les questions difficiles (D) des agents de niveau B. Pour ces analyses, les groupes d'agents de niveaux B et C ont été répartis proportionnellement selon le sexe et les tranches d'âge.

Pini, G. (2006). *A propos de la théorie des réponses aux items (TRI – IRT). Le cas d'items dichotomiques.* Groupe Edumétrie : Qualité de la mesure en éducation.

Raîche, G. (2004). *L'évaluation des compétences à l'enseignement supérieur :vers une vision intégratrice de l'évaluation des apprentissages*. Montréal, Québec : Université du Québec à Montréal.