Approches psychométriques de l'évaluation à l'aide de tests

Marianne Mavel, Pascal Bessonneau, Christophe Lalanne

Approches psychométriques de l'évaluation à l'aide de tests

Marianne Mavel[1]*, Pascal Bessonneau, Christophe Lalanne*
Centre International d'Études Pédagogiques (CIEP)

Introduction

Le ministère de l'Éducation nationale, de l'enseignement et de la recherche a confié au CIEP, Centre International d'Études Pédagogiques, établissement public national à caractère administratif sous tutelle de ce ministère, "la responsabilité de l'organisation hors de France des examens institués par le ministère de l'Éducation nationale pour évaluer l'enseignement du français langue étrangère" (article 2 du décret 2000 – 1017 du 12 octobre 2000).

Dans cette optique, le CIEP assure la conception et la gestion pédagogique et administrative de diplômes officiels destinés à tous les publics non francophones: le DELF (diplôme d'étude de langue française), le DALF (diplôme approfondi en langue française) et depuis peu le DILF (diplôme initial de langue française). Il gère également le test TCF (test de connaissance en français). Ces certifications évaluent les compétences langagières des candidat/es en français. Elles sont harmonisées sur le cadre européen commun de référence des langues (CECR), c'est-à-dire qu'elles sont calibrées et étalonnées sur des normes européennes et internationales. Le CECR a été élaboré par la division des politiques linguistiques du Conseil de l'Europe. Il s'agit d'un outil de référence européen pour l'évaluation, l'enseignement et l'apprentissage des langues. On retrouve également en annexe les six niveaux communs de référence établis par ALTE (Association of language testers in Europe – association des centres d'évaluation en langues en Europe). La qualité des certifications proposées par le CIEP est toujours vérifiée par une expertise pédagogique et, dans le cas du TCF et du DELF, une expertise psychométrique des épreuves de compréhension et de maîtrise des structures de la langue est aussi réalisée.

La psychométrie regroupe un ensemble de méthodes dont l'objectif est de mesurer de manière équitable et rigoureuse des capacités humaines qui ne sont pas directement observables (traits latents) telles que les compétences intellectuelles ou les connaissances en français, par exemple, (voir Bernier et Pietrulewicz, 1997). L'origine de la psychométrie remonte à la fin du 19ème siècle notamment avec les travaux de Wundt, Galton et Cattell. Cette discipline s'est développée au début du 20ème siècle avec la mise au point par Binet et Simon en 1905 de l'échelle métrique de l'intelligence. Cet outil permet d'évaluer le niveau mental d'un enfant en prenant compte de son âge. Les instruments de mesure utilisés pour évaluer ces capacités

[1] Cellule qualité et expertises - Psychométrie, département évaluation et certifications, Centre International d'Études Pédagogiques (CIEP), 1 avenue Léon-Journault, 92318 Sèvres cedex. Contact : mavel@ciep.fr.

humaines doivent posséder un certain nombre de qualités fondamentales: l'objectivité, la validité et la fidélité de la mesure. Ils prennent des formes très variées telles que des tests de connaissance, des questionnaires ou des échelles. Un test est un instrument d'évaluation dont l'objectif est de mesurer une faculté essentiellement cognitive telle que la compétence langagière. Un test se compose de plusieurs items dont les réponses sont binaires ou multiples. Pour quantifier les compétences individuelles, on peut appliquer un modèle mathématique de mesure sur les données recueillies à l'aide de l'instrument de mesure. Les constructeurs de tests ont souvent recours à deux modèles: la théorie classique des tests (TCT) et la théorie de réponse à l'item (IRT). La TCT fournit des informations sur la qualité fonctionnelle des items d'un test et l'IRT permet de donner aux candidat/es des scores fidèles et précis reflétant bien leurs positions par rapport au trait latent mesuré.

Cet article traitera de l'analyse psychométrique des items d'un test. La première partie sera consacrée au cadre général de la construction d'un instrument de mesure et visera à souligner les qualités fondamentales d'un instrument de mesure et la TCT. La deuxième partie développera le concept des modèles de l'IRT. La dernière partie présentera des exemples d'items du DELF analysés à l'aide de ces deux modèles de mesure.

I. Cadre général de la construction d'un instrument de mesure

1.1. Le trait latent

Un instrument de mesure cherche à mesurer une caractéristique chez un individu qui n'est pas directement mesurable, on appelle cette caractéristique un trait latent ou le construit de l'instrument. Par exemple, dans le cadre de l'évaluation des langues, la compétence langagière est assimilée à un trait latent.

1.2. L'objectivité

Un instrument de mesure est considéré comme étant objectif si les résultats qu'il fournit ne dépendent pas du correcteur ou de l'évaluateur qui l'utilise. Un individu devrait donc obtenir le même résultat à une épreuve quelle que soit la personne qui a corrigé sa copie. L'objectivité peut être assimilée à la fidélité inter-juges (inter-correcteurs). Dans le cadre de la correction d'épreuves de production écrite, par exemple, l'utilisation de grilles de correction standardisées et de manuels de correction permettent de limiter les biais classiques de correction (effet de halo, de compensation) et fournit un cadre de référence pour les correcteurs/rices. Il existe différentes méthodes pour s'assurer de leur degré d'objectivité: le calcul de coefficients de corrélation ou de concordance tel que le Kappa de Cohen, ou l'application d'un modèle de réponse à l'item, par exemple le modèle de Rasch à facettes multiples.

1.3. La fidélité

La fidélité ou fiabilité désigne la capacité d'un instrument de mesure à appréhender avec précision les caractéristiques qu'il est censé mesurer. Un

instrument de mesure est considéré fidèle si le résultat qu'il produit est reproductible. En théorie, un individu auquel on aura appliqué plusieurs fois le même instrument de mesure obtiendra à chaque fois le même score. L'analyse de la fidélité permet de s'assurer de l'équité des différentes versions d'un instrument de mesure. Il existe différentes méthodes pour évaluer la fidélité: les tests parallèles, le test-retest, la consistance interne du test (méthode des moitiés, alpha de Cronbach).

La méthode des tests parallèles donne une estimation du degré d'équivalence entre deux versions d'un même test composées d'items différents. Elle consiste à calculer la corrélation entre les scores obtenus aux deux versions du test qui auront été passées par les mêmes individus. La difficulté consiste à s'assurer de l'équivalence du contenu et du construit des deux versions.

La méthode du test-retest permet de mesurer la stabilité du test dans le temps. La même version du test est appliquée au même groupe de candidat/es à deux moments différents plus ou moins espacés dans le temps. Tout comme la méthode des tests parallèles, on calcule la corrélation entre les deux passations du test. Le principal problème de cette méthode est l'effet d'apprentissage chez les candidat/es dû au fait qu'ils/elles repassent la même version du test.

La méthode la plus souvent utilisée pour estimer la fidélité consiste à évaluer la consistance interne du test (ou son homogénéité) car elle ne nécessite qu'une seule administration du test et qu'elle s'applique aussi bien aux items dichotomiques[2] qu'aux items polytomiques[3]. Évaluer la consistance interne d'un test consiste à déterminer dans quelle mesure les items le composant évaluent la caractéristique (le trait latent) que le test doit mesurer. Comme il est impossible de déterminer la fidélité d'un instrument de mesure à partir d'une seule version de celui-ci, on obtient, en estimant la consistance interne du test, la limite inférieure de la fidélité et on sous-estime la fidélité de l'instrument. On juge qu'elle est satisfaisante si le coefficient de fidélité est assez élevé (alpha de Cronbach > 0,70).

1.4. La validité

La validité d'un instrument de mesure désigne son aptitude à bien mesurer ce qu'il est censé mesurer et uniquement cela. On distingue trois types de validité: la validité de contenu, la validité de construit et la validité prédictive. La vérification des deux premières repose essentiellement sur le jugement professionnel d'experts.

Un test est valide du point de vue de son *contenu* s'il est représentatif en tant qu'échantillon du domaine de comportement mesuré. On détermine la représentativité du test en s'assurant que ses contenus sont pertinents et couvrent bien le domaine que le test est censé mesurer. Dans le cadre de l'évaluation des compétences langagières, le CECR fournit un ensemble de recommandations pour les concepteurs de tests.

---

[2] Un item dichotomique est un item binaire, il est soit réussi, soit échoué. Le score de cet item ne peut alors prendre que deux valeurs: 0 pour l'échec et 1 pour le succès par exemple.

[3] Un item polytomique est un item dont le score prend plus de deux valeurs. Le score varie entre 0 et une valeur supérieure à 1 appelée score maximal de l'item. Par exemple, un item noté sur deux points pourra être codé de la manière suivante: 0, l'item est échoué; 1, l'item est partiellement réussi; 2, l'item est complètement réussi.

En ce qui concerne la validité de *construit*, il s'agit d'établir jusqu'à quel point le test fournit une mesure adéquate du construit théorique qui est mesuré (utilisation d'un ensemble de méthodes). Cette validation s'établit à partir d'une accumulation progressive de données qui viennent appuyer l'hypothèse selon laquelle "le test mesure bien ce qu'il prétend mesurer".

La validité *prédictive* s'applique lorsque l'on cherche à connaître le degré auquel les scores d'un test permettent de prédire la caractéristique étudiée. Par exemple, les tests d'aptitudes scolaires permettent de prédire les chances de réussite à l'école.

1.5. La théorie classique des tests

La théorie classique des tests (TCT), appelée aussi théorie du "score vrai" a été développée au début du 20ème siècle (Dickes *et al*., 1994). Elle est utilisée essentiellement pour détecter un certain nombre de dysfonctionnements concernant les composantes du test (les items) et à estimer la fidélité de l'instrument de mesure. Elle permet de calculer un ensemble d'indicateurs permettant de juger de la qualité du test: le score brut (score total[4]), la difficulté de l'item, le pouvoir discriminant de l'item et l'indice de consistance interne du test. Le modèle de mesure utilisé est un modèle linéaire. Cette théorie est basée sur la décomposition des scores observés en un score vrai et une erreur de mesure. La relation fondamentale postulée dans ce cadre théorique est alors: score observé = score vrai + erreur de mesure.

Le score observé d'une personne est le score qu'elle a obtenu en passant le test. On peut concevoir le score vrai d'un individu comme étant le score qu'il obtiendrait à un test parfait, c'est-à-dire un test qui mesure sans erreur. Dans le cas d'un test parfait, le score observé serait égal au score vrai. Cependant, en pratique, il existe toujours une marge d'erreur. On peut aussi exprimer le score vrai comme étant le score moyen d'une personne qui serait évaluée indéfiniment avec le même instrument de mesure. Le score vrai et l'erreur de mesure sont inconnus et l'on suppose l'erreur de mesure aléatoire.

1.5.1. Analyse des items

L'analyse des items consiste à étudier leurs qualités fonctionnelles à l'aide de deux indicateurs statistiques: l'indice de difficulté et le pouvoir discriminant. L'interprétation de ces indicateurs est facilitée par une analyse graphique. Cette analyse permet également de détecter un certain nombre de dysfonctionnements: des items trop faciles ou trop difficiles, des réponses ambigües, des distracteurs non choisis. L'ensemble de ces informations est évidemment très utile lors de l'élaboration d'un test.

Indice de difficulté d'un item

L'indice de difficulté d'un item calculé sur un échantillon donné est une estimation de la difficulté de l'item. L'indice de difficulté d'un item polytomique est égal au score moyen relatif des candidat/es à cet item (score moyen des candidat/es

[4] Le score total est le nombre total de points qu'a obtenu le/la candidat/e au test.

divisé par le score maximal de l'item). Dans le cas dichotomique, il s'agit plus simplement de la proportion de candidat/es ayant répondu correctement à l'item. Cet indice varie entre 0 et 1. Plus la valeur de l'indice est élevée, plus l'item est facile.

Indice de discrimination d'un item

Le pouvoir discriminant d'un item est sa capacité à différencier les candidat/es en fonction d'un critère donné, par exemple leur niveau de compétence. Il existe plusieurs types d'indices pour évaluer le pouvoir discriminant d'un item: la corrélation entre la réponse à l'item et le score total au test (item-test), la corrélation entre la réponse à l'item et le score au test sans cet item (item-reste) et pour les items à choix multiple, les corrélations entre les distracteurs et le score au test.

L'indice de discrimination est un indice d'homogénéité; il indique dans quelle mesure un item contribue à la détermination du score total. Cet indice varie entre -1 et 1. S'il est proche de zéro, l'item ne permet pas de différencier les candidat/es selon leur compétence. S'il est négatif, l'item présente alors une anomalie: il est mieux réussi par les candidat/es qui ont obtenu un score total faible.

Notons que l'utilisation de la corrélation item-test présente un inconvénient lié au fait que le score à l'item est compris dans le score total du test. Ainsi, la corrélation entre la réponse à l'item et le score total peut être positive alors que l'item ne contribue pas à augmenter le score total des candidat/es en fonction de leur compétence. Le coefficient de corrélation item-reste permet de corriger ce problème. On note que la différence entre la corrélation item-test et la corrélation item-reste est assez élevée quand un test est composé d'un petit nombre d'items.

En pratique, dans le cas des items dichotomiques, le coefficient de corrélation point-bisérial, noté rpbis, est très souvent utilisé comme indice de discrimination. Dans le cas d'un test composé d'items polytomiques, on utilise le coefficient de corrélation de Bravais-Pearson.

Analyse graphique

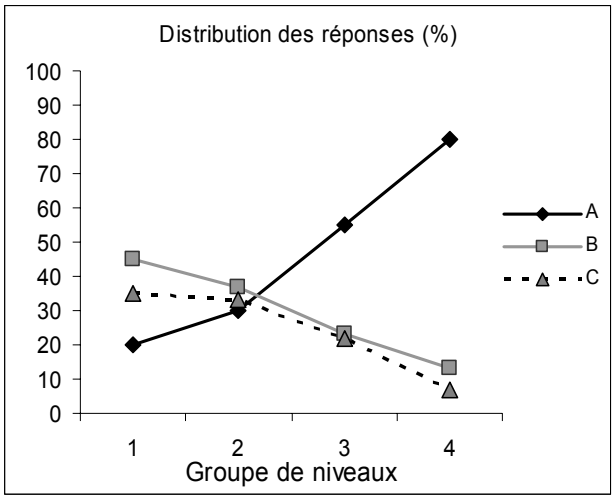
L'analyse graphique d'un item est très utile car elle permet de visualiser la difficulté et le pouvoir discriminant de l'item. Les figures 1a et 1b sont les représentations graphiques des réponses à deux items à choix multiples dont la clé[5] est A (B et C sont les distracteurs[6]). Nous pouvons lire sur ces graphiques la proportion de réponses correctes et incorrectes en fonction de quatre groupes de niveaux, de taille à peu près équivalente. Ces groupes de candidat/es sont ordonnés par le score total (score brut): le groupe 1 est ainsi composé des candidat/es dont les scores totaux sont les plus faibles et le groupe 4 des candidat/es dont les scores totaux sont les plus élevés. Pour chaque groupe de niveau, la proportion de réponses correctes (réponse A) est calculée (ici en pourcentages), ainsi que les proportions de réponses incorrectes B et C. On retrouve ces proportions en ordonnées sur les graphiques de la figure 1.

[5] La clé est la bonne réponse à l'item.
[6] Les distracteurs sont les réponses incorrectes d'un l'item à choix multiple.

Dans le cas de la figure 1a, le taux de réponses correctes (courbe A) augmente fortement en fonction du niveau des candidat/es: on passe d'un taux de réussite de 20 % pour le groupe 1 à un taux de 80% pour le groupe 4. À l'inverse, la proportion de réponses incorrectes (courbes B et C) diminue en fonction du niveau des candidat/es. Ainsi, lorsque l'item permet de bien distinguer les candidat/es en fonction de leur niveau, on dira que l'item est discriminant. Le coefficient de corrélation point bisérial de cet item est bien supérieur à 0; il vaut 0,50 ce qui indique que cet item a un fort pouvoir discriminant et confirme ce que l'on a observé graphiquement. Concernant la difficulté de l'item, elle semble plutôt moyenne car les candidat/es de bon niveau réussissent en moyenne très bien cet item contrairement aux candidat/es dont le niveau est plus faible: le taux de réussite des groupes 3 et 4 est supérieur à 50 % et celui des groupes 1 et 2 est inférieur à 30 %.

L'item de la figure 1b est un item très facile, son taux de réussite stagne autour de 95 % quel que soit le groupe de niveau. Cet item ne permet donc pas de différencier les candidat/es selon leur groupe de niveau, ce que l'on vérifie aisément à l'aide de l'indice de discrimination (cf. 1.5.1. Analyse des items, § Indice de discrimination d'un item), le coefficient de corrélation point-bisérial, qui vaut 0,08 (valeur très proche de 0).

(a)



(b)

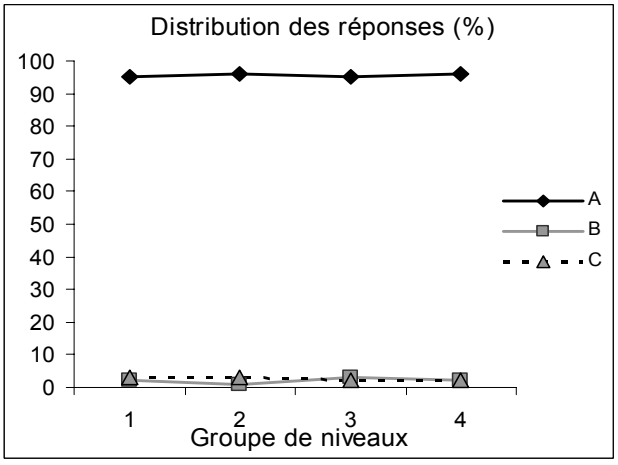


Figure 1 – **Graphique des proportions de réponses A, B et C des candidat/es pour un item donné.**

1.5.2. Fidélité – alpha de Cronbach

L'alpha de Cronbach[7] est un indice permettant de mesurer la consistance interne d'un test. Il permet de déterminer dans quelle mesure l'ensemble des items évalue bien la caractéristique que le test doit mesurer. La valeur de cet indice dépend à la fois de la longueur du test (nombre d'items) et de l'homogénéité des items (déterminée à partir des inter-corrélations entre items). Il est à noter qu'une façon d'augmenter la consistance interne d'un test est de rajouter des items.

L'alpha de Cronbach varie entre 0 et 1: plus sa valeur est proche de 1, plus le test est consistant. Comme nous l'avons dit plut haut (cf. § 1.3.), cet indice n'est pas une estimation de la fidélité, il s'agit de l'estimation de la valeur minimale que pourrait atteindre la fidélité. En règle générale, on juge qu'elle est satisfaisante si l'alpha de Cronbach est supérieur à 0,70.

1.6. Les limites du modèle

La théorie classique des tests se heurte, dans la pratique, à deux problèmes. Tout d'abord, elle fournit des résultats dépendants de l'échantillon d'individus. Ainsi, un item considéré comme un item facile pour un échantillon d'individus, pourrait apparaître difficile si l'on changeait l'échantillon. Il suffit de penser à un item dont la difficulté serait estimée chez des élèves de lycée puis chez des élèves de collège. De même, un exercice de conjugaison sur le plus-que-parfait sera plus facilement réussi par un groupe d'élèves qui a étudié ce temps en classe que par un autre groupe qui ne l'a jamais étudié.

D'autre part, il apparaît délicat de comparer les résultats obtenus à deux tests contenant des items de difficultés différentes. Par exemple, si deux candidat/es passent deux tests différents, il sera difficile d'affirmer que le/la premier/e candidat/e a une compétence deux fois plus élevée si son score est deux fois plus élevé que celui du/de la deuxième candidat/e. Il s'agit là d'un problème lié au fait que la TCT repose sur l'utilisation du score total, qui n'est rien d'autre qu'un rang (tel/le candidat/e a obtenu plus de bonnes réponses que tel/le autre) et ne possède pas de métrique propre (l'écart entre le nombre de bonnes réponses ne peut être interprété de manière directe et absolue).

Malgré ces inconvénients, cette théorie continue d'être très largement utilisée par les chercheurs/ses et les concepteurs/rices de tests car elle offre de très bons outils pour l'analyse statistique des items. La reconnaissance des limites de ce modèle a conduit à de nouveaux développements et à l'élaboration de nouveaux modèles psychométriques, notamment les modèles issus de la théorie de réponse à l'item (TRI) dont l'objectif est d'estimer les caractéristiques des items (par exemple la difficulté, le pouvoir discriminant) ainsi que le niveau de compétence des individus. La partie suivante se propose de présenter le concept de l'IRT et différents modèles de réponse à l'item.

[7] Le coefficient alpha de Cronbach est aussi appelé KR20 quand il est utilisé avec des items dichotomiques.

II. Les modèles de réponse à l'item

Les modèles de réponse à l'item font suite aux travaux de deux grandes écoles, l'une danoise et l'autre américaine composées de mathématicien/nes et de psychométricien/nes des années 60 (Rasch, 1960; Birnbaum, 1968). Nous décrirons leur intérêt par rapport à la théorie classique des tests (pour une présentation plus détaillée, consulter Embretson et Reise, 2000). Puis nous décrirons le plus simple d'entre eux, le modèle de Rasch, avant d'évoquer le cas des modèles de réponse à l'item plus complexes (nombre de paramètres plus élevés, utilisation d'items polytomiques).

À la différence de la théorie classique des tests, les modèles de réponse à l'item vont permettre d'obtenir des paramètres d'item indépendants de ceux de la population sur laquelle ils ont été obtenus et inversement: ils possèdent ce que l'on appelle la qualité d'*objectivité spécifique*. Par ailleurs ils vont permettre d'obtenir conjointement une mesure de difficulté (pour les items) et d'habileté (pour les individus), interprétable sur la même échelle assimilée à une mesure continue du trait latent (la compétence linguistique par exemple). Ceci est valable dans le cas où l'on ne considère qu'une seule dimension (cas unidimensionnel) mais peut être généralisé au cas où plusieurs dimensions sont susceptibles de rendre compte des performances des individus (cas multidimensionnel). Par souci de clarté, nous nous limiterons volontairement à l'étude du cas unidimensionnel.

2.1. Présentation générale

Les modèles de réponse à l'item sont avant tout des modèles probabilistes, mais également des modèles qui permettent de fournir une mesure objective de la performance des individus, en relation avec une théorie clairement circonscrite tant dans son domaine que dans les moyens qui permettent de la vérifier.

L'essence même de ce type de modèle repose sur l'utilisation non pas de la réponse de l'individu à un item donné mais plutôt de la probabilité de réponse associée à cet item. Généralement, cette probabilité est la probabilité que le/la candidat/e donne une réponse correcte, mais l'on peut très bien modéliser la probabilité que le/la candidat/e choisisse l'une des réponses incorrectes (distracteur). Il convient dès lors de noter que la relation entre le trait latent et la réponse des candidat/es n'est plus considérée comme linéaire, à la différence de la théorie classique des tests (cf. §1.5). Cette probabilité de réponse est modélisée par ce que l'on appelle la courbe caractéristique de l'item (CCI): celle-ci représente alors la probabilité d'une réponse correcte à un item donné en fonction du niveau présumé (habileté) du/de la candidat/e.

2.2. Le modèle de Rasch

Le modèle de Rasch est le modèle de réponse à l'item le plus simple. Il permet de traiter exclusivement les réponses dichotomiques (codées vrai ou faux, par exemple), et se limite à un seul paramètre: la difficulté de l'item. On suppose également que les réponses s'expliquent par un seul trait latent (hypothèse d'unidimensionnalité) et que les réponses sont indépendantes les unes des autres. En

d'autres termes, l'ensemble des items qui composent le test permet de mesurer la même dimension, et la réponse d'un/e candidat/e à un item donné n'est pas influencée par sa réponse à un autre item. Nous reviendrons sur ces deux postulats de base après avoir brièvement décrit le principe général de fonctionnement du modèle de Rasch.

2.2.1. Caractérisation d'un item sur une échelle de mesure

Comme on l'a indiqué plus haut, le lien entre le trait latent et le score observé peut être modélisé par la courbe caractéristique de l'item (Figure 2). Plus précisément, le modèle de Rasch vise à modéliser l'écart entre l'habileté de l'individu et la difficulté de l'item. Rappelons que ces deux mesures sont exprimées sur la même échelle (exprimée le plus généralement en unités logit, par construction du modèle statistique) qui est supposée mesurer de façon continue le trait latent étudié.

Le recours à un tel modèle permet un *calibrage* qui conduit à une échelle commune aux individus et aux items. Tous les items mesurant une compétence particulière peuvent être positionnés le long d'une échelle, leurs positions et leurs espacements étant déterminés par le niveau de difficulté auquel ils correspondent. La réussite d'une personne à une partie de ces items peut être exprimée au travers d'une valeur qui correspond à un point quelque part sur cette échelle. Les items situés à gauche de la position de cette personne sont des items plus faciles pour lesquels elle a une probabilité de réussite supérieure à 50 %. Les items situés à droite sont des items plus difficiles pour lesquels la personne a une probabilité de réussite inférieure à 50 %. La difficulté d'un item est ainsi définie comme la valeur sur le trait latent pour laquelle l'individu a 50% de chance de répondre correctement.

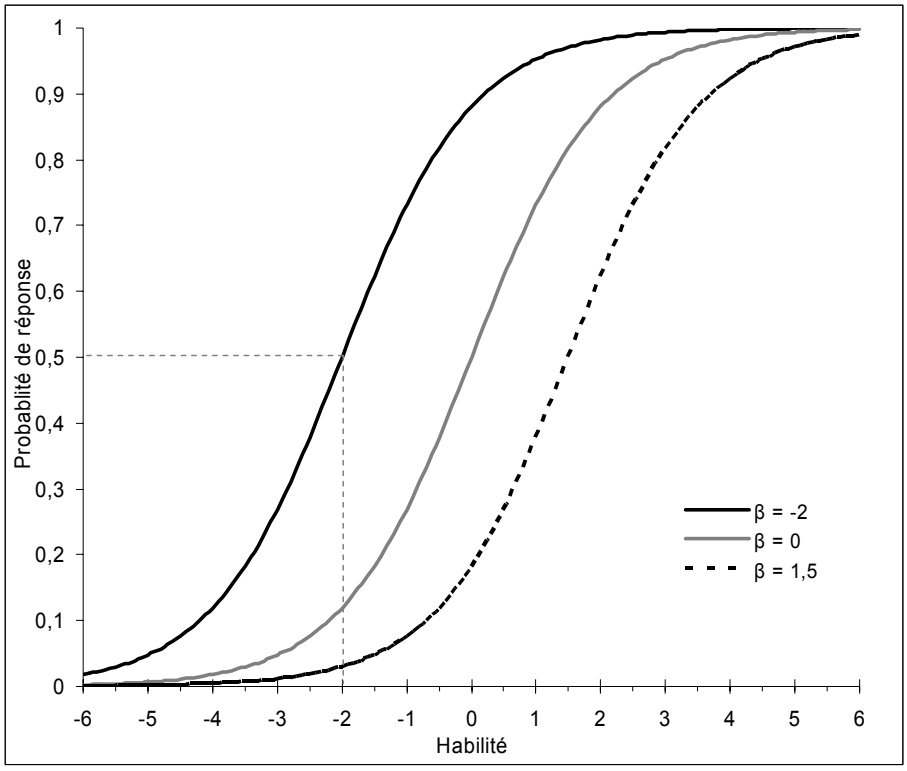


Figure 2 – **Courbes caractéristiques de trois items de difficulté croissante**

Comme on peut le voir dans la figure 2, la probabilité pour un/e candidat/e de répondre correctement (en ordonnées) augmente avec son habileté (en abscisses). Lorsque l'habileté est très inférieure à la difficulté, la probabilité de trouver la bonne réponse est pratiquement nulle. Inversement lorsque l'habileté est très supérieure à la difficulté, la probabilité de trouver la bonne réponse est proche de 1.

Ce modèle permet par conséquent d'ordonner les items en fonction de leur difficulté et d'évaluer le niveau des candidat/es en fonction des réponses correctes qu'ils/elles ont fournies. Naturellement, l'estimation de l'ensemble des paramètres définissant le modèle de Rasch pour un test donné (habileté de chacun des individus et difficulté de chacun des items) est une procédure assez lourde qui est, fort heureusement, confiée aux ordinateurs.

Ce type de modèle permet de fournir une indication précise concernant les caractéristiques intrinsèques d'un item, et ce de manière indépendante de l'échantillon d'individus auxquels les items ont été administrés. Chaque item apporte une certaine quantité d'information et celle-ci est maximale quand la difficulté est égale à l'habileté. On peut montrer que l'erreur de mesure, quant à elle, est inversement proportionnelle à la racine carrée de l'information. En d'autres termes, lorsque l'on cherche à situer précisément le niveau d'un/e candidat/e, on gagnera d'autant plus d'information en utilisant des items dont la difficulté est proche de son habileté réelle.

Étant donné que les items sont supposés indépendants, on va pouvoir additionner l'information apportée par chacun des items et calculer la courbe d'information du test. Celle-ci donne une idée de la gamme d'habiletés que l'on est susceptible d'évaluer de manière suffisamment précise avec le test considéré. Un test linguistique composé d'items de compréhension écrite ou orale de difficulté relativement peu élevée permettra de bien évaluer des candidat/es débutant/es dans cette langue. À l'inverse, un test linguistique similaire, composé uniquement d'items relativement difficiles, serait peu approprié pour évaluer ce même public de candidat/es: l'écart trop important entre le niveau réel des candidat/es (estimé au travers de leur habileté) et la difficulté des items rendrait très imprécis leur positionnement relatif sur une même échelle de mesure.

2.2.2. Hypothèses

Comme on l'a indiqué, le modèle de Rasch comme la plupart des modèles statistiques repose sur des hypothèses fortes: l'unidimensionnalité et l'indépendance locale

L*'unidimensionnalité* signifie qu'il n'y a qu'un seul et unique trait latent qui explique les réponses des individus. Cette hypothèse est difficile à vérifier en pratique. De nombreuses méthodes dans la littérature sont proposées. On trouvera des indices tels que l'indice de reproductibilité de Guttman ou l'alpha de Cronbach (cf. §1.5.2). Des tests statistiques ont également été créés (DIMTEST de Stout, par exemple). Les analyses factorielles et les analyses en composantes principales sont également largement utilisées. La plupart des méthodes développées ont pour objet,

sinon de prouver l'unidimensionnalité, du moins de démontrer la prédominance d'un seul et unique trait latent qui soit à même d'expliquer les performances observées.

La seconde hypothèse fondamentale est l'*indépendance locale*. Cette hypothèse stipule que, pour un niveau d'habilité donné, les réponses aux items sont indépendantes. En d'autres termes, pour l'ensemble des candidat/es possédant le même niveau de compétence, la réponse à un item n'a aucune influence sur la réponse à un autre item.

Ces deux hypothèses sont liées dans le sens où si le test est unidimensionnel alors l'hypothèse d'indépendance locale est vérifiée. Mais généralement la réciproque n'est pas vraie.

On notera enfin qu'une troisième hypothèse est également mise en avant par certains psychométriciens: il s'agit de la constance de la discrimination, c'est-à-dire de la capacité d'un item à bien discriminer les candidat/es selon leurs compétences et ce quelle que soit leur habileté. Ce point sera développé plus avant lors de la présentation des modèles plus complexes.

2.2.3. Propriétés

On a vu qu'une des propriétés d'un instrument de mesure était sa capacité à délivrer des scores reproductibles (cf. §1.3). Cette caractéristique est appelée propriété d'*objectivité spécifique*. Elle permet de comparer deux personnes indépendamment de l'instrument de mesure (avec des tests différents). L'une des conséquences de cette propriété est que pour un ensemble d'individus d'habileté différente, les difficultés des items demeurent les mêmes.

Par ailleurs, le score total possède ce que l'on appelle la propriété d'*exhaustivité*: toute l'information sur le trait latent est contenue dans le score. Ceci signifie que lorsque l'on considère qu'il n'y a qu'une seule dimension latente qui sous-tend les performances observées, la connaissance du score observé chez un ensemble de candidat/es suffit entièrement pour caractériser cette dimension et, par là, attribuer un niveau d'habileté propre à chaque personne et un niveau de difficulté propre à chaque item. Cette propriété est vraie pour les modèles de la famille du modèle de Rasch. Cette propriété est capitale pour certaines méthodes d'estimation des paramètres. Elle a, en revanche, une conséquence qui n'est pas intuitive en première approche. Deux candidat/es qui ont un même score global mais qui auront répondu correctement à des items différents auront la même estimation d'habileté. Si l'on exclut les scores parfaits (ensemble de réponses correctes ou incorrectes), on peut estimer 79 habiletés différentes à partir d'un test composé de 80 items. À chaque ensemble de réponses correctes correspond ainsi une habileté unique, nonobstant le pattern de réponses.

2.2.4. Adéquation entre les données réelles et les données modélisées

L'utilisation d'un modèle de réponse à l'item vise à fournir une mesure avec le moins d'erreurs possible (il s'agit du point de vue du statisticien), ou de manière équivalente à fournir un ensemble de prédictions ne s'écartant pas trop du modèle théorique supposé dès le départ (on parle de *construct*, et il s'agit plutôt du point de vue de l'expert en pédagogie). Il apparaît donc crucial de vérifier l'adéquation entre les données réelles et les données modélisées.

À l'image des tests portant sur la vérification de l'unidimensionnalité du test, les indices déterminant l'adéquation entre les données réelles et les données modélisées sont très nombreux. Ils dépendent étroitement du logiciel utilisé. Seuls deux indices simples et couramment employés seront abordés: l'*outfit*[8] et l'*infit*[9]. Ces indices sont calculés à partir de ce que l'on appelle les résidus ou écarts entre données observées et données prédites par le modèle: les résidus sont ainsi estimés par la différence entre la réponse observée (binaire dans le cas du modèle Rasch qui permet de modéliser des réponses dichotomiques) et la probabilité de réponse du/de la candidat/e à l'item. L'*outfit* permet de détecter les réponses improbables des individus, par exemple un individu de niveau élevé qui échouerait à un item facile. L'*infit* est pondéré par l'information fournie par l'item; ceci permet de minimiser l'information apportée par une réponse inattendue de la part d'un/e candidat/e pour lequel l'item est inadapté. Par exemple, un item trop facile auquel un/e candidat/e de bon niveau échouerait serait moins pris en compte dans l'évaluation du niveau du/de la candidat/e. D'une certaine manière, ces indicateurs permettent de "filtrer" les données aberrantes qui peuvent résulter, par exemple, d'une erreur d'inattention de la part du/de la candidat/e au moment de fournir sa réponse.

Ces indices prennent la valeur 1 si l'ajustement est "parfait". On retiendra que les valeurs que doivent prendre ces indices pour indiquer un bon ajustement sont empiriques et varient selon les auteur/es. Par consensus, un bon ajustement est associé à des valeurs d'*outfit* situées entre 0,5 et 1,5 et à des valeurs d'*infit* entre 0,8 et 1,2.

2.3. Modèles plus complexes

Le modèle de Rasch est très simple. Il est notamment très réducteur concernant le comportement de l'item qui n'est caractérisé que par un paramètre de difficulté. En outre, il ne s'applique qu'aux items dichotomiques. Pour ces raisons, d'autres modèles plus complexes ont été développés dont quelques exemples sont présentés dans les paragraphes suivants.

2.3.1. Modèles à 2 et 3 paramètres

Dans le cas du modèle à 2 paramètres, appelé encore modèle de Birnbaum, un paramètre supplémentaire est ajouté: il s'agit de la discrimination. Ce paramètre va permettre de moduler la pente de la courbe caractéristique de l'item (Figure 3), c'est-à-dire la capacité de l'item à discriminer plus finement des individus possédant des niveaux d'habileté proches. Dans ce cas, plus la pente est élevée (c'est-à-dire plus la partie centrale de la courbe apparaît orientée verticalement), plus l'item est discriminant.

Une des objections couramment faite au modèle de Rasch est la constance de la discrimination, usuellement fixée à 1. En effet la discrimination est généralement corrélée positivement au coefficient point bisérial utilisé dans la Théorie classique

[8] Moyenne non pondérée des carrés des résidus.
[9] Moyenne pondérée des carrés des résidus.

des tests (§1.5.1). En pratique, il est pratiquement impossible d'obtenir une grande homogénéité des coefficients des points bisériaux des items. C'est pour cette raison que certains auteurs préconisent l'utilisation de ce modèle plutôt que du modèle de Rasch.

Dans le modèle à 3 paramètres, on ajoute un paramètre parfois appelé "*guessing*", "conjoncture" ou "pseudo-chance", et qui constitue une manière de rendre compte de la probabilité pour le/la candidat/e de deviner une bonne réponse parmi plusieurs réponses possibles. Dans ce cas, même pour les plus bas niveaux de l'échelle d'habileté (vers la gauche, sur l'axe des abscisses), c'est-à-dire pour les candidat/es dont le niveau de compétence est le plus faible, la probabilité de répondre correctement n'est pas nulle, quelle que soit la difficulté de l'item.

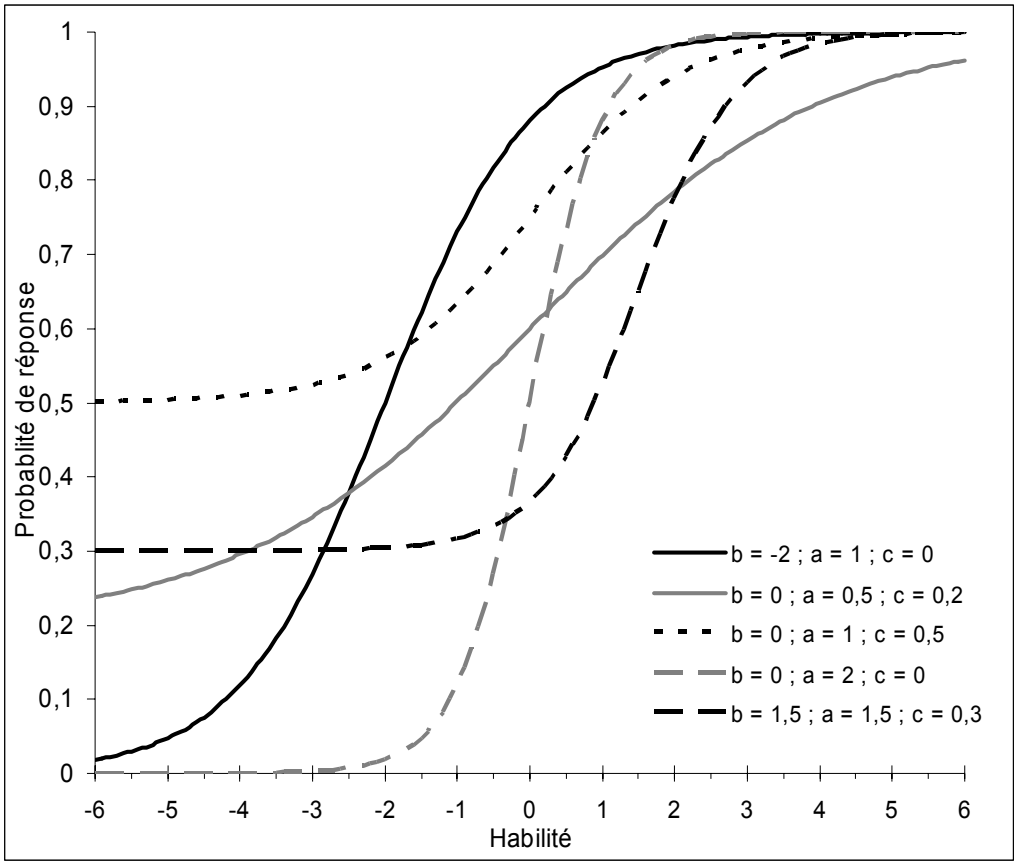


Figure 3 – **Modèle de Birnbaum: exemples de courbes caractéristiques d'items**
Le paramètre *b* correspond à la difficulté de l'item, le paramètre *a* à la discrimination et le paramètre *c* au paramètre de "pseudo-chance".

Avec ce type de modèles, on perd la propriété d'exhaustivité du score. Il n'est plus possible de calculer l'habileté simplement à partir des difficultés des items et du score des candidat/es. Il est alors nécessaire, pour chaque test, de calculer systématiquement, par une procédure assez lourde en temps de calcul, l'habileté du/de la candidat/e à partir de son "motif" de réponse. Enfin, il est nécessaire d'avoir un nombre plus important de réponses pour estimer les paramètres du modèle: de 600 à 800 individus au minimum pour le modèle à 2 paramètres et plus de 1000 individus

pour le modèle à trois paramètres, alors que pour utiliser le modèle de Rasch on peut se contenter de 250 individus.

2.3.2. Modèle pour items polytomiques

Dans le cas où le modèle est polytomique, la réponse à l'item n'est plus binaire et plus de deux réponses sont possibles. Le modèle de Rasch a donc été adapté pour ce type de situation, notamment quand la réponse est graduée, c'est-à-dire pour des échelles dite échelles de Lickert. Les réponses observées sur une échelle de ce type à 4 modalités de réponse peuvent être codées de la manière suivante: faux (0), faible (1), moyen (2), élevé (3), exact (4). Les modèles permettant de prendre en compte ce type de données dépendent de la façon d'interpréter le score associé aux réponses des candidat/es en relation avec le codage de cette échelle.

Le seul modèle qui sera évoqué ici est le modèle à crédit partiel (Masters, 1982) dont on trouvera un modèle d'application dans la partie suivante. Il est applicable lorsque les réponses sont ordonnées ou graduées. Contrairement à d'autres types de modèles élaborés pour traiter ce type de données (par exemple, les réponses graduées de Samejima, 1969), le modèle à crédit partiel est dérivé du modèle de Rasch et en conserve donc les propriétés: objectivité spécifique et exhaustivité du score. Tout comme dans le cas du modèle de Rasch, une difficulté est associée à chacune des modalités de réponse. Cette difficulté correspond au point pour lequel une réponse devient majoritaire par rapport à l'autre. En revanche, les courbes caractéristiques des items s'interprètent différemment car elles indiquent désormais la probabilité de fournir l'une des modalités de réponse (conditionnellement aux autres choix de réponse possibles) en fonction de l'habileté.

2.4. De la Théorie classique des tests aux modèles de réponse à l'item

Le modèle de réponse à l'item vient en complément de la théorie classique des tests. Si cette dernière permet de pointer les dysfonctionnements des items, elle ne permet pas de mesurer à la fois l'habileté des candidat/es et la difficulté des items. Or la relation qui lie ces deux grandeurs est la base même du comportement de l'individu. C'est parce qu'un/e candidat/e possède un bon niveau dans une langue étrangère, par exemple, qu'il/elle est à même de répondre correctement à des items dont le degré de difficulté est relativement peu élevé. C'est donc bien cette possibilité de modéliser la dépendance intrinsèque entre niveau d'habileté d'un individu et niveau de difficulté des items (dans le modèle de Rasch, cette dépendance ou relation fonctionnelle est objectivée par l'écart entre ces deux grandeurs) qui rend les modèles de réponse à l'item si attractifs et incomparablement plus utiles que la Théorie classique des tests lorsque l'on souhaite positionner des candidat/es les un/es par rapport aux autres à l'aide d'un test standardisé.

En revanche, comme tout modèle statistique, c'est-à-dire comme toute tentative de réduction de la réalité à un ensemble de processus objectivables au travers de mesures physiques, ce type de modèle possède des limitations. On peut en mentionner deux principales. La première concerne l'adéquation du modèle aux données: afin de fournir des estimations précises du niveau d'habileté, il semble évident que le modèle doit permettre de fournir des prédictions correctes et "se

conformer" aux données observées. En second lieu, les hypothèses d'application du modèle doivent être vérifiées. En effet, comment pourrait-on affirmer qu'un test mesure bien ce qu'il est censé mesurer, au travers de cet outil de mesure objectif que constitue le modèle de réponse à l'item, si en réalité les réponses observées ne peuvent être expliquées qu'à l'aide d'un faisceau de traits latents (violation du principe d'unidimensionnalité)?

III. DELF

Le DELF, Diplôme d'études en langue française et le DALF, Diplôme approfondi de langue française sont des diplômes du Ministère de l'éducation nationale, de l'enseignement supérieur et de la recherche. Ils s'adressent à des publics non francophones souhaitant faire valider leur connaissance du français. Ils existent depuis 1985 (arrêté du 22 mai 1985) et depuis 2005, ils sont harmonisés sur les 6 niveaux de l'échelle de compétence en langue définie dans le CECR (arrêté du 7 juillet 2005). Le DELF se compose de 4 diplômes indépendants correspondant aux niveaux A1, A2, B1 et B2 du CECR et le DALF de deux diplômes de niveaux C1 et C2[10]. Pour chaque niveau du DELF, quatre compétences sont évaluées: la compréhension orale, la production orale, la compréhension des écrits et la production écrite. Il existe pour le DELF une version "tous publics" et une version "junior et scolaire" destinée à des candidat/es scolarisé/es dans le secondaire. Régulièrement, des analyses psychométriques sont réalisées sur des épreuves de compréhension afin d'étudier la qualité fonctionnelle des items. Les psychométricien/nes étudient aussi le degré d'objectivité des correcteurs/trices pour certaines épreuves de production écrite. L'objectif à long terme est de pouvoir justifier de l'équité des épreuves du DELF.

Les épreuves de compréhension écrite et orale prennent le format d'un test composé de plusieurs exercices. Chaque exercice comprend plusieurs items portant sur un document ou un bloc de documents. La structure d'une épreuve est standardisée par le nombre, le type et le barème des exercices. Suivant le document sur lequel porte l'exercice, les concepteurs/trices d'items peuvent choisir le nombre et la forme des items. Les items peuvent prendre des formes assez variées: questions à choix multiples, questions ouvertes et exercices d'appariement.

Nous présentons, dans cette partie, l'analyse psychométrique de quatre items. Ces items proviennent d'une épreuve de compréhension écrite d'un sujet DELF niveau B2 version "scolaire et junior". Les items 1 et 2 portent sur le document 1 et les items 3 et 4 sur le document 2 (Figure 5). Les items 1 et 3 sont deux questions à choix multiples composées d'une clé C et de deux distracteurs A et B. Les items 2 et 4 sont deux questions ouvertes, le premier est noté sur 3 points et le deuxième sur 2 points. Nous commencerons par étudier les qualités fonctionnelles des items avec la TCT et puis nous terminerons par le modèle PCM (cf. 2.3.2. Modèle pour items

[10] A1, niveau élémentaire; A2, niveau élémentaire avancé; B1, niveau intermédiaire; B2, niveau intermédiaire avancé; C1, niveau supérieur; C2, niveau supérieur avancé.

polytomiques, § 0) avec lequel nous estimerons le niveau de difficulté des items et étudierons la qualité de la structure interne des items polytomiques. Ce modèle de réponse à l'item est utilisé car le DELF est composé d'items dichotomiques et polytomiques dont les modalités sont ordonnées (Bode, 2004; Linacre, 2004). Les modalités d'un item correspondent dans le cas du DELF aux différents scores qu'un individu peut obtenir à l'item. Elles sont ordonnées par ordre croissant de difficulté: un individu qui obtient 2 points à un item est censé avoir un niveau de compétence supérieur à celui qui n'obtiendra qu'1 point au même item.

Document 1: **SPORTIFS EN HERBE**

Qu'il court, nage, saute, tape dans un ballon ou pédale, l'enfant tire de multiples bénéfices – tant sur le plan physique que mental – de la pratique d'un sport. Ce dernier lui donne de l'assurance, modèle son corps, forge son esprit, facilite sa socialisation, développe son goût de l'effort. Et avec les diverses possibilités qui sont aujourd'hui proposées, il est rare qu'un enfant ne trouve pas une activité qui lui convienne. Le rôle des parents est ici primordial mais délicat: inciter sans forcer et savoir encourager l'athlète en herbe qui fait preuve de certaines qualités sans pour autant le lancer dans la course aux médailles.[…]

Contrairement à certaines rumeurs tenaces, le sport ne fait pas grandir. Et il est aujourd'hui prouvé qu'il ne bloque pas non plus la croissance. En revanche, certaines activités, comme la gymnastique, pratiquées de façon intensive peuvent ralentir cette dernière. Mais toutes les adorables petites gymnastes alliant grâce, souplesse et énergie finiront par avoir une taille parfaitement normale, seulement avec quelques années de retard. […] Leur squelette est-il plus fragile pour autant? Pour répondre à cette question, le Dr Michel Binder a une explication très imagée: "Chez les adultes, les os peuvent être comparés à des murs de béton dans lesquels sont fixées des ficelles soutenant les muscles. Lorsque l'on tire trop fort sur la ficelle, elle casse. L'adulte se fait ainsi des claquages, des élongations ou des tendinites. Pour un enfant, les os sont plutôt des murs de plâtre dans lesquels sont enfoncés des clous, c'est-à-dire des cartilages de croissance, qui retiennent les ficelles. Lorsque l'on tire trop fort sur la ficelle, ce sont les clous qui cèdent. La maladie de la sur-sollicitation chronique du surentraînement chez l'enfant est essentiellement liée au cartilage de croissance. On peut ainsi observer des fractures du cartilage ou des arrachements des points d'ancrage." D'où la nécessité de respecter la douleur, qui est le signal d'alarme le plus efficace.

Côté alimentation, il faut savoir que les besoins énergétiques de l'enfant – sportif ou non – sont voisins de ceux de l'adulte. À 5 ans, l'apport doit être en effet de 1 500 kilocalories par jour. À 9 ans, de 2 200, chez les garçons comme chez les filles. À l'adolescence, les premiers peuvent manger quotidiennement jusqu'à 3 000 kilocalories, les secondes 2 400. Et l'activité physique accroît les besoins, en fonction de son intensité. L'alimentation doit évidemment être équilibrée, à la fois au cours de chacun des trois repas – or 10 à 15 % des enfants et adolescents ne prennent pas de petit déjeuner – et sur l'ensemble de la journée. Mais, pendant la croissance, la

consommation de protéines doit être assez importante et, plus encore chez le jeune sportif car elles jouent un rôle essentiel dans la "construction et l'édification" de l'organisme. […] Et, comme tout sportif, l'enfant doit boire suffisamment, notamment pendant et après l'effort. D'autant plus qu'il transpire souvent beaucoup. Or c'est rarement le cas. "La peur de ne pas savoir où uriner à un âge où l'on devient très pudique l'incite à limiter sa consommation d'eau", regrettent de nombreux médecins du sport. […] Pour conserver "une âme saine dans un corps sain", les enfants pratiquant un sport doivent donc avoir une bonne discipline de vie et alimentaire.

Mais il ne faut pas oublier que cet édifice est fragile. Notamment au moment de l'adolescence, quand l'opposition devient la règle, les poussées de croissance perturbent le geste sportif – au risque de dégoûter les plus "accros" – et l'attrait pour le sexe opposé déplace les centres d'intérêt.

Anne Jeanblanc, *Le Point*, 17-06-03

Item 1 (1 point)
Lorsqu'ils s'entraînent trop, les enfants:
A. ☐ risquent moins de se blesser que les adultes car ils sont plus souples.
B. ☐ se blessent plus souvent mais moins gravement que les adultes.
C. ☒ se font des blessures de nature différente de celles des adultes.

Item 2 (3 points)
Quels facteurs sont déterminants dans l'évaluation des besoins caloriques de l'enfant?
L'âge (1 point), le sexe (1 point), l'intensité de l'activité sportive (1 point).

Document 2: **SURDOUÉS: TROP INTELLIGENTS POUR ÊTRE HEUREUX**

Depuis qu'elle a créé l'Association française pour les enfants précoces (Afep) en 1993, Sophie Côte a reçu toutes sortes de sollicitations des médias. Elle est d'ailleurs habituée à y répondre favorablement, afin de faire connaître au grand public la cause de ces petits qui sont "trop tout". Trop éveillés, trop curieux, trop exigeants, trop angoissés, et parfois trop mauvais en classe, où ils peuvent s'ennuyer à périr sous le regard exaspéré des enseignants. […]

Jean-Charles Terrassier est encore plus ancien dans le métier. Ce psychologue clinicien de Nice a, dès 1971, exploré le fonctionnement de ces têtes drôlement faites qui représentent tout de même plus de 2 % de la population. Lui qui s'est battu pour faire admettre l'existence des surdoués, pour encourager leur détection et éviter du même coup leur marginalisation a écrit des livres, multiplié les colloques, les interviews dans les journaux. Mais il n'en est pas revenu quand une grande chaîne de télévision publique a décidé de bâtir une émission autour d'un enfant qui serait testé "en direct" et dont les téléspectateurs connaîtraient le QI juste avant le générique de fin, au terme d'un suspense savamment orchestré...

Encore ignoré par l'institution scolaire, toujours nié par les tenants d'une psychanalyse pure et dure, l'enfant précoce, depuis quelque temps, est l'objet de

fantasmes collectifs de plus en plus tenaces. Dans une société où la performance est devenue une valeur en soi et l'enfant idéal un objet de désir narcissique, le mythe du petit génie envahit les têtes […]

De très rares collèges publics et des établissements privés en nombre plus important ont ouvert des classes spécifiques pour les enfants précoces, qui ont la possibilité de faire deux années en une. Ainsi du collège privé Saint-Louis, au Mans, où les professeurs, recrutés sur la base du volontariat, reçoivent une formation spécifique. Cet établissement, qui poursuit l'expérience depuis huit ans sans faire de tapage, doit refuser la plupart des demandes. "Pour l'entrée en quatrième, nous avons 10 places et 200 demandes", explique le directeur, avant de préciser que la plupart de ces enfants sont, d'une manière ou d'une autre, en difficulté scolaire, et qu'il ne pratique donc aucune sorte d'élitisme qui améliorerait ses résultats.

Mais, à côté d'institutions comme celle-ci, intéressées par l'épanouissement des enfants et non par leurs performances, d'autres ont vu dans cette population très spéciale un vivier propre à améliorer leurs résultats au brevet et au bac, tout en se situant sur un créneau porteur. […]

Dans l'imagerie populaire, l'enfant surdoué n'est pas à plaindre, c'est presque un nanti. Personne ne voit ses fragilités, et éventuellement ses souffrances. Résultat, plus d'un enfant sur trois n'a pas son baccalauréat, parce qu'il est rejeté du système scolaire avant le lycée […] Faute de recherches, faute de solutions pédagogiques adaptées, les surdoués risquent de demeurer ces objets de fantasme qui inspirent même les sectes, puisque des mouvements comme les Enfants indigo, avec des arguments à dormir debout, parviennent à convaincre des parents désemparés que leurs enfants sont des envoyés d'une autre civilisation et qu'il convient de les élever "autrement", loin des psys. Entre propositions délirantes et indifférence des institutions, les familles ont bien du mérite à se frayer un chemin vers la simple reconnaissance d'une particularité qui fascine et indispose.

Sophie Coignard, *Le Point*, 17-03-05

Item 3 (1 point)
Vrai ou faux? Cochez la case correspondante:
Les professeurs qui travaillent dans ces classes spécifiques reçoivent un salaire plus élevé.
A. ☐ Vrai.
B. ☐ Faux.
C. ☒ On ne sait pas.

Item 4 (2 points)
Que signifie Sophie Coignard lorsqu'elle écrit: "Dans l'imagerie populaire, l'enfant surdoué n'est pas à plaindre, c'est presque un nanti."?
Pour la plupart des gens (1 point), l'enfant surdoué a une situation enviable (1 point).

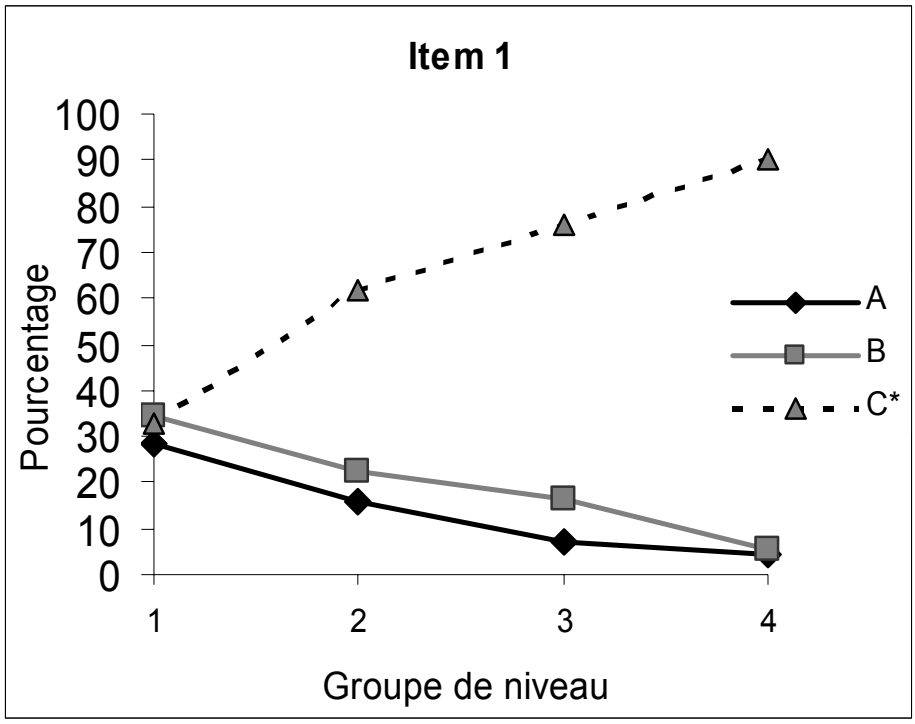
Figure 4 – **Documents et items extraits d'une épreuve du DELF junior et scolaire niveau B2**

3.1.1. Théorie classique des tests

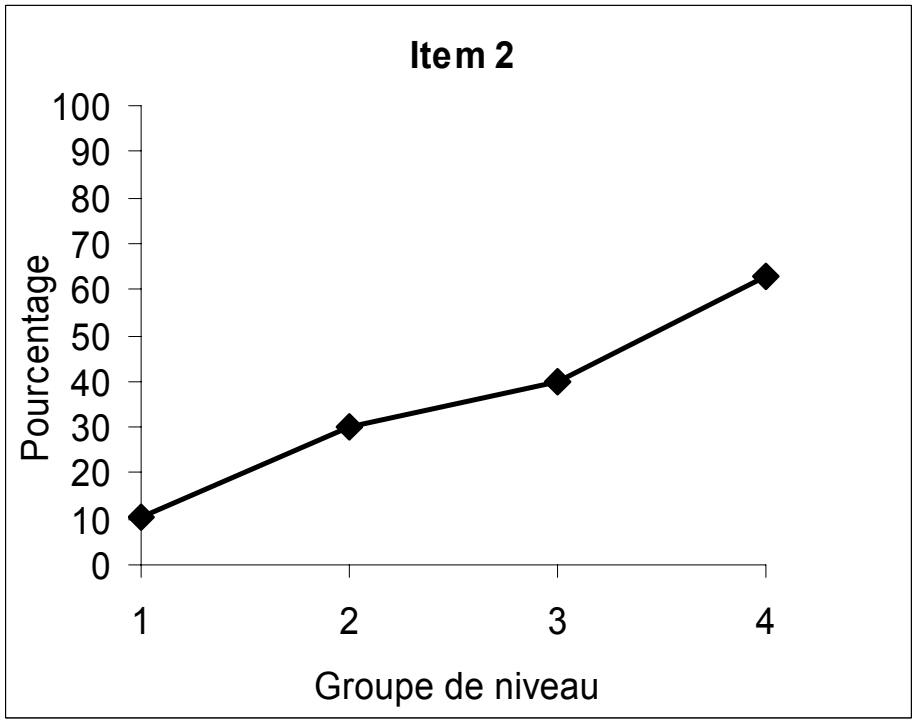
Les items 1 et 3 sont assez faciles, ils sont respectivement réussis par 66 % et 62 % des candidat/es (Tableau 1). Les deux autres items sont assez difficiles, le taux de réussite est de 37 % pour l'item 2 et 30 % pour l'item 4. Les deux questions ouvertes sont donc plus difficiles. Les indices de discrimination item-reste des items 1, 2 et 4 sont supérieurs à 0,30 et les "indices de discrimination" des distracteurs de l'item 1 sont négatifs (Tableau 1), ce qui suggère que ces items permettent de bien distinguer les candidat/es suivant leurs performances au test. Leur pouvoir discriminant est observable aussi graphiquement (Figure 6): le taux de réponses correctes augmente fortement en fonction du niveau des candidat/es. Par exemple pour l'item 1, on passe d'un taux de réussite de 35% pour le groupe 1 à un taux de 90 % pour le groupe 4 (Figure 6.a, courbe C) et les proportions de réponses incorrectes (courbes B et C) diminuent en fonction du niveau des candidat/es. En ce qui concerne l'item 3, il présente un certains nombre de dysfonctionnements. Son indice de discrimination item-reste est très proche de zéro, il est égal à 0,06 (Tableau 1), l'item ne permet donc pas de différencier les candidat/es selon leur compétence. Le graphique des proportions de réponses des candidat/es (Figure 5c) illustre bien ce problème: le taux de réponse correcte n'augmente pas continuellement du groupe 1 au groupe 4. Il passe de 60 % pour le groupe 1 à 50 % pour le groupe 2 alors que ce dernier devrait réussir cet item plus facilement que le groupe 1. Ensuite, le taux de réussite augmente jusqu'au groupe 4 et atteint les 71 %. En ce qui concerne les distracteurs, les proportions de réponses incorrectes, B et C, ne diminuent pas en fonction du niveau des candidat/es et le distracteur B est plus souvent choisi que le distracteur C (Tableau 1, figure 4): à la question "Vrai ou faux? Les professeurs qui travaillent dans ces classes spécifiques reçoivent un salaire plus élevé.", 33 % des candidat/es répondent que c'est faux (distracteur B) et 5 % pense que c'est vrai (distracteur C). Les qualités fonctionnelles des items 1, 2 et 4 sont très bonnes, par contre l'item 3 ne fonctionne pas correctement, il ne sera donc pas intégré à la suite de l'analyse.

Item	**Clé**	**Pourcentages**			**Score maximal**	**Score moyen**	**Difficulté**	**Indices de discrimination**				
		A	*B*	*C*				*Item-test*	*Item-reste*	*A*	*B*	*C*
1	C	13	19	66	1	0,66	66%	0,43	0,37	-0,26	-0,18	0,37
2					3	1,11	37%	0,48	0,32			
3	C	5	33	62	1	0,62	62%	0,13	0,06	-0,10	0	0,06
4					2	0,60	30%	0,46	0,35			

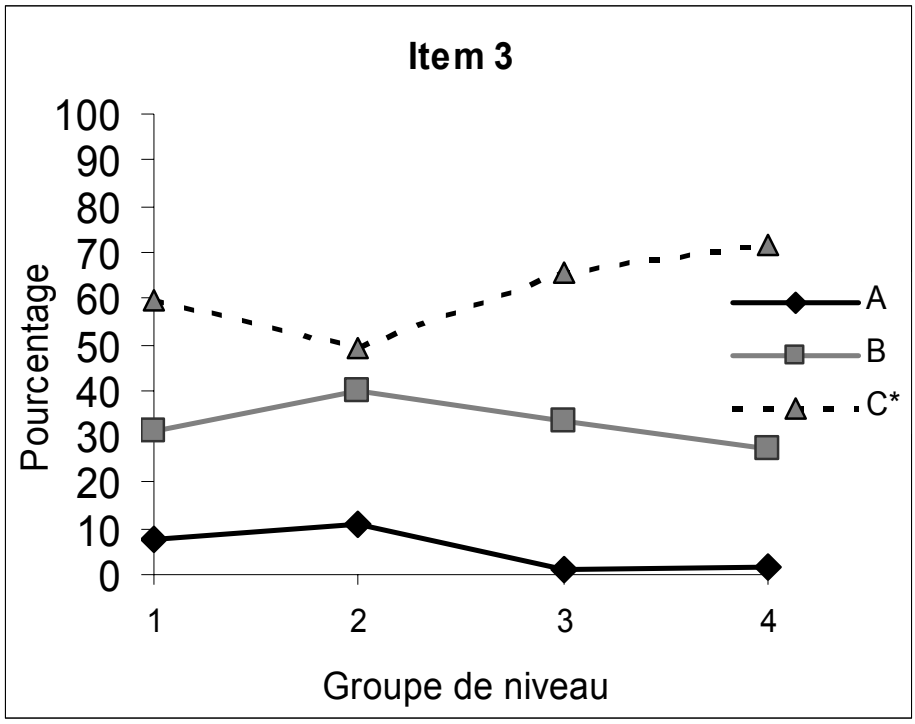
Tableau 2. **Résultats statistiques concernant les items obtenus avec la TCT**


Item 1
100
90
80
70
60
50
40
30
20
10
0
Pourcentage
1
2
3
4
Groupe de niveau
A
B
C*

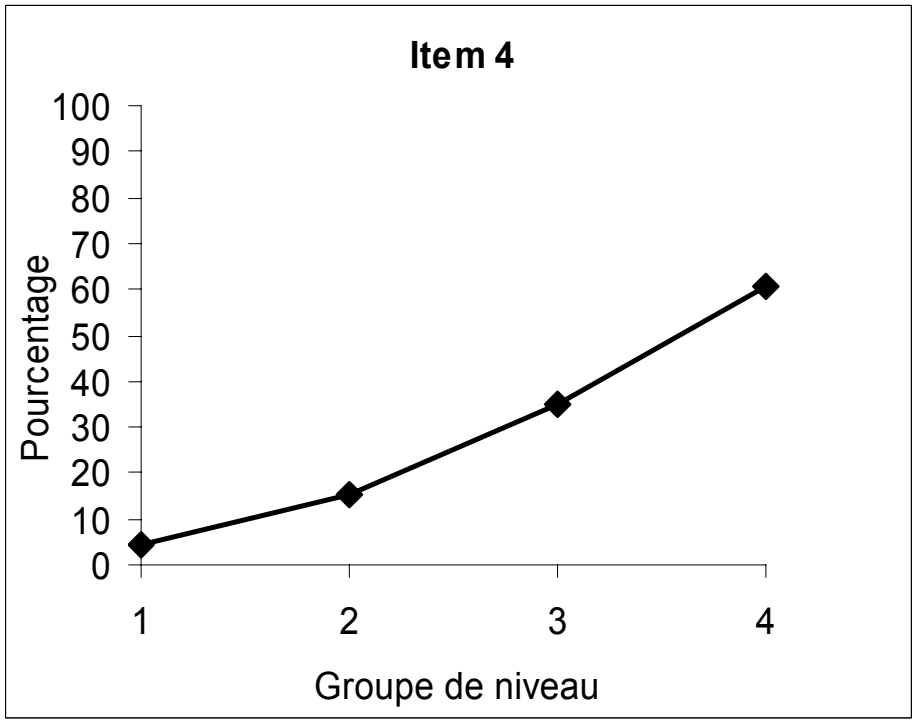
(a)


Item 2
100
90
80
70
60
50
40
30
20
10
0
Pourcentage
1
2
3
4
Groupe de niveau

(b)



(c)



(d)

Figures 5 – **Distribution des réponses des candidat/es à un item en fonction de leur niveau**

3.1.2. Modèle à crédit partiel

Pour estimer le niveau de difficulté des items, nous utilisons le modèle à crédit partiel (PCM) présenté dans la section précédente (§2.3.2). Comme on l'a vu, les modèles de réponses à l'item ont l'avantage, contrairement à la TCT, de fournir des estimations indépendantes de l'échantillon sur lequel on travaille. La difficulté des

items est estimée en logit (unité de mesure), et varie généralement entre -3 et +3 (plus la valeur de la difficulté est élevée plus l'item est jugé difficile). La compétence des candidat/es est aussi estimée en logit sur la même échelle que les items.

On constate que l'ordre de difficulté des items est identique à celui obtenu précédemment avec la TCT: l'item 1 est nettement plus facile que les items 2 et 4, l'item 4 est un peu plus difficile que l'item 2 (Tableau 2). L'erreur de mesure de la difficulté est très faible pour les items 2 et 4, un peu plus élevée pour l'item 1: l'estimation de la difficulté est donc plus précise pour les items 2 et 4. L'infit et l'outfit des trois items sont très proches de 1, ce qui indique que le modèle permet de prédire correctement les réponses des candidat/es pour ces items.

Les items 2 et 4 étant polytomiques, nous analysons chacune de leurs modalités. Les modalités d'un item sont censées être ordonnées, elles indiquent le degré de réussite à l'item (plus leur valeur est élevée, plus la réponse à l'item est correcte). Si elles ne sont pas ordonnées, l'item ne peut pas correctement évaluer le trait latent, ici la compétence langagière. Pour l'item 2, la compétence moyenne des candidat/es (exprimée en logit) augmente en fonction de la valeur de la modalité (Tableau 3), ce n'est pas le cas pour l'item 4, elle est de 0.78 logit pour la modalité 1 et de 0.71 pour la modalité 2. Les candidat/es qui ont eu 1 point à cet item ont en moyenne un niveau légèrement supérieur à celles/ceux qui ont eu 2 points. Cette anomalie peut indiquer que les modalités de l'item 4 ne sont pas ordonnées, et que cet item ne permet pas d'évaluer correctement la compétence langagière des candidat/es. Notons que leur compétence moyenne pour la modalité 1 est calculée sur un petit nombre de candidat/es, 15 sur 257. Pour l'item 2, l'infit et l'outfit associés à la modalité 1 sont élevés, 1,68 et 1,88, ce qui indique que l'on observe beaucoup de réponses improbables pour cette modalité. Par exemple, un individu qui aurait dû avoir 2 points à cet item n'en n'a eu qu'un. Pour finir l'analyse des modalités des items, nous étudions les seuils de transition11 entre les modalités voisines, c'est-à-dire entre les modalités 0 et 1 (seuil 1), entre les modalités 1 et 2 (seuil 2), et ainsi de suite. Le seuil de transition entre les modalités 0 et 1 d'un item correspond au niveau de compétence d'un/e candidat/e qui aurait autant de chance d'obtenir l'une ou autre de ces modalités. Il en va de même pour les autres seuils. Les seuils de transition des items 2 et 4 ne sont pas ordonnés par ordre croissant en fonction des valeurs des modalités (Tableau 3). Cela n'implique pas que leurs modalités soient désordonnées, mais qu'une ou plusieurs modalités ont une probabilité d'être observées toujours inférieure aux autres modalités. Les courbes caractéristiques, CCI (Figure 6), de ces items permettent de voir rapidement de quelles modalités il s'agit. Les CCI illustrent le lien qui existe entre la probabilité d'observer une modalité et le trait latent mesuré (i.e. la compétence langagière des candidat/es). Les courbes des modalités 1 et 2 de l'item 2 se situent toujours en dessous des courbes des modalités 0 et 3 : leurs probabilités d'être observées est donc toujours inférieure à celle des modalités 0 et 3. Ainsi la probabilité qu'un candidat obtienne 0 ou 3 points à l'item 2 est plus forte que la probabilité qu'il obtienne 1 ou 2 points. La courbe de la modalité 1 de l'item 4 est elle aussi située en dessous des courbes des modalités 0 et 2. Étant donné que les

[11] Les seuils de transition sont des paramètres du modèle à crédit partiel.

items 2 et 4 sont assez difficiles, la modalité 0 est en effet très souvent observée : 49% des candidats ont eu 0 points à l'item 2 et 67% ont eu 0 points à l'item 4 (Tableau 3).

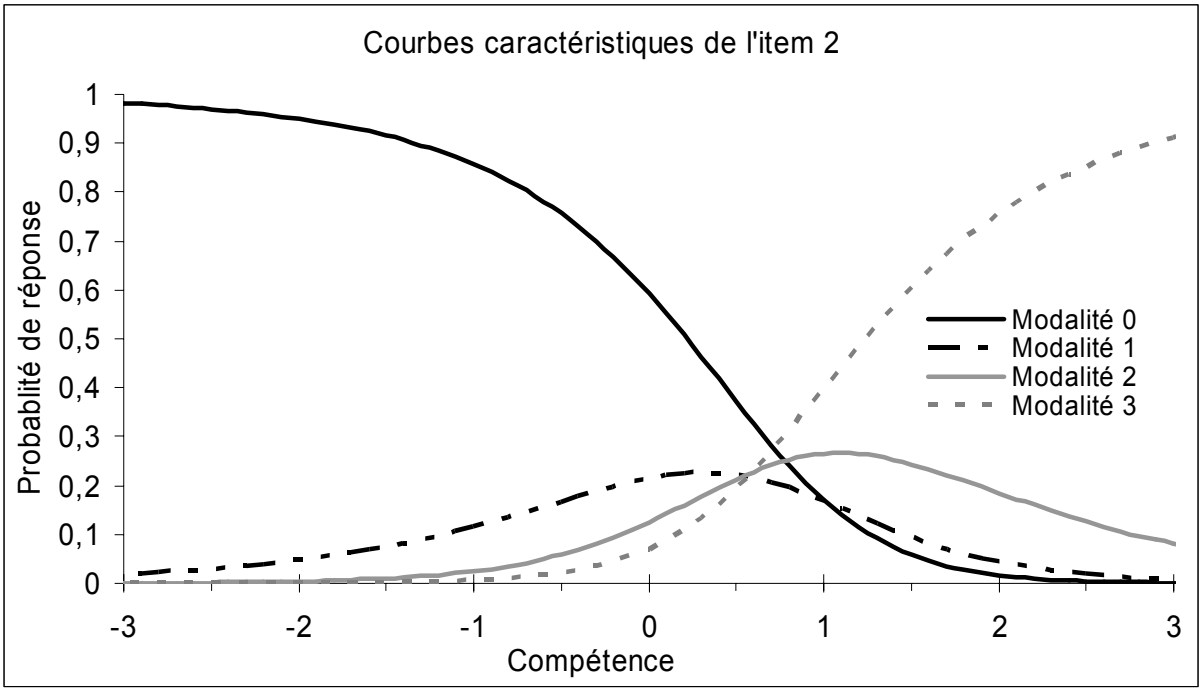
En conclusion, l'analyse psychométrique des items nous a permis d'étudier leur fonctionnement. L'item 3 ne permet pas de distinguer les candidat/es en fonction de leur compétence langagière, c'est pourquoi nous n'avons pas estimé sa difficulté avec le modèle de réponse à l'item PCM. L'item 4 pourrait ne pas évaluer correctement le niveau de compétence des candidat/es car celles/ceux qui réussissent partiellement l'item ont en moyenne un niveau légèrement supérieur à ceux/celles qui le réussissent complètement. L'item 2 fonctionne à peu près correctement, mais nous observons, parmi les candidat/es qui ont eu 1 point, beaucoup d'individus pour lesquels le modèle n'avait pas prévu qu'ils obtiennent ce score. Quant à l'item 1, il ne montre aucun dysfonctionnement.

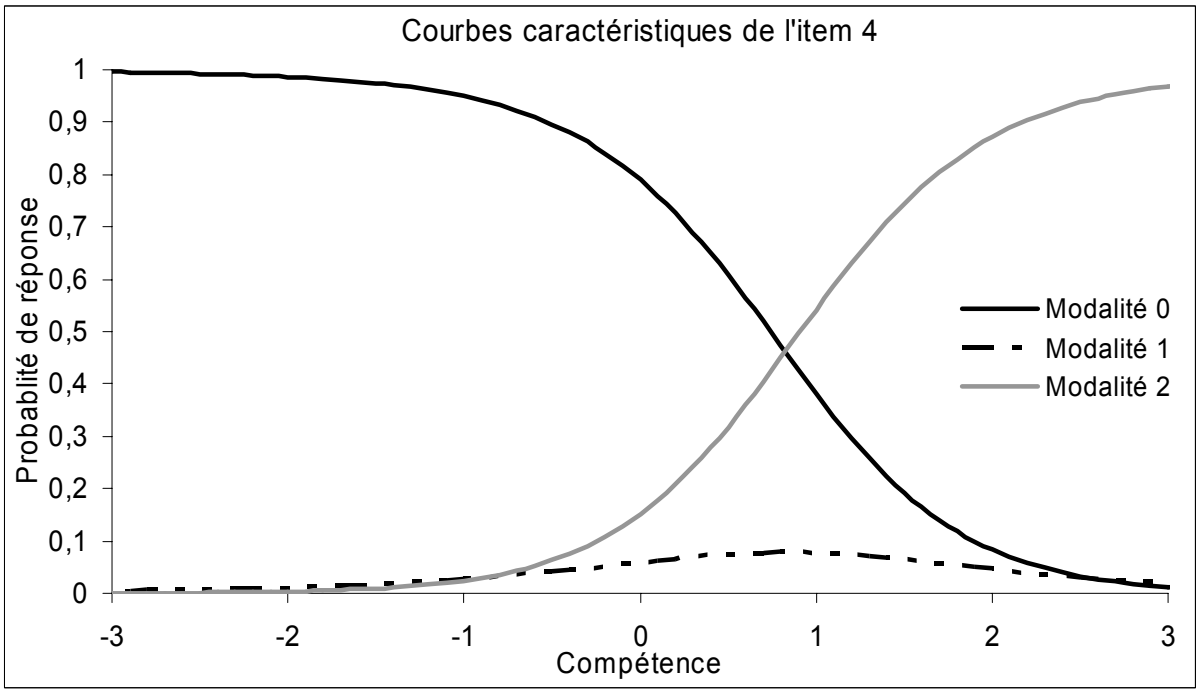
Item	**Mesure de la difficulté**	**Erreur de mesure de la difficulté**	***Infit***	***Outfit***
1	-0,58	0,14	0,94	0,92
2	0,72	0,06	1,05	1
4	0,82	0,08	1,04	0,91

Tableau 2. **Résultats statistiques des items obtenus avec le modèle PCM**

Item	**Modalité**	**Compétence moyenne**	***Infit***	***Outfit***	**Mesure du seuil de transition**	**Effectif**	**Pourcentage**
2	**0**	-0.16	1.02	0.97	Aucun	125	49 %
	1	0.11	1.68	1.88	1.02	46	18 %
	2	0.63	0.75	0.48	0.54	40	16 %
	3	0.86	1.07	1.00	0.59	46	18 %
4	0	-0.07	1.01	0.99	Aucun	172	67 %
	1	0.78	0.91	0.47	2.61	15	6 %
	2	0.71	1.06	1.06	-0.96	70	27 %

Tableau 3. **Résultats statistiques sur les modalités des items obtenus avec le modèle PCM**





Figures 6 – **Courbes caractéristiques pour les items polytomiques du DELF**

Conclusion

La psychométrie fournit un cadre méthodologique solide et apporte un éclairage particulier dans le domaine des sciences de l'éducation à l'aide de tests standardisés (pour une revue, Laveault et Grégoire, 2002). L'identification des dysfonctionnements que peuvent présenter les éléments constitutifs (les items) d'un instrument de mesure et leur remédiation contribuent grandement à améliorer la qualité des tests proposés, que ce soit dans le domaine de l'éducation ou plus généralement de la psychologie

appliquée. Notons toutefois que le résultat à un test ne donne pas une mesure absolue d'une capacité ou d'un trait de personnalité, comme une balance attribue à chacun un poids, mais que celui-ci permet une comparaison de la performance d'un individu avec les performances des individus auxquels il est légitimement comparable. Ainsi, puisque par définition la dimension latente (la compétence en langues, par exemple) n'est pas directement observable, il faut se donner les moyens, d'une part, d'assurer une mesure aussi précise que possible de cette dimension à l'aide d'une ou de plusieurs variables manifestes, et, d'autre part, de pouvoir évaluer les candidat/es, non pas nécessairement de manière absolue mais en tous les cas de manière équitable.

Les différentes techniques exposées dans cet article permettent d'assurer la fiabilité et la validité d'un test, au travers de sa standardisation et de son étalonnage, dans un souci constant d'assurer la qualité des items qui le composent. Le travail fourni par les experts du CIEP participe de cette volonté de délivrer des examens et des certifications correspondant aux standards d'élaboration d'un outil d'évaluation. Les processus de conception et de gestion de ces outils d'évaluation obéissent tous aux standards minimums établis par les membres de ALTE (Association of Language Testers in Europe) et sont évalués périodiquement par des experts qualiticiens dans le cadre de la certification ISO 9001: 2000 délivrée au Test de connaissance du français.

Bibliographie

BERNIER JEAN-JACQUES & BOGDAN PIETRULEWICZ. 1997. *La psychométrie. Traité de mesure appliquée*. Gaëtan Morin.

PAUL DICKES, JOCELYNE TOURNOIS, ANDRÉ FLIELLER & JEAN-LUC KOP. 1994. *La psychométrie. Théories et méthodes de la mesure en psychologie.* PUF, Le psychologue.

JOHN M. LINACRE. 2004. Rasch Model Estimation: Futher Topics. EVERETT V. SMITH JR. & RICHARD M. SMITH (dir.). *Introduction to Rasch Measurement*. Jam Press. 48-72.

RITA K. BODE. 2004. Partial Credit Model and Pivot Anchoring. EVERETT V. SMITH JR. & RICHARD M. SMITH (dir.*). Introduction to Rasch Measurement*. Jam Press. 279-295.

BIRNBAUM, A. 1968. Some latent trait models and their use in inferring an examinee's ability. LORD, F.M. & M.R.NOVICK. *Statistical theories of mental tests scores.* Addison-Wesley.

EMBRETSON, S.E. & S.P REISE. 2000. *Item response theory for psychologists*. Laurence Erlbaum Associates.

LAVEAULT D. & J.GRÉGOIRE. 2002. *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. De Boeck Université.

MASTERS, G.N. 1982. A Rasch model for partial credit scoring. *Psychometrika* 47. 149-174.

RASCH, G. 1960. *Probabilistic Models for Some Intelligence and Attainment Tests*. University of Chicago Press.

SAMEJIMA. 1969. Estimation of latent ability using a response pattern of graded scores. *Psychometrika* 17, Monograph Supplement.

Remerciements

Les auteur/es souhaitent remercier Patrick Riba, responsable du bureau DELF-DALF, Ingrid Jouette et Auréliane Baptiste, chargées de programme, pour leur aimable autorisation à diffuser et exploiter certains items composant les épreuves du DELF (avant leur révision finale).