Groupe–Edumétrie

Qualité de la mesure en éducation

A propos de la théorie des réponses aux items (TRI – IRT)

I. Le cas d'items dichotomiques

Gianreto PINI

Août 2012

1. Brève présentation de la méthode

La théorie des réponses aux items est un modèle statistique de la mesure relativement récent (deuxième moitié du 20e siècle) qui permet de faire face à des problèmes auxquels la psychométrie classique n'apporte pas toujours des réponses et des solutions satisfaisantes. Ainsi par exemple, l'évaluation des propriétés techniques d'un item (par le calcul de certains indices: de difficulté ou de discrimination notamment) fournit des résultats qui sont toujours relatifs à l'échantillon particulier d'individus auquel l'item a été administré. De ce fait, un item jugé facile ou difficile au sein d'un échantillon, peut ne plus l'être (ou ne plus l'être autant) s'il était appliqué à un échantillon différent.
Par rapport à ce genre de situation, la théorie des réponses aux items (TRI ou IRT dans la littérature anglophone) s'efforce de produire une estimation des propriétés de l'item qui soit indépendante d'un groupe particulier d'individus. En d'autres termes, elle cherche à élaborer des instruments de mesure dont les caractéristiques ne soient pas excessivement influencées par tel ou tel autre groupe de référence: ce qui, d'une certaine manière, conduit à définir des échelles qualifiées parfois d'"absolues".

Les premières tentatives visant à élaborer des échelles de ce genre remontent au début des années 50 (échelles de Guttman). A l'origine, elles reposaient sur un modèle (conceptuellement difficile à justifier) de nature entièrement déterministe, qui, par la suite, a été remplacé par des modèles beaucoup plus plausibles, de type probabiliste. Ces modèles sont fondés sur le postulat que la réponse d'un individu à l'item (et notamment sa probabilité de fournir une réponse correcte) est déterminée – ou peut être expliquée – par deux sortes de facteurs:

- d'une part, certains attributs du sujet (sa compétence par exemple), qui, n'étant pas directement accessibles à l'observation et à la mesure, sont généralement qualifiés de "traits latents";
- d'autre part, les propriétés de l'item lui-même: en particulier son degré de difficulté, son pouvoir de discrimination, sans oublier le rôle que la "chance" (réponses "au hasard") peut jouer dans certains cas.

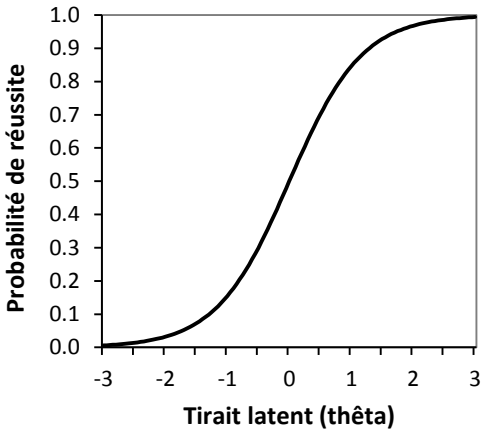
La réponse fournie à l'item est donc considérée comme une fonction des caractéristiques de l'individu et des caractéristiques de l'item. On postule par ailleurs (du moins dans la plupart des applications) que tous les items appartenant à un instrument donné (test, épreuve, échelle, etc.) permettent d'appréhender une même caractéristique "sous-jacente" (exigence d'unidimensionnalité: voir page 12 ci-après), et que les réponses à ces items sont affectées d'une erreur de mesure aléatoire.
Sur le plan technique et mathématique, la TRI utilise des modèles (équations) à un ou à plusieurs paramètre(s), qui établissent la relation entre le trait latent (niveau de compétence par exemple) et la probabilité de réussir un item. Cette relation est formalisée par une fonction (appelée fonction caractéristique de l'item), et peut être représentée graphiquement par une courbe (la courbe caractéristique de l'item).

Dans le contexte général qui vient d'être esquissé, l'objectif de la méthode est double, ces deux visées étant parfois poursuivies simultanément. Il s'agit, d'une part, d'estimer les propriétés métriques des items (calcul des paramètres dits de difficulté, de discrimination et, éventuellement, de pseudo-chance) et, d'autre part, d'estimer le niveau auquel chaque individu possède le trait latent (paramètre exprimant, par exemple, son niveau de compétence).
Par ailleurs, ces estimations sont supposées indépendantes des échantillons particuliers (d'individus d'une part et d'items de l'autre) à partir desquels l'étude est réalisée (propriété d'invariance: voir page 12).

L'exposé qui suit se réfère au cas (fréquent en sciences de l'éducation) où l'on est en présence d'un test ou d'une épreuve d'aptitude, de connaissance, de compétence, etc. Dans certains cas, la méthode peut toutefois s'appliquer aussi à d'autres instruments, et notamment à des échelles de toute sorte: d'intérêt, d'attitude ou de motivation par exemple.

2. Courbe caractéristique et fonction caractéristique d'un item

Comme nous venons de le signaler, la TRI repose sur le postulat que la réussite à un item dépend notamment du degré auquel l'individu possède une certaine caractéristique (compétence, aptitude, habileté, etc.): caractéristique désignée avec le terme de trait latent et représentée par la lettre grecque θ (thêta).
Dans ce cadre, la relation entre le trait latent et la performance (réussite à l'item) est modélisée par une fonction (appelée fonction caractéristique de l'item: voir ci-dessous), et elle est représentée graphiquement par une courbe: la courbe caractéristique de l'item (schéma ci-contre).



Cette courbe présente la forme d'un S plus ou moins allongé (sigmoïde) et décrit le lien qui existe entre la situation des individus par rapport au trait latent (leur niveau d'aptitude plus ou moins élevé par exemple) et la probabilité que ces individus ont de réussir l'item.

Le trait latent étant supposé normalement distribué au sein de la population, il est exprimé sur une échelle analogue à celle des scores *z*, dont les valeurs sont pratiquement comprises entre −3 et +3 (distribution centrée et réduite).
La définition de la courbe caractéristique d'un item peut être envisagée en considérant un ou plusieurs paramètres (modèles à un, deux ou trois paramètres), qui décrivent certaines propriétés importantes de l'item. Ce sont les paramètres dits de difficulté, de discrimination et de pseudo-chance [1].

Sur le plan mathématique, la fonction caractéristique de l'item est exprimée par une équation issue de la famille des modèles logistiques, dont la formulation dépend d'un certain nombre de facteurs et, notamment, du nombre de paramètres que l'équation comporte.
A titre d'illustration, voici l'équation qui définit la fonction considérée dans cet exemple, où $P_j(\theta)$ est la probabilité de réussite à l'item *j* pour un sujet possédant à un certain degré le trait latent thêta, et δ_j est un paramètre associé au même item (ici égal à 0):

$$P_j(\theta) = \frac{1}{1+e^{-1.7(\theta-\delta_j)}}$$

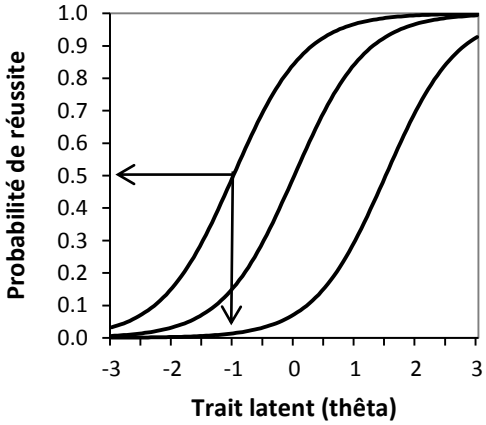
3. Les paramètres des items

3.1 Le paramètre de difficulté (δ_j)

Quel que soit le modèle utilisé pour définir la courbe caractéristique d'un item (à un, deux ou trois paramètres), le paramètre dit de difficulté est toujours présent. Dans le cadre de la TRI, il est défini par convention comme la valeur de thêta qui correspond à une probabilité de réussite exactement égale à 0.5.

[1] Des modèles comportant un nombre de paramètres supérieur à trois sont cités par certains auteurs. Nous n'en parlerons pas ici, car ces modèles sont très rarement utilisés en sciences de l'éducation.

C'est précisément cette valeur de thêta qu'on appelle paramètre de difficulté de l'item (δ_j).



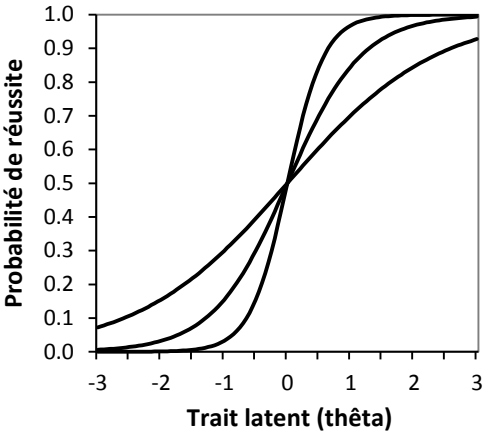
Dans l'exemple illustré par le schéma ci-contre les paramètres de difficulté des items représentés par les trois courbes sont respectivement, de gauche à droite: −1 (valeur lue sur l'axe horizontal pour une probabilité de réussite égale à 0.5), 0 et +1.5.
L'item décrit par la première courbe est donc plutôt facile, celui qui est décrit par la deuxième est de difficulté "moyenne" tandis que le dernier est le plus difficile.
On remarquera également que la mesure du trait latent chez les sujets et la difficulté des items sont exprimées sur la même échelle, allant de –3 et +3.

Signalons enfin qu'en vertu des propriétés de la distribution normale centrée et réduite, la valeur du paramètre de difficulté permet (dans certains cas tout au moins) de définir la proportion d'individus dont la probabilité de réussite est d'au moins 0.5. Ainsi, pour une valeur du paramètre égale à −1, cette proportion est de 84 %; pour une valeur égale à +1.5, la proportion est de 7 %; etc.

L'item sera jugé d'autant plus facile que la valeur du paramètre est proche de −3 et d'autant plus difficile que cette valeur est proche de +3.

3.2 Le paramètre de discrimination (α_j)

Une deuxième caractéristique importante de l'item est son pouvoir discriminatif, c'est-à-dire son aptitude à différencier les individus (distinguer ceux qui réussissent l'item de ceux qui y échouent) en fonction du degré auquel ils possèdent le trait latent. En TRI, cette propriété est fonction de la pente maximale de la courbe caractéristique.



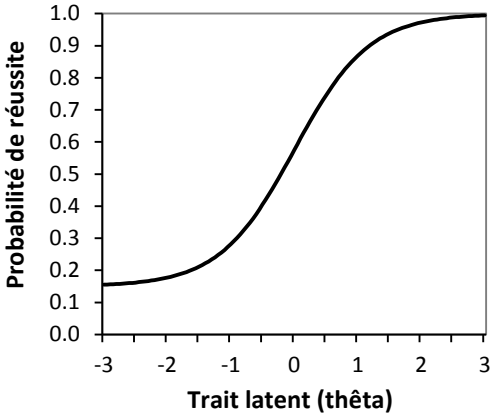
La valeur du paramètre de discrimination α_j est en effet proportionnelle à la pente de la tangente géométrique passant par le point d'inflexion de la courbe (on montre que la pente est maximale en ce point). Cette pente varie théoriquement entre 0 (lorsque l'angle formé avec l'axe horizontal est égal à 0°) et ∞ (angle égal à 90°). La pente peut donc être plus ou moins inclinée: plus la pente est abrupte, plus l'item est discriminatif et inversement [2].

L'exemple présenté ci-dessus montre les courbes caractéristiques de trois items de difficulté "moyenne" (paramètre de difficulté égal à 0), mais qui diffèrent quant à leur pouvoir de discrimination. Les paramètres α_j sont de 0.5 pour la courbe ayant la pente la plus faible, de 1 (cas intermédiaire) et de 2 pour la courbe avec la pente la plus accentuée.

[2] De manière plus formelle, on peut dire que la pente est exprimée par la tangente trigonométrique de l'angle que la tangente géométrique au point d'inflexion forme avec l'axe horizontal. On comprend dès lors pourquoi sa marge théorique de variation est comprise entre 0 et ∞ (elle est égale à 1 pour un angle de 45°). En pratique, il peut arriver que ce paramètre soit de signe négatif. Cela se produit lorsque l'item est mieux réussi par les moins bons élèves que par les meilleurs (moins bons et meilleurs en fonction de la position qu'ils occupent sur l'échelle des scores thêta).
Signalons enfin que, pour un modèle à deux paramètres, $\alpha_j = 4\ \Phi_j / D$, où Φ_j est la pente pour l'item *j* et *D* une valeur constante (voir plus loin).

3.3 Le paramètre de pseudo-chance (γ_j)

De nombreuses démarches d'évaluation ou de mesure sont effectuées en utilisant des items à choix multiple, où le sujet est invité à choisir la réponse qu'il juge correcte parmi deux ou plusieurs options soumises à son appréciation.
Dans de telles situations, on conçoit aisément que des facteurs aléatoires puissent influencer la performance. En effet, même s'il ne possède aucune compétence dans le domaine faisant l'objet de l'évaluation, l'individu a une probabilité non nulle de répondre correctement en choisissant au hasard l'une des options proposées.



Dans le jargon propre à la TRI, le terme de pseudo-chance désigne ce phénomène. Il se manifeste par le fait que la probabilité de réussite correspondant à une valeur de thêta égale à -3 est sensiblement supérieure à zéro (voir schéma).
En considérant la valeur de la courbe sur l'axe vertical au point -3 de l'échelle des scores θ, il est alors possible de définir un paramètre dit précisément de pseudo-chance (γ_j). Ce paramètre (dont la marge de variation théorique est comprise entre 0 et 1) peut s'interpréter comme la probabilité de réussir l'item *j* pour un individu d'habileté θ aussi faible que l'on puisse imaginer (dans cet exemple, $\gamma_j = 0.15$).

4. Les modèles

4.1 Le modèle de Rasch

Le modèle dit de Rasch constitue l'approche la plus simple utilisée dans le cadre de la TRI pour modéliser la relation entre le trait latent et la probabilité de réussir correctement un item (définition de sa fonction et de sa courbe caractéristiques).

La simplicité du modèle de Rasch va de pair avec une contrainte particulièrement exigeante, puisque tous les items d'un test sont supposés avoir le même pouvoir discriminatif (égal à 1). S'appuyant sur ce postulat, le modèle définit la fonction de l'item en considérant une seule caractéristique: sa difficulté. Pour cette raison, on parle parfois de modèle à un paramètre [3].

Selon ce modèle, la probabilité $P_j(\theta)$ de réussite à l'item *j* pour un individu qui possède le trait latent au niveau θ est définie par l'équation suivante (*e* est base des logarithmes naturels et δ_j le paramètre de difficulté de l'item):

$$P_j(\theta) = \frac{e^{(\theta-\delta_j)}}{1+e^{(\theta-\delta_j)}} = \frac{1}{1+e^{-(\theta-\delta_j)}}$$

[3] Signalons qu'en dépit de sa simplicité jugée parfois "excessive", le modèle de Rasch est très souvent utilisé dans toutes sortes de domaines, et notamment dans le cadre des grandes enquêtes internationales (PISA par exemple).

4.2 Modèles à un, deux ou trois paramètres

Comme nous venons de le voir, l'équation de Rasch est l'exemple typique d'un modèle à un paramètre. Au niveau de sa formulation, on peut y ajouter un constante *D* (voir formule de gauche), sorte de facteur d'"échelonnement" implicitement égal à 1 dans le modèle présenté au point précédent, mais qui peut assumer des valeurs différentes.
On montre notamment que lorsqu'on attribue à *D* la valeur 1.7 (voir formule de droite) la courbe caractéristique de l'item assume une allure très proche de l'ogive normale (intégrale de la courbe normale). Dans la pratique c'est donc souvent cette équation qui est appliquée lorsqu'on utilise le modèle à un paramètre.

$$P_j(\theta) = \frac{1}{1+e^{-D(\theta-\delta_j)}} \qquad\qquad P_j(\theta) = \frac{1}{1+e^{-1.7(\theta-\delta_j)}}$$

$P_j(\theta)$	probabilité qu'un individu possédant à un certain degré (entre –3 et + 3) la caractéristique θ réponde correctement à l'item;
δ_j	paramètre de difficulté de ce même item;
e	base des logarithmes naturels (népériens): 2.7182...;
1.7	valeur attribuée à la constante *D*.

Une approche plus complexe du problème consiste à définir $P_j(\theta)$ en faisant appel à deux paramètres: le paramètre de difficulté (δ_j) et le paramètre de discrimination (α_j). C'est le modèle dit précisément à deux paramètres, appelé aussi modèle de Birnbaum:

$$P_j(\theta) = \frac{1}{1+e^{-1.7\,\alpha_j(\theta-\delta_j)}}$$

Enfin, lorsque l'instrument est composé d'items à choix multiple (deux ou plusieurs options de réponse proposées), il est également possible d'ajouter au modèle un troisième paramètre (modèle à trois paramètres). Il s'agit du paramètre de pseudo-chance (γ_j), qui est supposé nul (égal à 0) dans les cas d'un modèle à un ou à deux paramètre(s):

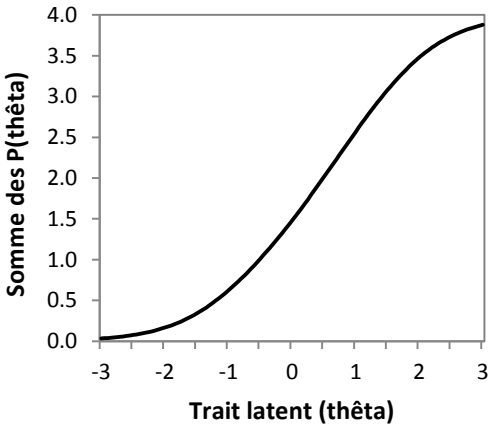
$$P_j(\theta) = \gamma_j + \frac{1-\gamma_j}{1+e^{-1.7\,\alpha_j(\theta-\delta_j)}}$$

5. La courbe caractéristique du test

Les modèles qui viennent d'être présentés permettent donc de construire les courbes caractéristiques pour tous les items qui figurent dans un test.
A partir de ces courbes, on peut définir la courbe caractéristique du test lui-même (voir schéma ci-après). Elle s'obtient simplement en additionnant, pour chaque valeur de thêta,

les probabilités relatives aux différents items. L'axe des ordonnées est une échelle dont les valeurs vont de *0 à n*, *n* étant le nombre total d'items (dans l'exemple, *n = 4*).

La forme de la courbe permet notamment de déterminer *(a)* quelle est la difficulté du test et *(b)* pour quelles valeurs de thêta (c'est-à-dire sur quelle portion de l'échelle des compétences) l'instrument discrimine le mieux les individus.



Dans l'exemple présenté ici, le test paraît de difficulté "moyenne" (à la valeur *n/2 = 2* sur l'axe vertical correspond une valeur de thêta comprise entre 0 et +1 sur l'axe horizontal).
Par ailleurs, le pouvoir de discrimination du test s'exerce surtout pour les individus dont le niveau de compétence est légèrement supérieur à la "moyenne" (valeurs de thêta entre 0 et +1.0).
C'est en effet dans cette zone que la pente de la courbe est la plus accentuée.

Enfin, la courbe caractéristique du test permet également d'effectuer une sorte de "conversion d'échelle", à partir de laquelle chaque résultat individuel (initialement exprimé par un score thêta compris entre −3 et +3) peut être interprété comme un pourcentage: plus exactement, comme le *pourcentage attendu* d'items réussis dans l'univers des items d'où proviennent ceux qui composent le test (échelle de "scores vrais" au sens usuel que ce terme assume en théorie de la mesure).
Pour obtenir ce pourcentage, il suffit de diviser par *n* et de multiplier par 100 la somme des probabilités obtenue pour une valeur donnée de thêta (calcul d'un pourcentage ordinaire).

6. Les paramètres des individus

Nous avons considéré jusqu'ici une première fonction de la TRI, qui permet de déterminer certaines caractéristiques des items et du test dans son ensemble. Une deuxième application importante est celle qui conduit à définir les paramètres des individus: c'est-à-dire, concrètement, leur niveau de compétence en fonction des réponses fournies aux différents items que le test comporte. Chaque individu obtiendra ainsi un score compris entre -3 et +3, sur une échelle qui est censée recouvrir l'éventail complet des niveaux possibles de compétence [4].

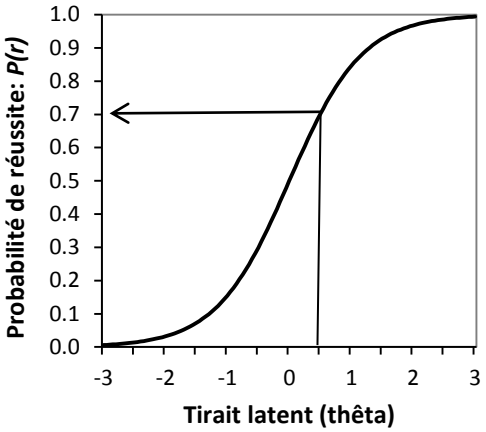
En ce qui concerne cette application de la TRI, deux cas différents peuvent se présenter.
Le cas le plus simple est celui où les paramètres des items sont connus (suite, par exemple, à des études effectuées antérieurement) et, sur la base de ces éléments, on procède à l'estimation des paramètres pour les individus.
Le deuxième cas est techniquement plus complexe et plus exigeant quant au nombre d'items et d'individus dont il faut disposer, car on ne connaît au départ ni les paramètres des items ni les paramètres des individus. Il s'agit donc d'estimer simultanément ces deux ensembles d'éléments [5].

Dans le cadre de cet exposé, seul le premier cas sera brièvement évoqué.

[4] Nous avons déjà signalé le fait que, dans le cadre de la TRI, la difficulté des items et la compétence des individus sont exprimées sur une même échelle: l'échelle des scores θ précisément.

[5] D'après certains auteurs, il faut disposer d'au moins 30 items et 500 sujets pour le modèle à deux paramètres et d'au moins 50 items et 1000 sujets pour le modèle à trois paramètres.

Comme nous l'avons vu, lorsqu'on connaît les paramètres d'un item on peut définir sa fonction et sa courbe caractéristiques en utilisant l'un ou l'autre des modèles existants. A l'aide de cette fonction et de cette courbe il est alors possible de déterminer, pour chaque valeur de thêta, la probabilité qu'un item soit réussi [*P(r)*] ou qu'il soit échoué [*P(e) = 1 – P(r)*].
Rappelons à ce propos que, sur le graphique qui représente la courbe caractéristique, la probabilité *P(r)* se lit sur l'axe vertical pour une valeur donnée de thêta.



Dans cet exemple (voir schéma), pour une valeur de thêta = +0.5, la probabilité de réussir l'item est *P(r) = 0.701*, tandis que la probabilité d'échouer sera *P(e) = 1 – 0.701 = 0.299*.

Si le test comporte *n* items, on aura donc *n* courbes caractéristiques et *n* probabilités pour chaque valeur de thêta. Ces *n* probabilités figurent dans les différentes lignes du tableau suivant, où $P(r)_j$ désigne la probabilité de réussir le *j-ième* item de la série:

θ	$P(r)_1$	$P(r)_2$	…	$P(r)_j$	…	$P(r)_n$
-3.0	0.008	0.002	⋯			0.015
-2.9	0.013					
-2.8						
…						
0.0	…					
…						
+2.8						
+2.9						
+3.0	0.987					

A partir de ces résultats on peut calculer la probabilité *P(C)* d'obtenir n'importe quelle configuration (ou combinaison) *C* de résultats.

Ainsi par exemple, dans le cas d'un instrument comportant 5 items, on peut déterminer la probabilité (pour chaque valeur de thêta) d'observer la configuration définie par une réussite aux items 1, 2, 4 et un échec aux items 3, 5: probabilité désignée par: $P(11010/\theta)$.

Pour traiter ce problème, on fait appel à la propriété d'indépendance locale (voir page 13). Cette propriété étant supposée acquise, on peut appliquer un théorème fondamental de la théorie des probabilités, selon lequel la probabilité que se réalisent deux ou plusieurs événements aléatoires indépendants est égale au produit des probabilités associées à la réalisation de chacun de ces événements.

Pour *n* événements indépendants de terme générique E_j ($j = 1$ à n) on a donc:

$$P(E_1 \ldots E_j \ldots E_n) = \Pi_j P(E_j)$$

Ainsi, à propos de l'exemple précédent, la probabilité d'obtenir la configuration (*11010*) pour une valeur donnée de thêta sera [*r* = réussite et *e* = échec, avec *P(e) = 1 – P(r)*]:

$$P(11010/\theta) = P(r)_1 \times P(r)_2 \times P(e)_3 \times P(r)_4 \times P(e)_5$$

Imaginons que, pour une valeur de $\theta = 0$, ces probabilités soient respectivement:

$$P(r)_1 = 0.846 \quad P(r)_2 = 0.701 \quad P(e)_3 = (1 - 0.395) \quad P(r)_4 = 0.155 \quad P(e)_1 = (1 - 0.023)$$

La probabilité d'obtenir la configuration (*11010)* sera donc (pour $\theta = 0$):

$$P(11010/0) = 0.846 \times 0.701 \times 0.605 \times 0.155 \times 0.977 = 0.0543$$

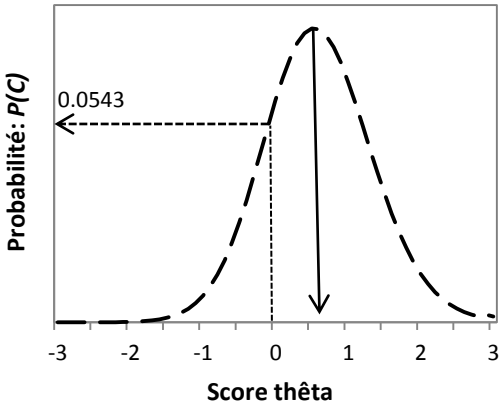
Les résultats ainsi obtenus permettent de construire une courbe dite de vraisemblance, dont ils définissent l'ordonnée (axe vertical) pour chaque valeur de thêta (axe horizontal).

Pour $\theta = 0$: *P(C) = 0.0543* dans l'exemple précédent (voir schéma).

A toute configuration de résultats [6] on peut donc associer une courbe de vraisemblance, et on attribuera aux individus qui possèdent cette configuration la valeur de thêta pour laquelle la probabilité *P(C)* est maximale (maximum de la fonction de vraisemblance). [7]

Dans l'exemple présenté ici le score thêta attribué aux individus ayant la configuration *(11010)* est +0.7: voir schéma.

Courbe de vraisemblance
pour une configuration *C* de résultats
Par exemple: *C = (11010)*



Une conséquence importante de cette démarche est que (sauf dans le cas du modèle à un paramètre), deux individus caractérisés par le même résultat global (même nombre d'items réussis) n'obtiennent pas forcément le même score thêta.
La contribution des réussites et des échecs à la définition du score thêta dépend en effet du pouvoir de discrimination des items, ainsi que du facteur de pseudo-chance qui leur est associé (tous les items ne "paient" pas de la même manière).
Le paramètre de difficulté a aussi une influence sur les scores thêta, mais cette influence est la même pour tous les individus.

7. La notion d'information

La notion d'information assume un rôle essentiel dans la TRI, où elle est parfois considérée comme l'équivalent des concepts de fidélité en théorie classique de la mesure ou de généralisabilité en théorie de la généralisabilité.
A l'origine de nombreuses applications, elle renseigne l'utilisateur de la méthode sur le "pouvoir informatif" d'un item ou du test dans son ensemble. Elle indique notamment sur quelle portion de l'échelle du trait latent (scores θ) le pouvoir informatif de l'item ou du test est le plus élevé: donc, pour quelle(s) catégorie(s) d'individus l'item ou le test est le plus précis (dans ce cadre, en effet, les termes d'information et de précision peuvent être considérés comme synonymes).

A cet égard, on comprend sans peine que, par exemple, un item "difficile" donne peu d'informations sur les compétences des élèves les plus faibles. En particulier, il ne permet pas de repérer des différences qui pourraient exister entre les sujets appartenant à cette catégorie: la seule information (en un sens sommaire) fournie par un item de cette nature est que tous les élèves situés en dessous d'un certain niveau de compétence échouent. Dans la catégorie générique des "bons élèves", en revanche, ce même item aura un pouvoir

[6] A l'aide du calcul combinatoire on vérifie que, dans le cas de 5 items, 32 configurations différentes peuvent théoriquement être observées (pour *n* items: 2^n).

[7] Sur le plan mathématique, le problème consiste à déterminer le maximum d'une fonction. Il est donc résolu en ayant recours aux méthodes du calcul différentiel. Techniquement l'estimation est exécutée en appliquant la méthode statistique dite du maximum de vraisemblance, dont la mise en œuvre comporte notamment une transformation logarithmique.

d'information plus élevé, permettant une réelle différenciation des individus en fonction de leur résultat (selon qu'ils réussissent l'item ou qu'ils y échouent).

De manière générale, on peut donc considérer qu'un item exerce souvent son pouvoir d'information dans une zone particulière de l'échelle des compétences. On montre d'ailleurs que ce pouvoir dépend, entre autres, du pouvoir de discrimination de l'item, qui est maximum pour les valeurs de thêta correspondant au point d'inflexion de la courbe caractéristique.
Sur le plan pratique, les algorithmes de la méthode permettent de calculer:

- pour un item donné et pour chaque point de l'échelle des compétences (chaque valeur de thêta), quel est son pouvoir informatif: définition de la fonction et de la courbe dites d'information (voir *point 8*);
- pour un item donné, à quel endroit sur l'échelle des compétences (pour quelle valeur de thêta) son pouvoir d'information est maximum.

Les éléments qui viennent d'être évoqués montrent que la notion d'information peut jouer un rôle important dans l'élaboration d'un instrument de mesure ou d'évaluation. En effet, lorsque les caractéristiques des items sont connues (sur la base d'analyses réalisées lors d'études précédentes par exemple), on pourra effectuer un choix d'items aussi conforme que possible aux besoins et aux exigences de chaque situation particulière.

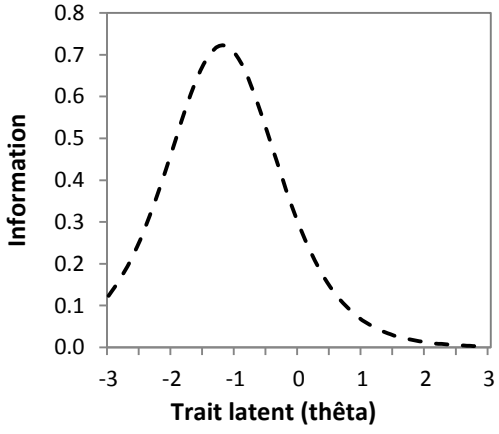
On retiendra également que la notion d'information entretient une relation étroite avec l'erreur qui affecte la précision de la mesure. Elle permet notamment d'estimer l'erreur standard associée aux paramètres (résultats) des individus sur l'échelle des scores thêta.

8. Courbes d'information et erreur standard

8.1 La courbe d'information de l'item

Nous avons vu précédemment que la fonction caractéristique d'un item permet de calculer, pour chaque niveau de compétence (pour chaque valeur de θ), la probabilité de réussir (et donc aussi de ne pas réussir) l'item.
A l'aide de ces probabilités et des paramètres du modèle utilisé (difficulté, discrimination et pseudo-chance), on peut définir la fonction d'information de l'item, représentée graphiquement par la courbe dite d'information (voir équation et schéma ci-après).

Cette courbe montre sur quelle portion de l'échelle des compétences (axe horizontal) le pouvoir informatif de l'item est le plus élevé.
Dans l'exemple présenté ici, les individus pour lesquels l'item est le plus informatif (et donc le plus précis) sont ceux dont le score θ est compris entre −2.0 et −0.5: niveaux de compétence assez nettement inférieurs à la "moyenne".



On peut d'ailleurs déterminer que, dans ce cas, le pouvoir informatif de l'item est maximum pour une valeur de θ égale à −1.2 (avec des modèles à un ou à deux paramètres, l'information est maximale pour la valeur de θ égale au paramètre de difficulté de l'item).

Voici l'équation qui définit la fonction d'information d'un item et qui permet de construire sa courbe d'information:

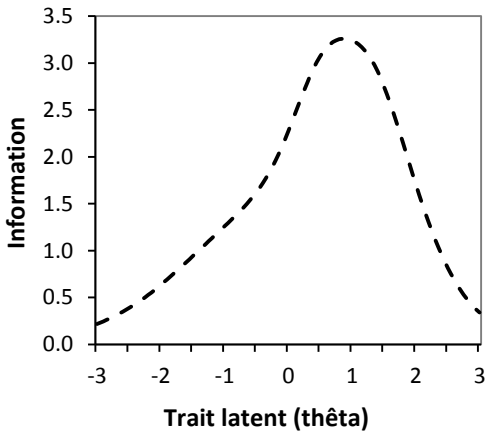
$$I_j(\theta) = D^2 \alpha_j^2 \frac{1-P_j(\theta)}{P_j(\theta)} \left[\frac{P_j(\theta)-\gamma_j}{1-\gamma_j} \right]^2$$

- $I_j(\theta)$ Information associée à l'item *j* au point thêta sur l'échelle du trait latent;
- $P_j(\theta)$ probabilité qu'un individu possédant à un certain degré la caractéristique θ réponde correctement à l'item;
- α_j paramètre de discrimination (égal à 1 dans le modèle à un paramètre);
- γ_j paramètre de pseudo-chance (égal à 0 dans les modèles à un ou deux paramètres);
- *D* constante à laquelle on attribue souvent la valeur 1.7.

8.2 La courbe d'information du test

A partir des courbes d'information établies pour tous les items qui composent l'instrument, on peut définir la courbe d'information du test lui-même (voir schéma).
Elle s'obtient simplement en additionnant, pour chaque valeur de thêta, les $I_j(\theta)$ calculés pour les *n* items considérés (*j = 1 à n*):

$$I(\theta) = \sum_j I_j(\theta)$$



Cette courbe montre notamment sur quelle(s) portion(s) de l'échelle des compétences (axe horizontal) le test est le plus informatif (le plus précis).
Dans l'exemple présenté ici, les individus pour lesquels le test est le plus informatif sont ceux dont les valeurs de thêta se situent entre 0 et +1.75 approximativement.

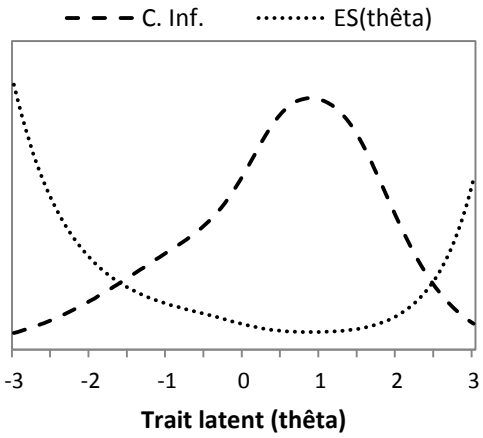
Il est par ailleurs à remarquer que, dans certains cas, on peut obtenir des courbes qui présentent deux sommets (un peu comme dans le cas de courbes bimodales), traduisant le fait qu'en raison de sa composition, le test exerce son pouvoir informatif dans deux zones différentes de l'échelle des compétences.

8.3 L'erreur standard de la mesure

Comme nous l'avons déjà signalé, la notion d'information permet également de définir l'erreur standard (ou erreur type) de la mesure.
En TRI, cette erreur n'est pas la même pour toutes les valeurs possibles de l'échelle des compétences. Contrairement à ce que postulent d'autres modèles de la mesure (théorie classique ou théorie de la généralisabilité notamment), l'erreur varie en effet d'une valeur à l'autre de thêta, comme on peut aisément le constater en considérant la formule suivante ($ES(\theta)$ désigne l'erreur standard de la mesure pour une valeur donnée de thêta):

$$ES(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

On vérifie en particulier que l'erreur est minimale pour les scores situés dans la zone de l'échelle où le test est le plus informatif (voir point *8.2* ci-dessus), et qu'elle augmente lorsqu'on s'éloigne de cette zone. De ce point de vue, l'erreur standard peut également être interprétée comme une mesure de l'incertitude associée à la mesure, celle-ci étant d'autant plus grande que le pouvoir d'information du test est faible.



En raison de ce lien étroit qui existe entre les notions d'information et d'erreur standard, certains logiciels indiquent sur un même graphique les courbes qui décrivent l'évolution de ces deux grandeurs tout au long de l'échelle définie par le trait latent. On remarquera que les échelles relative à ces deux courbes (axe vertical) ne sont pas les mêmes.

9. Caractéristiques de la méthode et conditions d'application

9.1 La propriété d'invariance

La propriété dite d'invariance est probablement la caractéristique principale de la TRI: pour certains auteurs, et dans certains cas tout au moins, c'est elle qui confère à cette méthode une certaine forme de "supériorité" par rapport à d'autres théories de la mesure (théorie classique et théorie de la généralisabilité). Cette propriété postule que:

- les estimations relatives aux items (paramètres de difficulté, de discrimination et de pseudo-chance) sont indépendantes de l'échantillon particulier d'individus à partir duquel elles sont effectuées;
- les estimations relatives aux individus (leur niveau de compétence, d'habileté ou d'"endossement": scores θ) sont indépendantes de l'échantillon d'items utilisé dans le cadre d'une étude particulière.

L'importance, théorique et pratique, de cette propriété est considérable. Il importe toutefois de préciser que l'invariance des estimations est toujours "relative", car l'"ancrage" des échelles (le choix de l'origine) est arbitraire. De ce fait, les valeurs des paramètres ne sont définies qu'à une transformation affine près.
On retiendra par ailleurs que l'invariance des estimations n'est assurée que si certaines conditions relativement strictes sont satisfaites: en particulier, il faut que l'ajustement du modèle aux données soit satisfaisant pour la population dans son ensemble ainsi que pour différents sous-groupes provenant de cette population (garçons et filles, etc.): voir page 14.

9.2 L'unidimensionnalité

Comme cela est souvent le cas dans la construction de dispositifs d'évaluation ou de mesure, la TRI postule l'unidimensionnalité de l'instrument (test ou échelle) auquel elle est appliquée. Concrètement cela suppose que tous les éléments qui le composent (tous les items,

toutes les questions) contribuent à appréhender chez les individus une seule et même caractéristique (ou dimension) "sous-jacente": un même type de compétence, l'attitude manifestée à l'égard d'un même objet, l'intérêt pour telle forme d'activité, etc. [8]
Toutefois, dans la pratique courante de la mesure, il est rare que l'on soit en présence d'une dimension réellement unique. Ainsi, selon une définition plus large, on parlera d'unidimensionnalité lorsqu'il existe une dimension clairement dominante (un facteur "général" dans le langage de l'analyse factorielle) par rapport à d'autres dimensions qui peuvent également intervenir et exercer une certaine influence (supposée de nature aléatoire la plupart du temps) sur les résultats individuels.
Il existe différentes approches qui permettent d'évaluer cette caractéristique de l'instrument, parmi lesquelles on peut citer le coefficient alpha de Cronbach, le modèle de la généralisabilité ou l'analyse factorielle.

9.3 L'indépendance locale

En TRI, la validité des estimations relatives aux caractéristiques des individus (leur niveau de compétence, d'habileté ou d'"endossement": paramètres des individus) suppose que la condition dite d'indépendance locale soit satisfaite.
Cette condition postule que la performance (réussite ou échec) à chaque item n'est pas influencée par la performance relative aux autres items qui figurent dans le même instrument (d'où précisément la notion d'indépendance).
Techniquement cela signifie que, pour un niveau de compétence donné (c'est-à-dire pour une certaine valeur de thêta), la corrélation entre les résultats des individus à deux items quelconques doit être nulle ou proche de zéro.
La notion d'indépendance locale entretient certaines relations avec celle d'unidimensionnalité. On peut d'ailleurs montrer que lorsque la condition d'unidimensionnalité est satisfaite, celle d'indépendance locale l'est également.

10. Quelques domaines d'application

10.1 Etude du fonctionnement différentiel des items

On dit qu'un item présente un "fonctionnement différentiel" lorsque son comportement varie d'un groupe d'individus à l'autre au sein d'une même population (garçon et filles; élèves issus de telle ou telle autre catégorie sociale, etc.). Source de biais qui risquent d'affecter la précision de la mesure, la présence d'items de ce type est généralement révélée par le fait que des individus possédant un même niveau global de compétence les réussissent systématiquement mieux ou moins bien selon le groupe auquel ils appartiennent (par exemple les garçons systématiquement mieux que les filles).
Le fonctionnement différentiel des items est un aspect que les spécialistes de l'évaluation et de la mesure connaissent depuis longtemps, et qui a donné lieu à plusieurs stratégies visant à en estimer l'importance et à en corriger les effets. Dans ce contexte, la comparaison des courbes caractéristiques d'un même item définies séparément pour différents sous-groupes d'individus constitue une démarche d'analyse souvent très utile.

[8] Certains travaux récents dans ce domaine visent à développer des modèles multidimensionnels, qui devraient permettre de traiter des situations caractérisées par la présence de plusieurs dimensions et exigeant de ce fait le recours à plusieurs paramètres de compétence (d'habileté).

10.2 Construction et gestion d'une banque d'items

On désigne avec le terme "banque d'items" un répertoire (liste, catalogue, recueil) d'items qui, le plus souvent, ont déjà été utilisés à des fins d'évaluation ou de recherche, et que l'on souhaite mettre à la disposition de personnes (chercheurs ou enseignants) susceptibles de s'en servir dans le cadre de leurs activités. Classés selon différents critères techniques et pédagogiques, ces items ont parfois déjà fait l'objet d'analyses antérieures, qui ont permis d'en déterminer certaines caractéristiques: notamment leur niveau de difficulté et leur pouvoir de discrimination. Toutefois, dans la mesure où ces items ont été appliqués à des groupes d'élèves, d'étudiants ou de candidats très divers, les caractéristiques techniques qui leur sont conférées par les méthodes traditionnelles pourraient ne pas être pertinentes lorsqu'ils sont utilisés avec d'autres groupes d'individus. De ce point de vue, la propriété d'invariance que la TRI permet d'assurer est un élément qui rend cette approche particulièrement adaptée pour l'élaboration et la gestion d'un dispositif de cette nature.

10.3 Application du testing adaptatif

En psychologie, le "testing" est une pratique déjà ancienne, dont l'origine remonte à la fin du 19e siècle, et qui, pendant des décennies, a été réalisée selon le modèle classique du "test papier-crayon". Face aux problèmes méthodologiques que cette approche ne parvient pas toujours à résoudre de manière satisfaisante, les spécialistes du domaine se sont efforcés de concevoir une forme de testing dit "sur mesure" (*tailored testing*), dont la spécificité est de pouvoir s'adapter en cours de route aux caractéristiques de chaque individu particulier. Concrètement cela suppose qu'à partir des réponses initialement fournies par le sujet, une première estimation de son niveau de compétence puisse être établie, et que les questions posées ultérieurement soient choisies (par un dispositif automatisé et souvent informatisé) en tenant compte de la compétence préalablement estimée. Dans la suite de la démarche, cette estimation sera progressivement précisée et affinée, jusqu'au moment où des "règles d'arrêt" sont activées lorsque le processus d'estimation semble avoir atteint un degré de stabilité et de précision satisfaisants. L'avènement de la TRI a contribué de manière décisive au développement du testing adaptatif, au sein duquel elle joue un rôle tout à fait essentiel.

10.4 Traitement des données dans les enquêtes internationales

Bien que les spécialistes de l'évaluation et de la mesure s'en servent depuis quelques décennies déjà, la TRI a acquis une notoriété considérable au cours des 15 dernières années en raison du rôle qu'elle joue au sein des grandes enquêtes internationales (PISA par exemple). Dans le cadre de ces études, en effet, le modèle de Rasch est appliqué pour déterminer les caractéristiques des items ("calibrage") et pour établir les scores individuels. Cette démarche permet notamment de situer sur une même échelle de compétences des élèves qui (mis à part un sous-ensemble relativement restreint d'items communs) ont été confrontés et ont répondu à des items différents. Il est ainsi possible d'effectuer l'étude en considérant un vaste répertoire d'items, qui garantit un échantillonnage aussi exhaustif que possible des domaines de compétence considérés.

11. Ajustement du modèle aux données

Les algorithmes et les techniques présentés dans les pages précédentes permettent toujours de construire la courbe caractéristique pour chacun des items considérés. De plus, nous savons que la solution ainsi obtenue constitue en général le meilleur ajustement possible du

modèle aux données dont on dispose.
Il arrive toutefois que même le meilleur ajustement possible demeure insatisfaisant et objectivement de mauvaise qualité, ce qui peut avoir des conséquences fâcheuses au niveau des interprétations que l'on sera amené à formuler. Il est donc nécessaire que la qualité de l'ajustement fasse l'objet d'une étude minutieuse, à l'aide de démarches spécifiquement conçues à cet effet.
Une vérification de ce type peut d'ailleurs être envisagée à plusieurs niveaux. On peut en effet considérer l'ajustement global du modèle, l'ajustement pour chaque item et l'ajustement pour chaque individu. En général, lorsqu'on évoque la qualité de l'ajustement sans autres précisions, on se réfère à l'ajustement global du modèle (ajustement pour l'ensemble des items et pour l'ensemble des individus).
Dans le cadre de ce bref exposé nous n'entrerons pas dans le détail des procédures auxquelles on peut avoir recours. Signalons simplement que ces procédures sont essentiellement de deux sortes: graphiques les unes et numériques les autres.

Les méthodes de vérification graphiques présentent de nombreuses analogies avec celles qu'on utilise dans d'autres domaines de l'analyse statistique (analyse de régression par exemple). Elles consistent pour l'essentiel à comparer les courbes théoriques fournies par le modèle aux résultats empiriques réellement observés (à cet égard, les individus sont souvent répartis en un certain nombre de catégories, en fonction de leur position sur l'échelle des scores thêta). Dans cette perspective, les écarts observés dans chaque catégorie entre la proportion d'individus qui s'y trouve et la proportion d'individus prédite par le modèle peut faire l'objet de représentations graphiques diverses et fournir des renseignements précieux concernant la qualité de l'ajustement.

En revanche, les méthodes numériques conduisent à calculer différents indices à partir des écarts (souvent désignés avec le terme de résidus) qui existent entre résultats observés et résultats attendus. Certains de ces indices possèdent une distribution d'échantillonnage connue. Suivant une démarche largement répandue, il est alors possible de tester l'hypothèse nulle selon laquelle (mises à part les fluctuations aléatoires dues à l'échantillonnage) il n'y a aucune différence systématique entre résultats observés et résultats prédits par le modèle. Si l'hypothèse nulle est rejetée aux seuils de signification usuels, on considère que l'ajustement du modèle risque de ne pas être satisfaisant.
Pour d'autres indices, en revanche, l'interprétation se fait en comparant la valeur calculée à des seuils ou critères de "décision", fixés le plus souvent de manière empirique sur la base de considérations dictées par la pratique et par l'expérience (l'ajustement est considéré comme étant satisfaisant si l'indice calculé est inférieur / supérieur à une certaine valeur, ou s'il est compris entre telle et telle autre limites).

12. Références bibliographiques utiles

Bertrand, R. & Blais, J.-G. (2004).

Modèles de mesure. L'apport de la théorie des réponses aux items.

Sainte-Foy, Presses de l'Université du Québec, 376 p.

Laveault, D. & Grégoire, J. (2002).

Introduction aux théories des tests en sciences humaines.

Paris, De Boeck, 2e éd., 336 p.