Tutorials in Quantitative Methods for Psychology
2011, Vol. 7(2), p. 42-53.

DOI: 10.20982/tqmp.07.2.p042

Les estimateurs de capacité dans la théorie des réponses aux items et leur biais

Louis Laurencelle
Université du Québec à Trois-Rivières

Stéphane Germain
Université Laval

Dans le modèle paramétrique de la théorie des réponses aux items (TRI), l'estimation du niveau d'aptitude (θ) du répondant a fait l'objet de plusieurs procédures différentes, certaines basées uniquement sur le protocole de réponses observé et d'autres, appuyées sur un modèle de population contraignant. Peu d'études cependant ont entrepris d'établir les mérites comparatifs de ces procédures. Au moyen d'une large expérimentation Monte Carlo, nous étudions le biais, la précision et l'efficacité de capture de quatre estimateurs de capacité : MV, BME, EAP, WARM, auxquels nous ajoutons un estimateur MV winsorisé (WINS) et un autre extrapolé (EXT), qui permettent tous deux l'utilisation de l'estimateur MV pour les scores parfaits (tous items réussis) et nuls (tous items échoués).

In the parametric setting of item response theory (IRT), several procedures have been put up to estimate the examinee's ability level (θ), some based solely on the information in the response protocol while others are grounded on a restrictive population model. Few studies however have endeavoured to compare the relative merits of these procedures. By way of an extensive Monte Carlo experiment, we studied the bias, precision and capture efficiency of four ability estimators : ML, BME, EAP, WARM, to which we added a winsorized ML estimator (WINS) and an extrapoled one (EXT), both enabling ML estimation for perfect (all items passed) and null (all items failed) scores.

Malgré la grande utilisation qui en a été faite depuis les premières apparitions de la théorie des réponses aux items (TRI), dans Lord (1977), peu de chercheurs se sont intéressés à l'étude des propriétés statistiques des estimateurs du niveau d'aptitude du répondant. Il existe d'excellents ouvrages d'introduction à la TRI, parmi lesquels nous signalerons celui, en français, de Bertrand et Blais (2004) et l'ouvrage déjà classique de Hambleton, Swaminathan et Rogers (1991). Dans ces ouvrages, pas un mot, ou presque, sur le biais, la variance et l'efficacité des différents estimateurs proposés. Plus étonnant encore, l'ouvrage en seconde édition de Baker et Kim (2004), dont le sous-titre se lit « Parameter estimation techniques », reste muet sur la question.

Ce n'est pas dire que les auteurs ont ignoré l'affaire. Au contraire, dès 1983, Lord développe une expression pour quantifier approximativement le biais de l'estimateur du maximum de vraisemblance (MV) : nous y revenons plus loin. Warm (1989) intercepte la fonction de biais de Lord et s'en sert pour proposer un estimateur débiaisé. Samejima (1993a, 1993b) reprend et prolonge l'étude de Lord, toujours pour l'estimateur MV. Enfin, nous trouvons une étude comparative, celle de Kim et Nicewander (1993) : on y étudie différents estimateurs du paramètre de capacité par le moyen de simulations Monte Carlo, en explicitant certaines de leurs propriétés psychométriques. En dehors de cette parution, difficile de trouver des données permettant d'évaluer et de comparer les différents estimateurs ayant cours dans la pratique, notamment les estimateurs bayésiens BME et EAP, en plus de MV et des estimateurs dérivés.

C'est une telle étude comparative que nous entreprenons

ici, en recourant au modèle logistique à deux paramètres[1] de la théorie des tests, exprimé par :

$$P_i(\theta) = \frac{1}{1 + \exp\left(-a_i(\theta - b_i)\right)} \qquad (1)$$

où θ, notre paramètre d'intérêt, dénote la « capacité », ou niveau d'aptitude du répondant ; i (variant de 1 à n) est un numéro d'item, à réponse 0 ou 1 ; $P_i(\theta)$ est la probabilité que le répondant à niveau θ produise la réponse « 1 » ; a_i et b_i sont respectivement les paramètres de discrimination et de difficulté de l'item i.

On pose de coutume que la capacité θ se distribue normalement dans la population, et le modèle de métrique proposé est la distribution normale de moyenne μ et de variance σ². À toutes fins pratiques, la distribution normale standard ($\mu_\theta = 0$, $\sigma_\theta^2 = 1$) est retenue, dont la densité de probabilité est indiquée par :

$$\phi(\theta) = \frac{1}{\sqrt{2\pi}} \exp(-\theta^2/2) \qquad (2)$$

Quant aux paramètres d'items a_i et b_i, leurs distributions ne sont pas prédéfinies. On pose ordinairement $0 < a_i$ et $b_i \in \{ -4, +4 \}$, les b_i ayant la même métrique que θ.

Objectifs de l'étude

L'objectif premier de l'étude est de mettre en jeu différents estimateurs de θ, dénotés $\hat{\theta}$, dans l'arène d'une simulation Monte Carlo à grande échelle, et d'en faire apparaître les propriétés psychométriques comparatives. Ces propriétés concernent le biais et la variance, auxquels nous ajouterons le taux de capture. L'organisation de l'étude Monte Carlo et la définition des mesures observées se trouvent plus bas.

Comme objectif auxiliaire, nous souhaitons évaluer deux estimateurs inédits, basés sur l'estimateur MV, que nous proposons au lecteur : un estimateur « winsorisé » (WINS) et un estimateur « extrapolé » (EXT, Germain et Laurencelle, 2010), ces deux estimateurs combattant (efficacement?) une tare rédhibitoire de MV que nous explicitons ci-après.

Examinons d'abord les estimateurs en lice.

Les estimateurs retenus

Estimateur par maximum de vraisemblance (MV)

Le principe d'estimation par le maximum de vraisemblance, très important en statistique (Kendall et Stuart, 1979), est basé sur la structure paramétrique du problème, c'est-à-dire essentiellement sur le modèle de mesure (1). La fonction de vraisemblance de la variable de capacité (θ) est donnée par :

$$L(\theta; \mathbf{x}) \prod_{i=1}^{n} P_i(\theta)^{x_i} \left(1 - P_i(\theta)\right)^{1-x_i}$$

où **x** est le vecteur de réponse du sujet, x_i étant la réponse du sujet à l'item i. L'estimateur de θ par maximum de vraisemblance est la valeur qui maximise cette fonction et est noté par :

$$\hat{\theta}_{\mathrm{MV}} = \operatorname{argmax}[L(\theta; \mathbf{x})]$$

La valeur cherchée, $\hat{\theta}_{\mathrm{MV}}$, est la solution en θ de $\sum a_i(1 - P_i(\theta)) \cdot x_i - a_i P_i(\theta)(1 - x_i) = 0$.

Remarquons que cet estimateur (MV) n'est pas défini lorsque le score du sujet est 0 ou n. En effet, par exemple pour un score total parfait de $X = n$, la fonction de vraisemblance converge vers 1 pour un argument θ allant à l'infini. Certains logiciels (comme LOGIST[MD], voir Wingersly, Barton et Lord, 1982) posent dans ce cas des valeurs « poteaux », comme par exemple $\hat{\theta} = +5$ pour $X = n$.

L'estimation a généralement un biais s'atténuant avec le nombre d'items n. Lord (1983, p. 237, éq. 28) a identifié ce biais : nous y revenons plus bas, à la faveur de l'estimateur de Warm (1989).

Estimateur par mode bayésien (BME[2])

Admettant que nous disposions de la distribution a priori de la variable θ, soit $\phi(\theta)$ (2), sa fonction de densité a posteriori selon le vecteur de réponse $\underline{x}$ sera donnée par :

$$g(\theta|\mathbf{x}) = \frac{L(\theta, \mathbf{x})\phi(\theta)}{\int_{-\infty}^{\infty} L(\theta, \mathbf{x})\phi\theta d\theta}.$$

L'estimateur de θ par mode bayésien est la valeur qui maximise cette fonction (ou son numérateur) et il est noté par :

$$\hat{\theta}_{\mathrm{BME}} = \operatorname{argmax}(g(\theta|\mathbf{x}));$$

sa valeur est la solution en θ de $\sum_{i=1}^{n} a_i(1 - P_i(\theta)) \cdot x_i - a_i P_i(\theta)(1 - x_i) - \theta = 0$. Cet estimateur est toujours bien défini même pour les scores parfaits car la densité normale tend vers 0 pour θ allant à plus ou moins l'infini.

Estimateur par espérance a posteriori (EAP)

L'estimateur de θ désigné par l'expression « espérance a posteriori » (EAP) utilise la moyenne (ou espérance) de la distribution a posteriori plutôt que son mode. Il est donné par :

$$\hat{\theta}_{\mathrm{EAP}} = \int_{-\infty}^{\infty} \theta \cdot g(\theta|\mathbf{x}) d\theta.$$

[1] C'est le modèle apparemment le plus en usage, par rapport au modèle à 1 paramètre (ou modèle de Rasch), $1/[1 + \exp(-(\theta - b_i))]$, et celui à trois paramètres, $c_i + (1 - c_i) / [1 + \exp(-a_i \cdot (\theta - b_i))]$, c_i étant désigné « pseudo chance » ou « gratuité ».

[2] Aussi désigné MAP (« maximum a posteriori »).

Cet estimateur est aussi toujours bien défini même pour les scores parfaits.

Estimateur par maximum de vraisemblance pondérée de Warm (WARM)

Afin de corriger le biais de l'estimateur par maximum de vraisemblance, Warm (1989) pondère la fonction de vraisemblance par sa fonction de biais $w(\theta)$ (Lord, 1983), soit:

$$\hat{\theta}_{\mathrm{WARM}} = \mathrm{argmax}([w(\theta)L(\theta;\mathbf{x})])$$

La fonction de poids doit satisfaire l'équation suivante :

$$\frac{d}{d\theta}\ln w(\theta) = \frac{I'(\theta)}{2I(\theta)},$$

où $I(\theta)$ est la fonction d'information du test et $I'(\theta)$ est sa dérivée. Explicitement, ces deux fonctions sont :

$$I(\theta) = \sum^{n} a_i^2 P_i(\theta)\,(1 - P_i(\theta))\ ;$$

$$I'(\theta) = \sum_{i=1}^{n} a_i^3 P_i(\theta)\,(1 - P_i(\theta))\,(1 - 2P_i(\theta))$$

et la valeur cherchée est la solution en θ de :

$$\sum_{i=1}^{n} a_i\,(1 - P_i(\theta)) \cdot x_i - a_i P_i(\theta)(1 - x_i) + \frac{I'(\theta)}{2I(\theta)} = 0.$$

La pondération ajoutée par Warm a l'effet de rendre cet estimateur du maximum de vraisemblance bien défini pour les scores parfaits ou nuls. Notons qu'il ne s'agit pas d'un estimateur bayésien puisque rien n'y est explicitement supposé sur la distribution propre de θ.

Estimateur du maximum de vraisemblance winsorisé (WINS)

L'estimateur MV, on l'a vu, diverge vers $+\infty$ lorsque $X = n$ (ou $-\infty$ lorsque $X = 0$) et n'est alors plus utilisable. Il existe, pour de tels cas, un principe statistique qu'on peut appeler à l'aide, la « winsorisation » (voir p. ex. Tukey, 1962) : ce principe consiste à donner à l'estimateur $\hat{\theta}$ la valeur plausible et observée la plus proche. L'estimateur WINS est ainsi défini par :

$$\hat{\theta}_{\mathrm{WINS}} = \begin{cases} \hat{\theta}_{\mathrm{MV}} & \text{si } 0 < X < n \\ \mathrm{argmax}[L(\theta;\mathbf{x})|X = n-1] & \text{si } X = n \\ \mathrm{argmin}[L(\theta;\mathbf{x})|X = 1] & \text{si } X = 0 \end{cases}$$

Noter que, pour $X = n$ (ou 0), la valeur maximale (ou minimale) de θ se produit lorsque l'item supprimé (ou ajouté) correspondant est celui doté du coefficient a_i le plus faible.

Estimateur du maximum de vraisemblance extrapolé (EXT)

Tout comme l'estimateur WINS, ce nouvel estimateur est identique à MV, sauf dans le cas d'un score parfait ($X = n$) ou nul ($X = 0$). Pour un score parfait, basé sur n items réussis, la valeur θ est extrapolée en invoquant un n+1e item imaginaire, plus difficile que les n items présentés, et qui serait échoué. Pour qu'un tel événement soit statistiquement plausible, les conditions suivantes s'imposent :

1) $\hat{\theta}_{\mathrm{EXT}}(X = n) > \hat{\theta}_{\mathrm{MV}}(X = n-1)$ { ou $\hat{\theta}_{\mathrm{EXT}}(X = 0) < \hat{\theta}_{\mathrm{MV}}(X = 1)$ }
   La valeur $\hat{\theta}$ pour un score parfait doit être plus forte que celle accordée pour n'importe quel score partiel (ou le contraire, pour un score nul) ;
2) *$P_{n+1}(\hat{\theta}_{\mathrm{EXT}}) \le P_i(\hat{\theta}_{\mathrm{EXT}})$ pour $i = 1 .. n$*
   La probabilité de « réussir » l'item virtuel ajouté doit être plus basse que ou égale à celle de réussir n'importe quel parmi les n items du test (ou plus haute, pour un score nul) ;
3) *$\hat{\theta}_{\mathrm{EXT}}$ ($X = n$) est la valeur la plus centrée (la moins extrême) qui satisfait à la fois les conditions 1 et 2. Cette clause a pour seul but de garantir l'unicité de la solution.*

Le système à résoudre pour obtenir $\hat{\theta}_{\mathrm{EXT}}$ est alors la solution de :

$$\text{pour } 0 < X < n: \quad \sum_{i=1}^{n} a_i\,(1 - P_i(\theta)) \cdot x_i - a_i P_i(\theta)(1 - x_i) = 0$$

$$\text{pour } X = n: \quad \sum_{i=1}^{n} a_i\,(1 - P_i(\theta)) - a^* \cdot C_{bas} = 0$$

$$\text{pour } X = 0: \quad a^* \cdot C_{haut} - \sum_{i=1}^{n} a_i\,(1 - P_i(\theta)) = 0$$

où $a^* = \min(a_i)$, $C_{bas} = min(P_i(\theta))$ et $C_{haut} = max(P_i(\theta))$. Quant au paramètre b_{n+1}, on peut le déduire par $b = \hat{\theta}_{\mathrm{EXT}} + \log_e[(1 - C)/C] / a^*$.

Un exemple

Un exemple simple servira à illustrer les différents estimateurs. L'exemple porte sur un test comportant $n = 5$ items, dont les caractéristiques établies sont :

Item	1	2	3	4	5
a_i	0,8	1,1	0,5	0,7	1,2
b_i	–3,5	–1,75	0,0	1,75	3,5

Pour ce test, nous proposons les protocoles de réponses suivants :

Protocole	item 1	item 2	Item 3	item 4	item 5
$X = 3$	1	1	1	0	0
$X = 5$	1	1	1	1	1
$X = 0$	0	0	0	0	0

pour lesquels les différents estimateurs en lice présentent les valeurs indiquées.

Protocole	MV	BME	EAP	WARM	WINS	EXT
$X = 3$	0,82	0,19	0,21	0,82	0,82	0,82
$X = 5$	indét.	1,64	1,62	4,72	4,27	4,58
$X = 0$	indét.	–1,51	–1,50	–4,93	–3,92	–4,53

Noter que, pour l'estimateur fondamental du maximum de vraisemblance (MV), la valeur θ est indéterminée lorsque le protocole est uniforme, tendant vers $+\infty$ pour $X = n$ et vers $-\infty$ pour $X = 0$.

Conditions et paramètres de l'étude

Pour garder à l'étude comparative un format raisonnable, nous imposons d'emblée deux limites importantes : nous n'étudions que le modèle TRI à deux paramètres, les paramètres a_i et b_i, et nous nous plaçons dans un contexte où le test et ses items sont déjà « calibrés », c'est-à-dire, les valeurs des indices a_i et b_i sont connues et n'ont pas à être estimées[3].

Globalement, par des simulations Monte Carlo, nous avons étudié :

- des tests comportant $n = 11$, 31 et 101 items, à réponses 0 ou 1 ;
- le paramètre a_i a été généré selon une distribution Gamma(10 ; 0,1), produisant des valeurs de caractéristiques $\mu = 1$, $\sigma^2 = 0{,}100$, $\gamma_1 = 0{,}632$ et $\gamma_2 = 0{,}600$ (pour γ_1 et γ_2, voir (3)), soit des valeurs plausibles dans le domaine des applications TRI ; les valeurs a_i présentées dans le petit exemple à 5 items ci-dessus sont représentatives ;
- le paramètre b_i a été étalé régulièrement de $b_1 = -3{,}5$ à $b_n = +3{,}5$ par intervalles égaux à $7 / (n-1)$; cette structure représente un test équilibré, de niveau de difficulté modéré selon Kim et Nicewander (1993) ;
- le paramètre θ a été étalé régulièrement de $\theta = -3{,}5$ à $\theta = 3{,}5$ par intervalles de 0,1, plutôt que d'être généré par le biais d'une variable de loi normale (2).

Les ingrédients du modèle TRI à deux paramètres étant ainsi définis, le programme informatique simulait alors des protocoles aléatoires répondant à ce contexte.

La structure des simulations a été la suivante, pour chacun des formats de tests ($n = 11$, 31 et 101 items). Pour 100 cycles, un ensemble de n valeurs a_i et b_i étaient générées, puis, pour chaque cycle, 100 protocoles de réponses simulés étaient générés pour chacun des 71 niveaux de valeur θ.

L'ensemble des protocoles simulés pour chaque niveau de n, au nombre de 710 000, repose donc sur les 10000 protocoles associés à chacune des 71 valeurs prescrites de θ.

Enfin, pour chaque protocole, ont été obtenues les valeurs des estimateurs MV, BME, EAP, WARM, WINS et EXT, en plus du score brut X, obtenu en sommant les items.

Résultats

Le plan d'analyse et les estimateurs comparés

Parmi les sept estimateurs du niveau d'aptitude mentionnés ci-dessus, les sorties informatiques de l'étude font clairement apparaître trois groupes, pour chacun desquels nous proposons de retenir un estimateur type. Ce sont :

- BME et EAP, dont nous retenons BME
  La corrélation (r) entre les 710 000 estimations produites pour les deux estimateurs déborde toujours 0,9999 pour les protocoles à 11, 31 et 101 items, et elle déborde 0,99999 pour les moyennes regroupées aux 71 niveaux θ. De plus, elle excède 0,9995 pour les 71 niveaux de biais. Sans être identique, le comportement des deux estimateurs est hautement comparable, et nous retenons BME à toutes fins pratiques.
- MV, WINS, EXT, dont nous retenons EXT
  Rappelons d'abord que WINS et EXT sont des estimateurs du maximum de vraisemblance (MV) *prolongés*, en ce qu'ils habilitent l'estimation pour des protocoles à score parfait ou nul, rendant obsolète l'estimation MV simple. D'autre part, la corrélation (r) entre les estimations WINS et EXT atteint 0,99950, 0,99999 et 1 pour les protocoles de 11, 31 et 101 items, alors qu'elle atteint 0,99995, 0,99999 et 1 pour les moyennes associées aux 71 niveaux θ. Quant aux biais associés aux 71 niveaux θ, les corrélations respectives sont de 0,97236, 0,99998 et 1. En fait, dans le contexte paramétrique de l'étude, nous avons observé 3,46% de scores limites ($X = 0$ ou $X = n$) pour les protocoles à $n = 11$ items, 0,14% pour 31 items et 0% pour 101 items. Des 3,46% scores associés à $n = 11$ items, 2,36% proviennent des niveaux θ égaux ou supérieurs (en valeur absolue) à 3,0, ce qui, dans le modèle normal, représenterait 0,27% de la population. Donc, un phénomène assez marginal, même pour un petit protocole de 11 items. À toutes fins pratiques, nous retenons EXT à titre d'estimateur *complet*, c.-à-d. applicable à tous les protocoles, et qui assigne équitablement un score distinctif aux protocoles extrêmes, ce qui n'est pas le cas de WINS.
- WARM
  Estimateur du maximum de vraisemblance lui aussi, WARM, bien que « corrélé » aux estimateurs MV, WINS et EXT, a un comportement divergent en ce qui concerne

[3] Le lecteur est renvoyé à la documentation citée, particulièrement Bertrand et Blais (2004) et Baker et Kim (2004), pour une revue des méthodes d'estimation conjointe (paramètres d'items et paramètre de capacité) disponibles.

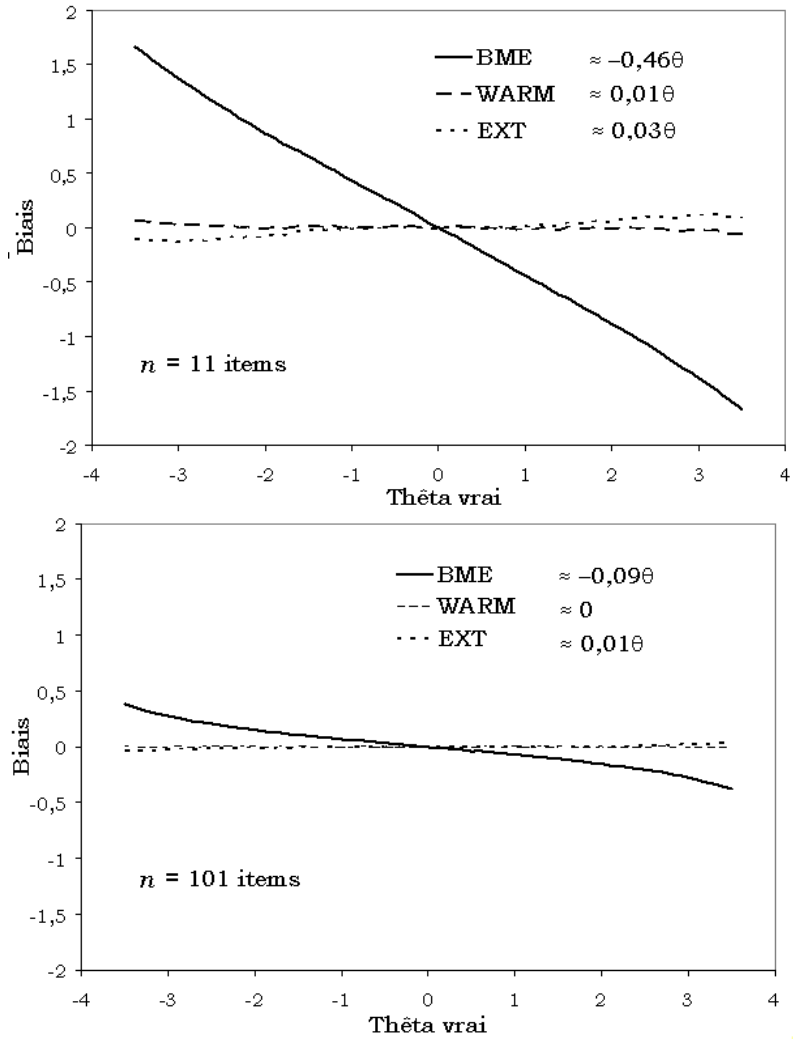

Figure 1. Biais des estimateurs en fonction de θ (haut: n = 11 items; bas: n = 101 items)

le biais. Ainsi, les corrélations (r) entre EXT et WARM pour 710 000 estimations sont de 0,99788, 0,99973 et 0,99999, alors qu'en moyenne pour les 71 niveaux θ, elles sont de 0,99998, 0,99995 et ~1. Cependant, quant aux biais à travers les 71 niveaux θ, ces corrélations sont de –0,70722, 0,37749 et 0,04536. Nous relèverons plus loin d'autres caractéristiques distinctives de l'estimateur WARM.

La rétention de nos trois estimateurs types (BME, EXT et WARM) allégera à la fois nos tableaux, graphiques et discussions ; nous reviendrons sur chaque estimateur en conclusion.

Dans les sections qui suivent, nous comparons les trois estimateurs retenus selon différentes perspectives : d'abord, leurs propriétés de biais, de variabilité et de précision, puis leur capacité prédictive, et enfin leur efficacité de capture.

Biais, variabilité et caractéristiques de distribution

Tableau 1. Valeurs moyenne, minimale et maximale de l'erreur-type des estimateurs selon θ, pour des protocoles à 11, 31 et 101 items

	BME			WARM			EXT		
items	Moy	Min	Max	Moy	Min	Max	Moy	Min	Max
11	0,47	0,40	0,50	0,92	0,87	1,00	0,95	0,88	1,03
31	0,37	0,36	0,41	0,55	0,49	0,71	0,57	0,50	0,80
101	0,26	0,25	0,28	0,30	0,27	0,37	0,30	0,29	0,39

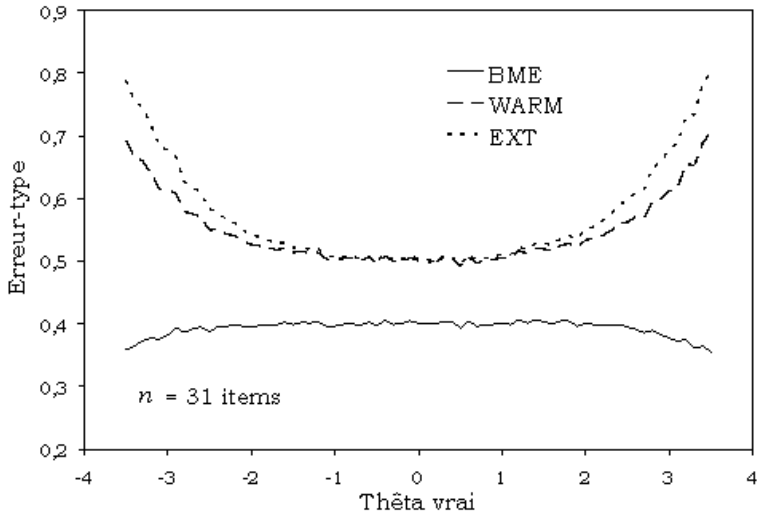

Figure 2. Erreur-type des estimateurs en fonction de θ (n = 31 items)

Tout estimateur $\hat{\theta}$ est une statistique, c.-à-d. une fonction d'estimation basée sur la réalisation d'un événement comportant des ingrédients de hasard, et dont on peut juger la qualité à travers différentes caractéristiques. Dans le modèle paramétrique (1) étudié, θ dénote la capacité, ou niveau d'aptitude, du répondant, et $\hat{\theta}$ en est un estimateur, dont on escompte que la valeur soit proche de la valeur vraie θ. Pour apprécier cette proximité et juger du comportement statistique de $\hat{\theta}$, les différents indices possibles sont le biais, l'erreur-type (ou variance), le REQM (racine de l'erreur quadratique moyenne) et les indices de forme.

Biais. Le biais dénote la différence entre l'espérance, ou moyenne, d'un estimateur et sa valeur vraie, ici E $\{\hat{\theta}\} - \theta$. La Figure 1 illustre ce biais, en fonction des 71 paliers de θ (de –3,5 à 3,5).

Il apparaît clairement que BME, tout comme EAP, a un biais prononcé, un biais centripète ayant pour effet principal de tasser les estimations $\hat{\theta}$ vers la moyenne. La portion linéaire de ce biais engendre un R^2 de plus de 0,999, à quoi on peut ajouter une légère tendance sigmoïde, plus visible dans la Figure 1 (bas) que dans Figure 1 (haut). Le nombre d'items a une influence sur ce biais, avec un coefficient linéaire passant de –0,46 pour 11 items, à –0,23 pour 31 et –0,09 pour 101. Les estimateurs WARM et EXT sont, quant à eux, pratiquement dépourvus de biais.

Erreur-type. Pour une condition paramétrique donnée, p. ex. une valeur θ vraie, et toutes choses étant égales par ailleurs, on préfère un estimateur stable, moins variable, à un estimateur qui fluctue. L'erreur-type est une mesure de cette variabilité, définie comme :

$$\sigma(\hat{\theta}) = \sqrt{E\left\{\hat{\theta} - E(\hat{\theta})\right\}^2},$$

quantité qui est estimée ici par l'écart-type des valeurs $\hat{\theta}$ obtenues. Nous avons choisi d'illustrer cette variabilité en affichant les résultats pour des protocoles à 31 items, à la Figure 2.

Deux constatations ressortent de l'examen de la Figure 2 : le niveau de variabilité de l'estimateur bayésien BME est plus bas que ceux des estimateurs WARM et EXT, et il est aussi plus stable d'un niveau θ à l'autre. Ces constatations s'appliquent aussi à nos résultats pour 11 et pour 101 items. Le Tableau 1 documente ces données.

L'estimateur BME, on le voit, conserve sa plus grande précision par rapport aux autres estimateurs, mais cette supériorité s'amenuise avec un nombre croissant d'items. Par exemple, le ratio des erreurs-types de WARM sur BME passe par 1,96, 1,49 et 1,15 quand le nombre d'items passe par 11, 31 et 101.

REQM. La statistique REQM, qui combine l'information de biais et celle de variabilité, est un indice global de la précision reconnu pour la précision résultante d'un estimateur. On l'obtient simplement par :

$$\text{REQM} = \sqrt{(\text{Biais})^2 + (\text{Erreur-type})^2};$$

cet indice nous informe de quelle distance métrique l'estimateur $\hat{\theta}$ peut s'éloigner de sa cible θ. La Figure 3 montre l'évolution du REQM pour nos trois estimateurs.

Dans le cas des protocoles comptant 11 items, l'estimateur BME affiche un REQM plus bas que WARM dans l'intervalle θ = ±1,7, et que EXT dans l'intervalle ±1,9. Cet intervalle préférentiel se réduit à θ = ±1,6 pour les deux estimateurs avec des protocoles à 31 et 101 items. Rappelons que, dans une population normale standard, censée représenter la répartition des capacités θ, l'intervalle (–1,6 ; 1,6) contient environ 89% des effectifs, de sorte que, du point

Tableau 2. Corrélation entre la valeur vraie θ et chacun des estimateurs, incluant le score brut X

items	X	BME	WARM	EXT
11	0,913	0,917	0,910	0,910
31	0,967	0,969	0,966	0,964
101	0,989	0,990	0,989	0,989

Tableau 3. Erreur-type d'estimation et REQM* associées à la valeur $\hat{\theta}$ prédite depuis θ, sur la base des données issues des estimateurs BME et EXT

	BME	REQM				EXT	REQM			
items	s_e	θ = 0	θ = 1	θ = 2	θ = 3	s_e	θ = 0	θ = 1	θ = 2	θ = 3
11	0,82	0,82	0,94	1,23	1,59	0,85	0,85	0,85	0,85	0,86
31	0,50	0,50	0,56	0,69	0,86	0,55	0,55	0,55	0,55	0,55
101	0,29	0,29	0,30	0,34	0,39	0,30	0,30	0,30	0,30	0,30

*Cet indice est obtenu comme d'habitude, en additionnant s_e^2 et $(\hat{\theta} - \theta)^2$ et en en prenant la racine carrée.

de vue de la *précision de mesure*, l'estimateur BME (comme son associé, EAP) a le plus grand mérite.

À ce moment-ci, il est intéressant de noter que, si l'erreur de mesure dans $\hat{\theta}$, concrétisée par l'erreur-type, fluctue d'une estimation à l'autre, ce n'est pas le cas du biais, lequel est une constante pour un contexte de mesure donné. Ainsi, si on prend, disons, 4 estimations $\hat{\theta}$ d'un répondant, basées sur 4 administrations de protocoles de mesure semblables, l'erreur moyenne résultante aura une erreur-type divisée par $\sqrt{4}$ alors que le biais restera inchangé. La Figure 4 présente un graphique des données correspondantes.

Comme on peut voir, le niveau d'imprécision des trois estimateurs a considérablement baissé et ce, davantage pour les estimateurs WARM et EXT que pour BME, qui reste pénalisé par son biais, lequel est lié seulement au nombre d'items dans le protocole.

Indices g_1 et g_2. Une description complète du comportement des estimateurs doit inclure la forme approximative de leur distribution, telle que révélée par les indices g_1 et g_2. Ces indices sont définis par leurs équivalents paramétriques, γ_1 dénotant l'asymétrie et γ_2 dénotant l'aplatissement (ou voussure) de la forme distributionnelle, soit :

$$\gamma_1 = E(X-\mu)^3/\sigma^3; \gamma_2 = E(X-\mu)^4/\sigma^4 - 3, \qquad (3)$$

leurs valeurs étant estimées par les moyennes équivalentes. La forme « normale », généralement appréciée pour ses propriétés statistiques, correspond à $\gamma_1 = \gamma_2 = 0$.

L'estimateur BME, qui est obtenu par le moyen d'une pondération normale de la fonction de vraisemblance, affiche une asymétrie (g_1) contraire, c.-à-d. d'un signe contraire à θ : positive pour $\theta < 0$ et négative pour $\theta > 0$; comprise dans l'intervalle ±0,35 pour n = 11 items, cette asymétrie s'attenue fortement pour les protocoles plus longs, se confinant alors dans l'intervalle ±0,15. La mesure d'aplatissement, quant à elle, avoisine le zéro.

Les estimateurs WARM et EXT, semblables l'un à l'autre, diffèrent cependant de BME. Leur asymétrie est de même signe que θ, présente une forme sigmoïde et est généralement de plus grande amplitude que celle de BME, l'atténuation selon le nombre d'items étant beaucoup moins forte. Quant à l'indice d'aplatissement, il est globalement plus important que pour BME et, avec le nombre d'items croissant, il évolue vers une courbe en U, avec de fortes pointes de leptokurtose dans le voisinage de θ = ±3,5.

Pour résumer les données se rapportant à la forme de distribution, l'estimateur pondéré BME apparaît le plus « sympathique », proche d'une forme normale, avec une évolution rapide vers cette forme selon le nombre d'items, ce qui n'est pas le cas des estimateurs WARM et BME. Encore faut-il rappeler que, au contraire de BME, les estimations de WARM et EXT jouxtent les extrémités du domaine θ, extrémités qui constituent un mur difficilement franchissable[4] et qui expliquent sans doute la leptokurtose constatée dans ces segments du domaine θ ; BME, d'autre part, grâce à son biais centripète, se tient prudemment en retrait de ces limites et peut donc déployer ses estimations sans contrainte.

Prédiction linéaire à partir du θ vrai

Nos données comportent des séries de 710 000 valeurs $\hat{\theta}$ estimées à partir d'un répertoire de valeurs vraies θ, de sorte que nous pouvons étudier la prédictivité des unes par les autres.

Dans un premier temps, sur la base des 710 000

[4] D'autant plus que les indices de difficulté b_i sont eux aussi bornés aux limites ±3,5.

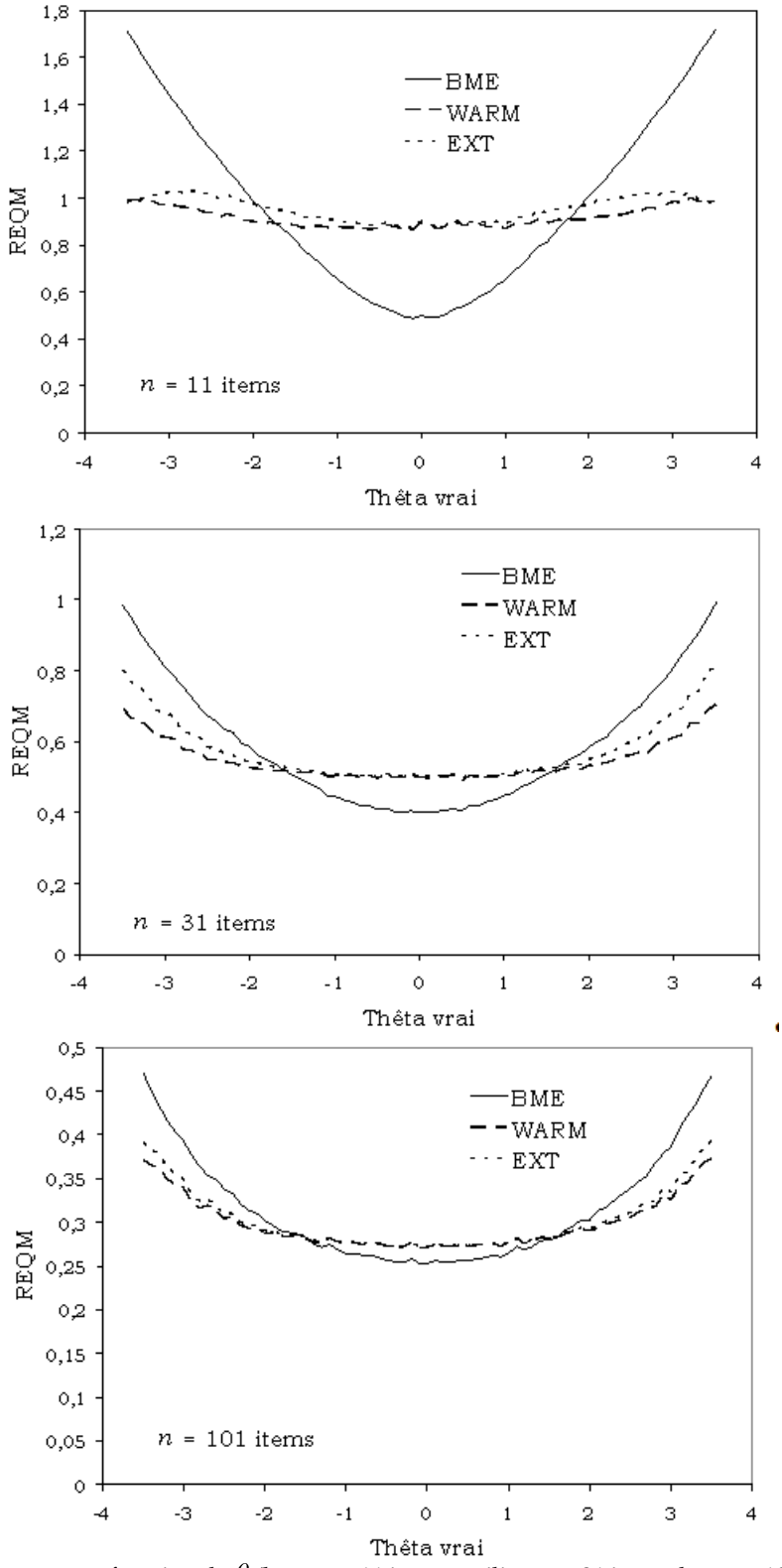

Figure 3. REQM des estimateurs en fonction de θ (haut: n = 11 items ; milieu: n = 31 items; bas: n = 101 items)

estimations produites, nous avons calculé la corrélation simple entre le θ vrai et chacun des estimateurs, BME, WARM et EXT, en plus du score brut, X : le Tableau 2 rapporte les coefficients.

Cette corrélation globale donne un léger avantage à l'estimateur BME, tandis le score brut X se tient tout près des autres, comme il a souvent été démontré (Kim et Nicewander, 1993; Galdin et Laurencelle, 2010).

Nous avons donc établi les fonctions de régression linéaire « $\hat{\theta}' = b \cdot \theta + a$ » pour les $\hat{\theta}$ issus de chacun de nos

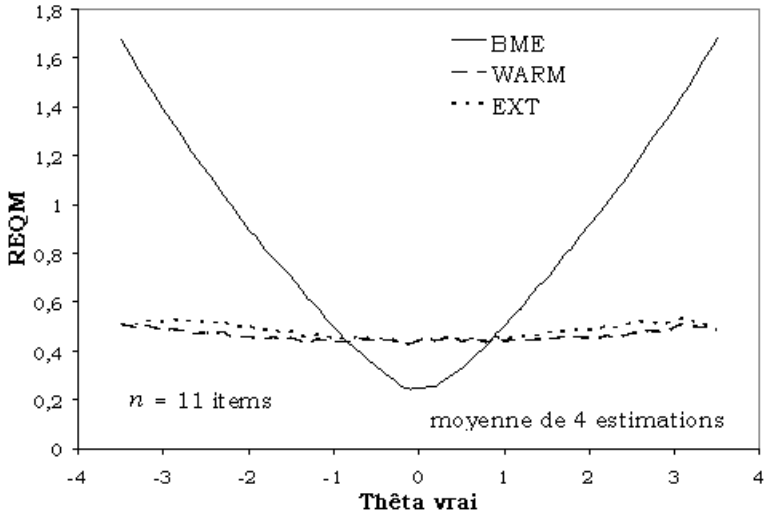

Figure 4. REQM des estimateurs en fonction de θ (n = 11 items) pour la moyenne de 4 estimations semblables

estimateurs ; le Tableau 3 résume l'information en comparant les résultats des estimateurs BME et EXT, le comportement de EXT mimant exactement celui de WARM pour cette analyse.

L'erreur-type d'estimation, une caractéristique globale de la fonction de régression, est moindre pour BME que pour EXT, et cet avantage, bien sûr, s'affirme pour la prédiction à $\theta = 0$, valeur à laquelle le biais de BME est nul. On voit cependant que, dès $\theta = 1$ (ou -1), l'avantage au REQM passe à EXT, qui reste non biaisé alors que BME pâtit par son biais centripète. L'allure des résultats serait-elle un effet de l'inhomogénéité de variance, tel qu'illustré à la Figure 2 pour les protocoles de 31 items? L'analyse détaillée par paliers de θ montre des patterns du REQM semblables à ceux présentés ci-dessus : même si la variance d'erreur de EXT augmente pour un niveau θ s'écartant de 0, cet effet est dominé par celui du biais de BME, la prédominance de BME apparaissant alors à partir de $\theta > 1$. Le Tableau 3 montre aussi que, tel qu'attendu, la différence entre les estimateurs s'estompe pour des protocoles plus longs.

Ces résultats font écho à ceux déjà présentés sur l'erreur-type et le REQM des estimateurs.

Taux de capture

Posant pour vrai le modèle paramétrique de la TRI, nous sommes intéressés, à partir d'un protocole de réponses et de l'ensemble des paramètres (a_i, b_i) d'items, à obtenir une estimation $\hat{\theta}$ et ainsi *cerner* le mieux possible le niveau d'aptitude θ du répondant. Nous avons concrétisé cette idée de cerner par le concept opérationnel de « capture » et par une mesure du taux de capture, définie comme :

$$\Pr\left\{\hat{\theta} \in]\theta - d; \theta + d]\right\}, \qquad (4)$$

soit la probabilité que l'intervalle borné par $\theta - d$ et $\theta + d$ contienne la valeur estimée, où $d > 0$ et dénote la grandeur du demi-intervalle θ : cette probabilité est évaluée pour chaque niveau θ par la proportion d'estimations $\hat{\theta}$ qui s'inscrivent dans l'intervalle ainsi borné. Noter que la capture inverse, à savoir celle pour laquelle l'intervalle ($\hat{\theta} - d$; $\hat{\theta} + d$) contient θ, est aussi possible; cependant, son évaluation est plus difficultueuse, puisque les données ne sont pas structurées selon $\hat{\theta}$, et son résultat est globalement équivalent.

Illustrons l'efficacité de capture de nos trois estimateurs pour un demi-intervalle $d = 0{,}5$: la Figure 5 présente les résultats.

Deux caractéristiques ressortent d'abord de ces figures. Premièrement, le taux de capture de BME est le plus grand, et le meilleur, dans la zone centrale du domaine θ ; c'est aussi la zone où le biais de BME est quasi nul. Secondement, l'efficacité de WARM et EXT est globalement plus haute et plus constante que celle de BME à travers le domaine. D'un autre côté, il n'est que juste de signaler que, si on prend en compte la distribution normale des valeurs θ dans la population, la *moyenne pondérée* du taux de capture de BME apparaît plus élevée. Ainsi, pour 11, 31 et 101 items, elle est de 0,55, 0,73 et 0,94, alors que, pour WARM et EXT, elle est de 0,43, 0,68 et 0,93 : un avantage donc, mais surtout favorable à la portion centrale, moyenne, de la population.

Les analyses faites en appliquant d'autres valeurs du demi-intervalle d aboutissent globalement aux mêmes constatations.

Discussion et conclusion

Les différentes analyses rapportées ci-dessus à propos des estimateurs BME, WARM et EXT ont fait apparaître deux grandes différences :

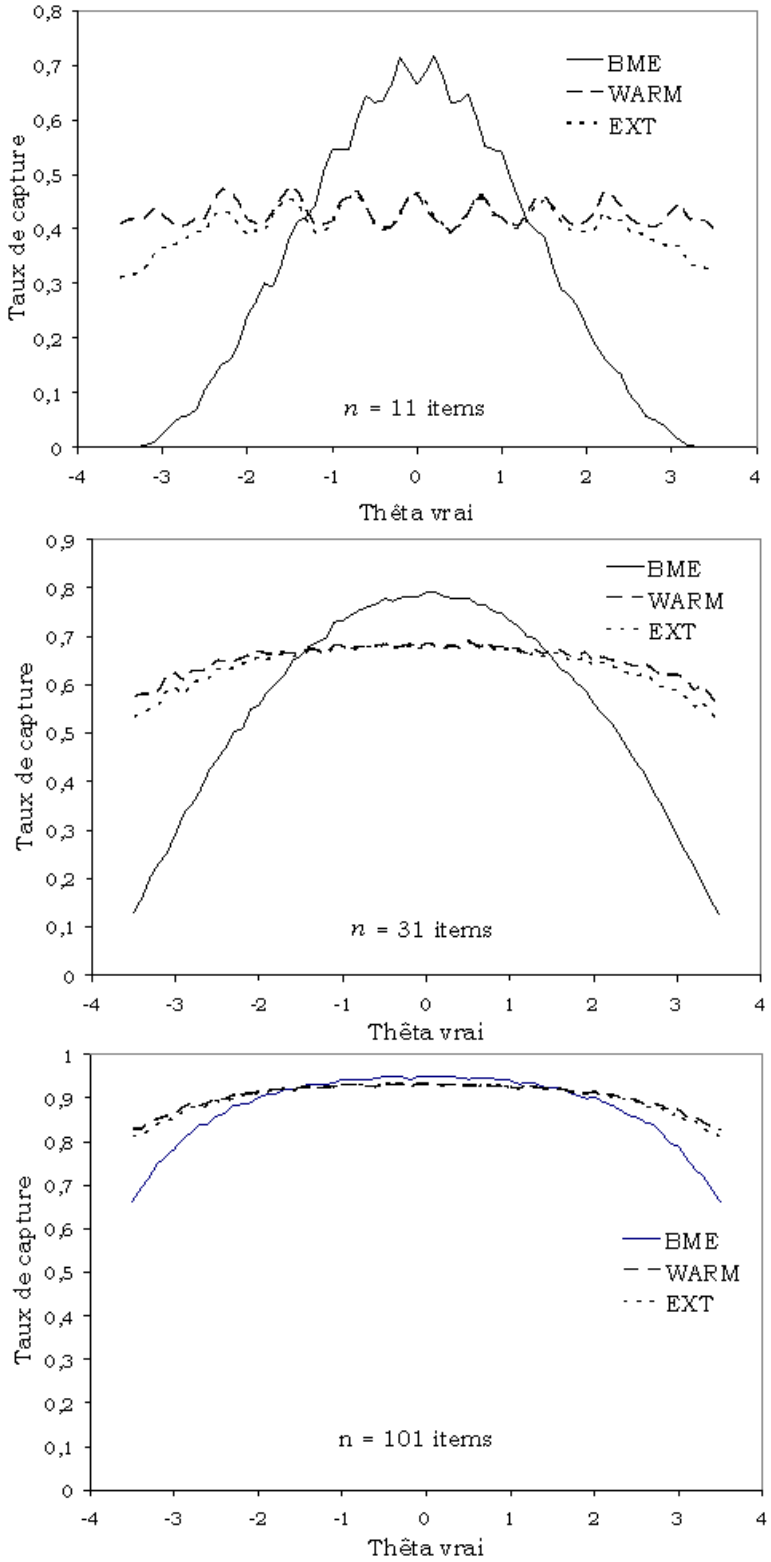

Figure 5. Taux de capture (d = 0,5) des estimateurs pour des protocoles à 11 items (haut), 31 items (milieu) et 101 items (bas), selon le niveau θ

- L'estimateur bayésien BME a de belles propriétés statistiques, qui se sont montrées meilleures que WARM et EXT. En résumé, son erreur-type est plus basse ; cette erreur-type est relativement homogène à travers le domaine θ, au contraire de WARM et EXT pour lesquels elle croît aux deux extrémités du domaine. Finalement, le taux de capture, qui reflète la capacité de l'estimateur à cerner la valeur vraie du niveau d'aptitude θ du

Tableau 4. Moyennes de $\hat{\theta}_{\mathrm{BME}}$ pour différents paliers de θ (moyennes basées chacune sur 10 000 estimations)

items	$\theta = 0$	$\theta = 1$	$\theta = 2$	$\theta = 3$
11	0,00	0,56	1,11	1,62
31	0,00	0,80	1,57	2,28
101	0,00	0,92	1,85	2,72

répondant, est *globalement* plus élevé pour BME que pour WARM et EXT, grâce à une performance excellente au centre du domaine θ. D'un autre côté, même s'il est légèrement plus faible en moyenne pondérée (au contraire de la moyenne simple), le taux de capture de WARM et EXT reste à peu près égal tout le long du domaine θ.

- L'estimateur bayésien BME est fortement biaisé, au contraire de WARM et EXT : ce biais est centripète et s'atténue en fonction inverse du nombre d'items. Pour cette raison, l'erreur résultante, que représente l'indice REQM, est plus élevée pour BME que pour les autres estimateurs, sauf dans la portion centrale du domaine θ, là où le biais joue le moins.

Lord (1983), on l'a vu, a identifié le biais de l'estimateur MV, que d'autres tels que Warm (1989) et Samejima (1993a, 1993b) se sont ingéniés à corriger. Il s'agit là toutefois d'un biais relativement léger[5], difficilement comparable à celui de BME, comme le montre la Figure 1. D'ailleurs, Samejima (1969) et Lord (1986) constatent que la pondération bayésienne, p. ex. dans les estimateurs BME et EAP, *ajoute du biais*. Ce biais provient de la fonction de pondération, laquelle agit de deux manières : elle répartit les valeurs $\hat{\theta}$ possibles autour d'une loi (normale) unimodale et symétrique, ce qui amène un tassement de $\hat{\theta}$ vers le mode de cette fonction, et elle fixe (arbitrairement) ce mode, ordinairement sur la valeur $\mu(\theta) = 0$. La Figure 6 illustre ces effets.

L'exemple illustré en Figure 6 correspond à un test comprenant 11 items, les paramètres d'items étant semblables à ceux exploités dans nos simulations ($0{,}5 \leq a_i \leq 1{,}3$; $-3{,}5 \leq b_i \leq 3{,}5$), et un protocole avec un seul item échoué ($X = n-1 = 10$). La fonction de vraisemblance apparaît en ligne hachurée. L'estimateur du maximum de vraisemblance (MV), identique ici à EXT et WINS, est de 3,98. La fonction de pondération $\phi(\theta)$, tracée en pointillé, est appliquée à la fonction de vraisemblance et produit la fonction bayésienne, en ligne pleine, ramenant $\hat{\theta}$ vers le centre, en fait vers $\mu(\theta) = 0$, ce qui produit $\hat{\theta}_{\mathrm{BME}} = 1{,}76$: cet effet centripète est analogue au *shrinkage* décrit dans Baker et Kim (2004) pour l'estimation conjointe des paramètres d'items. Comme le signale Lord (1986), la « population » à laquelle l'estimation bayésienne confronte le répondant (et son protocole) est posée, voire fixée, par $\mu(\theta)$, une valeur arbitraire. Par exemple, eût-on utilisé $\mu(\theta) = 1$ que la courbe pointillée à la Figure 6 se fût déplacée d'une unité vers la droite, et $\hat{\theta}_{\mathrm{BME}}$ eût été de 2,27 !

Dans le cas d'une pondération telle que celle appliquée dans les estimateurs BME et EAP, qu'entend-on par « estimation », que veut-on estimer, et à quel type de population faisons-nous référence? Si le but de l'estimation était, à partir d'un ou de quelques protocoles tirés au hasard, de trouver le niveau d'aptitude typique d'une *population générale* et d'en délimiter la moyenne, alors l'application d'un modèle plausible de population générale, comme la normale N_θ (0 ; 1), conviendrait. Au contraire, en psychologie différentielle comme en psychométrie, le but est de repérer le niveau d'aptitude particulier d'un répondant afin de le situer le plus précisément possible parmi la population générale et trouver sa valeur distinctive. Le protocole de réponses fournies est une réalisation stochastique qui dépend du niveau θ, une variable latente libre : le fait de rapporter le protocole à une population arbitraire et de lui asservir l'estimation de θ contrevient à cet objectif. C'est pourtant ainsi qu'opère l'estimation bayésienne, créant un biais tel qu'illustré à nouveau pour l'estimateur BME, au Tableau 4.

Que penser d'un estimateur, ou d'une méthode de mesure, qui, pour $\theta \neq 0$, non seulement vise à côté (et en-dessous) de la valeur réelle du paramètre, mais qui fournit des estimations *dont le niveau change en fonction du nombre d'items*? Non seulement l'invariance prétendue du modèle est sérieusement mise à mal (Hambleton et coll., 1991 ; Galdin et Laurencelle, 2010), mais l'utilité même de la mesure devient douteuse, notamment pour les répondants qui croiraient posséder un niveau d'aptitude éloigné de 0.

Quant aux estimateurs non bayésiens, soit WINS, EXT et WARM, leurs comportements et propriétés sont assez proches pour rendre le choix difficile. Leurs biais, relativement petits mais non nuls, sont contraires : alors que WARM affiche un biais léger et centripète, le biais de EXT (ou WINS) est centrifuge et plus important. De plus, et globalement, l'estimateur WARM a un niveau d'erreur

[5] Les courbes de biais présentées dans Warm (1989), voir p. 433, figure 1, pour $n = 10$ items, $a_i = 1$ (tous i) et des b_i distribués, sont invraisemblablement développées et n'ont pu être reproduites.

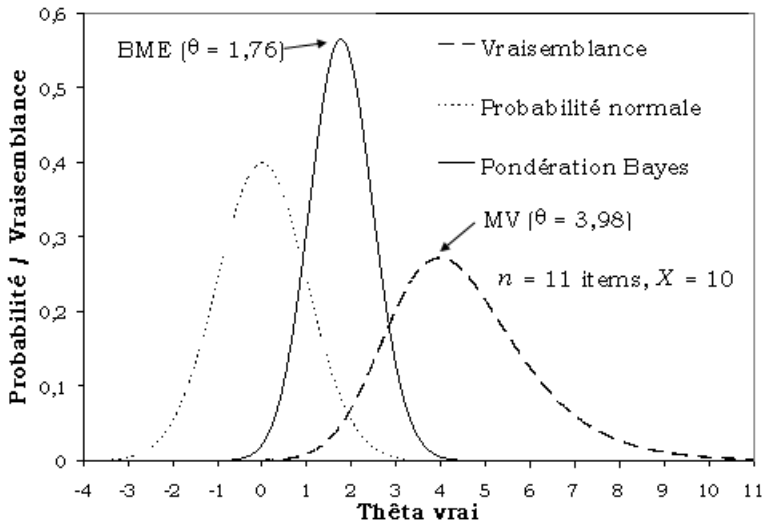

Figure 6. Illustration du déplacement (vers la gauche) de l'estimation $\hat{\theta}_{\mathrm{BME}}$ après pondération par une normale (en pointillé) de moyenne $\mu(\theta) = 0$

légèrement plus bas que ses compétiteurs, et c'est définitivement l'estimateur qu'il nous faut recommander, selon les données disponibles. D'un autre côté, l'estimateur winsorisé (WINS), de calcul très facile, et notre estimateur extrapolé EXT, qui rend justice aux protocoles parfaits ou nuls, tiennent aussi très bien la route.

References

Baker, F. B., Kim, S.-H. (2004). *Item response theory : Parameter estimation techniques* (2e édition). New York : Marcel Dekker.

Bertrand, R., Blais, J.-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items*. Québec : Presses de l'Université du Québec.

Galdin, M., Laurencelle, L. (2010). Assessing parameter invariance in item response theory's logistic two item parameter model : a Monte Carlo investigation. *Tutorials in Quantitative Methods for Psychology*, *6*, 39-51.

Germain, S., Laurencelle, L. (2010, juin). Biais et fidélité de cinq estimateurs de capacité thêta en TRI. Communication présentée au deuxième colloque Méthodes Quantitatives et Sciences Humaines, Université de Montréal, 7 juin 2010.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). *Fundamentals of items response theory*. Newbury Park (CA) : Sage.

Kim, J. K., Nicewander, W. A. (1993). Ability estimation for conventional tests. *Psychometrika*, *58*, 587-599.

Kendall, M., Stuart, A. (1979). The advanced theory of statistics. Volume 2: Inference and relationship (4e édition). New York : MacMillan.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, *14*, 117-138.

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika*, *48*, 233-245.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Psychometrika*, 51, 157-162.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17*, *34* (4, Pt. 20).

Samejima, F. (1993a). An approximation for the bias function of the maximum likelihood estimate of a latent variable for the general case where the item responses are discrete. *Psychometrika*, *58*, 119-138.

Samejima, F. (1993b). The bias function of the maximum likelihood estimate of ability for the dichotomous response level. *Psychometrika*, *58*, 195-209.

Samejima, F. (1998). Some considerations for eliminating biases in ability estimation in computerized adaptive testing. Document présenté au congrès annuel de la American Educational Research Association, San Diego, CA, 13 au 17 avril.

Tukey, J. W. (1962). The future of data analysis. *Annals of Matthematical Statistics*, *33*,1-67.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Wingersly, M. S., Barton, M. A., Lord, F. M. (1982). LOGIST users' guide. Princeton (NJ) : Educational Testing Service.

Manuscript received 12 May 2011.
Manuscript accepted 16 May 2011.