

The distribution of person-fit indices conditional on the estimated proficiency level and the detection of underachievement at a placement test

Gilles Raïche, Université de Moncton
Jean-Guy Blais, Université de Montréal

68th International Meeting of the Psychometric Society

Cagliari
July 07, 2003

Abstract

With some placement tests, many students are seemingly unconcerned with true performance and prefer to underachieve in order to be placed in a less demanding course. Can we detect them? In order to study this problem, we considered three indices frequently used in person-fit research, l_z , *Zeta* and *W*, as well as three others conceived specifically for the underachievement detection, I_{random} , $I_{inverse}$, and $I_{combined}$. One thousand administrations of the placement test TCALS II were simulated for proficiency levels, \mathbf{q} , varying between -2.00 and 2.00 , by increments of 0.25 . A recently developed variation of Bock and Mislevy's EAP estimation with adaptive a priori and adaptive integration interval, was used to estimate the proficiency level. The obtained adaptive EAP estimator seems to be even less biased than Warm's weighted maximum likelihood estimator. Of all the studied indices, l_z was the most efficient to detect underachievement response patterns. I_{random} , $I_{inverse}$, and $I_{combined}$, were less effective than hoped. At the levels where the students are most likely to be interested in being underplaced, the detection rate is higher than 97% when we consider a pattern of 10% hypothetical targetting underachievement aberrant responses. At a 20% aberrance response pattern, the detection is always successful.

Objectives

Molenaar and Hoijtink (1990) discussed the fallacy of the $N(0,1)$ probability distribution assumption of many person-fit indices based on item response theory modelizations. They suggested different methods of dealing with the problem and to obtain useful information about these probability distributions, specifically the criterion value used to detect an aberrant response pattern at a specific false detection rate. One of these methods was to consider a simulation of the null distribution of the person-fit index based on the proficiency level estimate for a specific test.

This strategy was applied to a problem we have been facing for many years in post secondary education in Québec, Canada. In 1993, two English as a second language courses were introduced into the curriculum. A placement test, the TCALS II, is currently administered upon admission to the institution and students are directed to one of the 5 course levels, based on language competency. Many students are seemingly unconcerned with truly learning and prefer to underachieve during their placement test in order to be placed in a less demanding course. This leaves them with more time to take jobs or more demanding courses. Meanwhile, these courses' teachers complain about the disturbing effects of behaviours associated with this so-called underplacement within the classroom.

In order to meet this need, we considered three indices frequently used in person-fit research as well as three others we conceived specifically for this project. The first three indices are the I_z index suggested by Drasgow and Levine (1986; Drasgow, Levine and Williams, 1985), the W index proposed by Wright et Stone (1979) and the *Zeta* index of Tatsuoka (1996).

The three others, specific to this project, are propositions of person-fit indices tailored for the detection of a random response pattern, I_{random} , and of an incorrect response pattern, $I_{incorrect}$, inspired by the optimal indices suggested by Drasgow, Levine and Zickar (1996).

These indices are

$$I_{random} = \frac{P_{random}}{P_{normal}} \quad \text{and} \quad I_{incorrect} = \frac{P_{incorrect}}{P_{normal}}, \quad \text{Equations 1 and 2}$$

where P_{random} and $P_{incorrect}$ are the marginal likelihoods of the aberrant response patterns:

$$P_{random} = \int_{q=-\infty}^{\infty} \prod_{g=1}^n \left(\frac{1}{k}\right)^{X_g} \left(\frac{k-1}{k}\right)^{1-X_g} dq \quad \text{Equation 3}$$

and

$$P_{incorrect} = \int_{q=-\infty}^{\infty} \prod_{g=1}^n P(X_g = 1 | \mathbf{q})^{1-X_g} P(X_g = 0 | \mathbf{q})^{X_g} dq. \quad \text{Equation 4}$$

Note that for equation 4, the exponents of the usual IRT probability function are modified so that an incorrect response is now more probable for a high ability subject.

We also considered a combined index, $I_{combined}$, based on the joint likelihood of random responses and incorrect responses:

$$P_{combined} = \int_{q=-\infty}^{\infty} \prod_{g=1}^n [P_{random} + P_{incorrect}] dq \quad \text{Equation 5}$$

Method

One thousand administrations of the placement test TCALS II were simulated for proficiency levels, \boldsymbol{q} , varying between -2.00 and 2.00 , by increments of 0.25 . Having recently developed our own variation of Bock and Mislevy's EAP estimation with adaptive a priori and adaptive integration interval of an unidimensional Rasch model with 40 quadrature points (Raïche and Blais, submitted), we estimate the proficiency level. With this adaptive EAP estimation method, AEAP, the proficiency level is computed as with adaptive testing, after each item. In order to be adaptive, the a priori estimate is updated to the previous proficiency level estimated at the $k-1$ item and the integration interval is shifted around this new a priori estimate.

The usual EAP estimate is obtained by :

$$\hat{\boldsymbol{q}}_{eap} = \frac{\int_{-\infty}^{\infty} \boldsymbol{q} f(\boldsymbol{q}_{apriori}) \prod_{i=1}^n P(x_i = 1 | \boldsymbol{q})^{x_i} P(x_i = 0 | \boldsymbol{q})^{1-x_i}}{\int_{-\infty}^{\infty} f(\boldsymbol{q}_{apriori}) \prod_{i=1}^n P(x_i = 1 | \boldsymbol{q})^{x_i} P(x_i = 0 | \boldsymbol{q})^{1-x_i}} \quad \text{Equation 6}$$

While the AEAP estimate is computed according to :

$$\hat{\boldsymbol{q}}_{aeap_k} = \frac{\int_{-\infty}^{\infty} \boldsymbol{q} f(\boldsymbol{q}_{k-1}) \prod_{i=1}^k P(x_i = 1 | \boldsymbol{q})^{x_i} P(x_i = 0 | \boldsymbol{q})^{1-x_i}}{\int_{-\infty}^{\infty} f(\boldsymbol{q}_{k-1}) \prod_{i=1}^k P(x_i = 1 | \boldsymbol{q})^{x_i} P(x_i = 0 | \boldsymbol{q})^{1-x_i}} \quad \text{Equation 7}$$

According to Raïche and Blais (submitted, 2001), in an adaptive testing context, the adaptive EAP estimation method could provide a less biased method to estimate the proficiency level. The obtained estimator seems to be less biased than Warm's weighted maximum likelihood estimator (1989).

Each of the 1000 simulations considered the five following strategies for the provisional proficiency estimation: (1) normal response pattern, (2) 10% random response pattern, (3) 20% random response pattern, (4) 10% incorrect response pattern, and (5) 20% incorrect response pattern.

A normal response pattern was generated following:

$$\text{If } P(X_g = 1 | \mathbf{q}) \geq U(0,1), \text{ Then } X_g = 1; \text{ Else } X_g = 0. \quad \text{Equation 8}$$

A random response pattern was generated following:

$$\text{If } \frac{1}{k} \geq U(0,1), \text{ Then } X_g = 1; \text{ Else } X_g = 0. \quad \text{Equation 9}$$

An incorrect response pattern was generated following:

$$\text{If } P(X_g = 1 | \mathbf{q}) \leq U(0,1), \text{ Then } X_g = 1; \text{ Else } X_g = 0. \quad \text{Equation 10}$$

For each person-fit index and each estimated proficiency level, the criterion for a 5% false detection rate was based on the normal responses pattern. Subsequently, this criteria is applied to the detection of underplacement behaviour using each of the four aberrant response strategies.

Results

Generally, the criteria for a 5% false detection rate is far from the theoretical value that would be predicted by a $N(0,1)$ probability distribution. Results are presented in table 1. In fact, looking at the l_z index, for example, its value never reaches the $N(0,1)$ criteria of -1.65 and is mostly far removed from this value. The use of a -1.65 criteria would clearly under estimate the frequency of inappropriate response patterns. Only Tatsuoka's *Zeta* sometimes approaches or is equal to the $N(0,1)$ criteria value of 1.65 .

Table 1
Criteria for a 5% false detection rate for each person-fit index and each estimated proficiency level

$\hat{\epsilon}$	L_z	W	<i>Zeta</i>	I_{random}	$I_{incorrect}$	$I_{combined}$
-2.00	-1.01	1.17	1.45	0.258179	4.29 E-08	0.258179
-1.75	-1.24	1.16	1.54	5.151367	1.25 E-07	5.151367
-1.50	-1.41	1.16	1.60	1.331787	2.67 E-09	1.331787
-1.25	-1.45	1.14	1.60	0.029076	9.54 E-12	0.029076
-1.00	-1.53	1.14	1.49	0.000088	3.75 E-13	0.000088
-0.75	-1.46	1.15	1.54	3.17 E-08	3.61 E-13	3.67 E-08
-0.50	-1.57	1.16	1.65	9.18 E-12	2.29 E-12	6.51 E-11
-0.25	-1.27	1.17	1.49	1.93 E-16	5.11 E-10	5.11 E-10
0.00	-1.21	1.22	1.54	1.33 E-21	0.001713	0.001713
0.25	-0.87	1.25	1.10	1.50 E-27	0.001758	0.001758
0.50	-0.88	1.23	1.26	1.04 E-34	0.001713	0.001713
0.75	-0.94	1.26	1.36	2.37 E-37	0.001713	0.001713
1.00	-0.94	1.26	1.36	2.94 E-40	0.001713	0.001713
1.25	-0.87	1.27	1.14	3.65 E-42	0.003540	0.003540
1.50	-0.81	1.27	1.12	1.38 E-44	0.006632	0.006632
1.75	(n.a.)	(n.a.)	(n.a.)	(n.a.)	(n.a.)	(n.a.)
2.00	-0.77	1.27	0.97	1.31 E-48	0.007862	0.007862

Tables 2 and 3 let us show the efficacy for simulated students of random and inverse strategies to successfully underachieve on the TCALS II. Globally, the strategies are very successful, mostly for course level 4 and 5. At course level 5, many simulated students can even be placed two levels under their true level. Of course, 20% inverse response strategy is the most successful one.

Table 2

Student underplacements according to normal responses, 10 % random responses, and 20 % random responses

Course level	Estimated course level				
	1	2	3	4	5
1 (0 %)	98,57 %	1,43 %	0,00 %	0,00 %	0,00 %
(10 %)	99,63 %	0,37 %	0,00 %	0,00 %	0,00 %
(20 %)	99,93 %	0,07 %	0,00 %	0,00 %	n.d.
2 (0 %)	17,60 %	81,20 %	1,20 %	0,00 %	0,00 %
(10 %)	33,90 %	66,07 %	0,03 %	0,00 %	0,00 %
(20 %)	50,03 %	49,97 %	0,00 %	0,00 %	n.d.
3 (0 %)	0,00 %	12,40 %	80,24 %	7,14 %	0,22 %
(10 %)	0,00 %	31,98 %	67,98 %	0,04 %	0,00 %
(20 %)	0,02 %	57,60 %	42,38 %	0,00 %	n.d.
4 (0 %)	0,00 %	0,00 %	10,70 %	59,93 %	29,37 %
(10 %)	0,00 %	0,00 %	95,00 %	5,00 %	0,00 %
(20 %)	0,00 %	0,17 %	99,83 %	0,00 %	n.d.
5 (0 %)	0,00 %	0,00 %	0,00 %	12,53 %	87,47 %
(10 %)	0,00 %	0,00 %	74,90 %	25,03 %	0,07 %
(20 %)	0,00 %	0,00 %	99,97 %	0,03 %	n.d.

Table 3

Student underplacements according to normal responses, 10 % inverse responses, and 20 % inverse responses

Course level	Estimated course level				
	1	2	3	4	5
1 (0 %)	98,57 %	1,43 %	0,00 %	0,00 %	0,00 %
(10 %)	99,47 %	0,53 %	0,00 %	0,00 %	0,00 %
(20 %)	98,20 %	1,80 %	0,00 %	n.d.	n.d.
2 (0 %)	17,60 %	81,20 %	1,20 %	0,00 %	0,00 %
(10 %)	32,37 %	67,00 %	0,03 %	0,00 %	0,00 %
(20 %)	31,27 %	68,70 %	0,03 %	n.d.	n.d.
3 (0 %)	0,00 %	12,40 %	80,24 %	7,14 %	0,22 %
(10 %)	0,00 %	33,96 %	66,04 %	0,00 %	0,00 %
(20 %)	0,02 %	66,98 %	33,00 %	n.d.	n.d.
4 (0 %)	0,00 %	0,00 %	10,70 %	59,93 %	29,37 %
(10 %)	0,00 %	0,00 %	99,97 %	0,03 %	0,00 %
(20 %)	0,00 %	1,80 %	98,20 %	n.d.	n.d.
5 (0 %)	0,00 %	0,00 %	0,00 %	12,53 %	87,47 %
(10 %)	0,00 %	0,00 %	99,97 %	0,03 %	0,00 %
(20 %)	0,00 %	0,00 %	100,00 %	n.d.	n.d.

Of all the indices, l_z was the most efficient to detect undeplacement response patterns. I_{random} , $I_{incorrect}$ and $I_{combined}$, were less effective that we had hoped. This is why we present only those results obtained with the l_z index in figures 1 and 2. At each of the placement levels we observed the detection rate of successful underplacements, as simulated in this study. At the levels where the students are most likely to be interested in being underplaced, levels 4 and 5, the detection rate is higher than 97 % when we consider a pattern of 10% aberrant responses, either random or incorrect. At a 20% aberrance response pattern, the detection is always successful. At placement level 3, a detection rate of 99% is obtained with a pattern of 10% incorrect responses. With a pattern of 10% random responses this rate drop to 52%. When a 20% aberrant response pattern is considered and the placement levels are over level one, the detection rate is always equal or over

87%. It is therefore clear that the use of the l_z index with this placement test is very effective at the levels of interest.

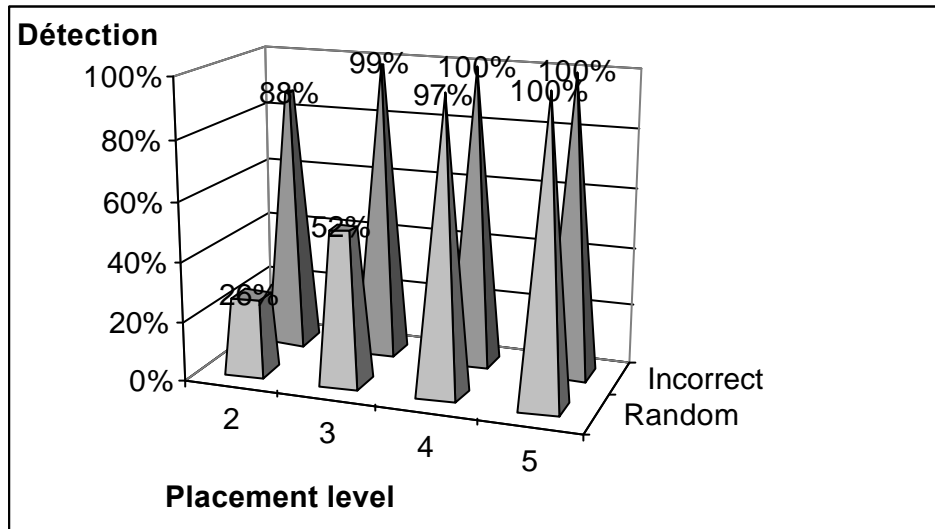


Figure 1. Detection rate of successful underplacements with the l_z index (10 % aberrant responses)

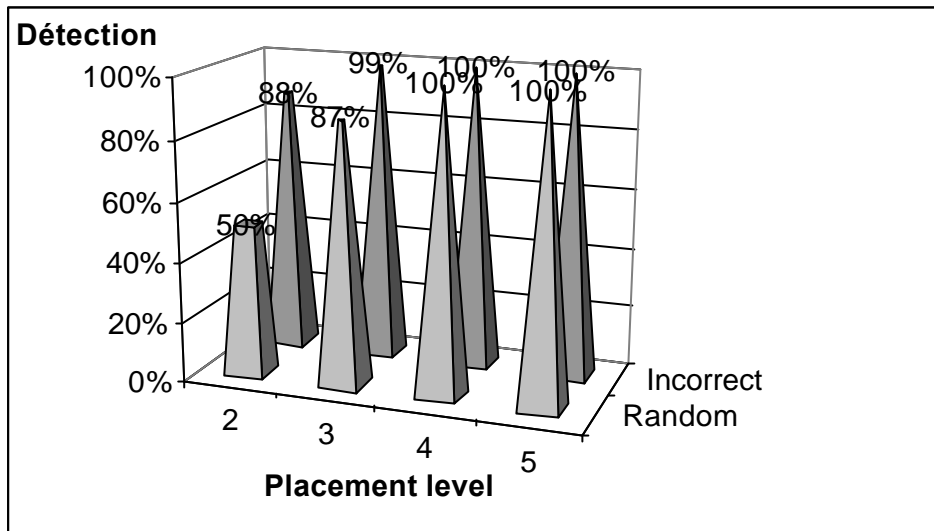


Figure 2. Detection rate of successful underplacements with the I_z index (20 % aberrant responses)

Tables 4 and 5 give the detail detection rate for course level.

Table 4

Detection of underachievement with normal, 10% random responses, and 20% random responses

Course level	Detection efficacy of I_z	
	Non detected	Detected
2 (0 %)	95,08 %	4,92 %
(10 %)	74,24 %	25,76 %
(20 %)	49,63 %	50,37 %
3 (0 %)	95,48 %	4,52 %
(10 %)	48,41 %	51,59 %
(20 %)	13,09 %	86,91 %
4 (0 %)	93,15 %	6,85 %
(10 %)	2,95 %	97,05 %
(20 %)	0,30 %	99,70 %
5 (0 %)	97,07 %	2,93 %
(10 %)	0,27 %	99,73 %
(20 %)	0,00 %	100,00 %

Table 5

Detection of underachievement with normal, 10% inverse responses, and 20% inverse responses

Course level	Detection efficacy of I_z	
	Non detected	Detected
2 (0 %)	94,73 %	5,27 %
(10 %)	69,72 %	30,28 %
(20 %)	12,47 %	87,53 %
3 (0 %)	95,06 %	4,94 %
(10 %)	33,80 %	66,20 %
(20 %)	1,16 %	98,84 %
4 (0 %)	94,33 %	5,67 %
(10 %)	0,00 %	100,00 %
(20 %)	0,00 %	100,00 %
5 (0 %)	95,73 %	4,27 %
(10 %)	0,00 %	100,00 %
(20 %)	0,00 %	100,00 %

Implications

In many studies, the person-fit indices fail to perform satisfactorily because of the erroneous $N(0,1)$ assumption. When the criteria is obtained from the one null probability distribution of the estimated proficiency level specific to a test, the detection rate of aberrant response patterns could be much improved and find more appropriate use in everyday practice. In fact, considering the results obtained, the I_z index will be very efficient in detecting students who succeed in underplacement in English as a second language courses, as measured by the TCALS II administered in many post secondary institutions in Québec.

References

- Bock, R.D., and Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a micro computer environment. *Applied Psychological Measurement*, 6(4), 431-444.
- Dragow, F. et Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, 10(1), 59-67.

- Drasgow, F., Levine, M.V. et Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Drasgow, F., Levine, M.V. et Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9(1), 47-64.
- Molenaar, I.W. et Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55(1), 73-106.
- Raîche, G., Blais, J.-G. (submitted). Considerations about expected a posteriori estimation in adaptive testing: Adaptive a priori, adaptive cotrection for bias, and adaptive integration interval. In G. Englehard (Ed.): *Objective measurement, Theory into practice – Volume 6* Stamford, Connecticut: ABLEX.
- Raîche, G., Blais, J.-G. (2001). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en fonction de deux règles d'arrêt dans le contexte de l'application du modèle de Rasch. *Mesure et évaluation en éducation*, 24(2-3), 23-40.
- Tatsuoka, K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education*, 9(1), 65-76.
- Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54(3), 427-450.
- Wright, B.D. et Stone, M.H. (1979). *Best test design : Rasch measurement*. Chicago : MESA Press.