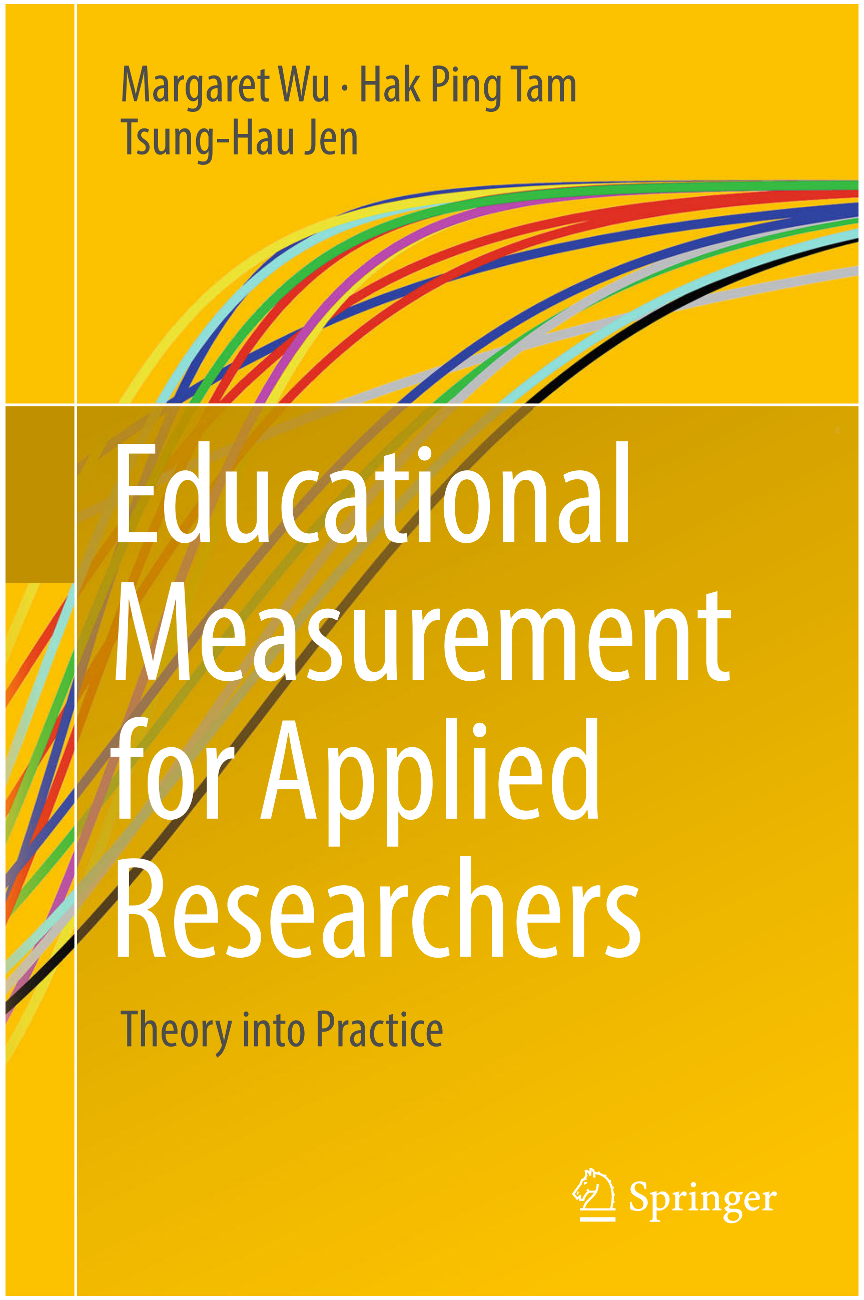
Ajuste ao modelo de Rasch: *residual based statistics infit e outfit*

Ricardo Primi
USF 2019
Programa de Pós Graduação Stricto Sensu em Psicologia

Capítulo 8



Margaret Wu · Hak Ping Tam
Tsung-Hau Jen
Educational Measurement for Applied Researchers
Theory into Practice
Springer

Modelo *vs* dados observados

- Rasch: modelo probabilístico do acerto $P = 1$ em função de dois efeitos principais: theta θ e delta δ
- Mas outros fatores podem afetar a probabilidade de acerto:
  - chute (*guessing*)
  - dependência entre os itens
  - DIF
  - Outras dimensões

Estatísticas de ajuste do modelo (*residual based fit statistics*)

$$z_{ni} = \frac{(x_{ni} - E(X_{ni}))}{(Var(X_{ni}))^{\frac{1}{2}}}$$

where $E(X_{ni})$ is the expected value of the item response, and $Var(X_{ni})$ is the variance of the item response. In the case of the dichotomous Rasch model, $E(X_{ni}) = P_{ni}$ and $Var(X_{ni}) = P_{ni}(1 - P_{ni})$. The standardised residuals have served

Squaring z_{ni} and summing over n (persons), a statistic is derived that can be used as a fit index for item i. Squaring z_{ni} and summing over i (items), a statistic is derived that can be used as a fit index for person n (Wright and Masters, 1982). For

$$\text{Unweighted mean square (outfit)} = \frac{\sum_n z_{ni}^2}{N} = \frac{1}{N}\sum_n \frac{(x_{ni} - E(X_{ni}))^2}{Var(X_{ni})} \qquad (8.3)$$

$$\text{Weighted mean square (infit)} = \frac{\sum_n z_{ni}^2 Var(X_{ni})}{\sum_n Var(X_{ni})} = \frac{\sum_n (x_{ni} - E(X_{ni}))^2}{\sum_n Var(X_{ni})} \qquad (8.4)$$

Interpretando medidas de ajuste

- O valor esperado = 1
- underfit e overfit
- Esses índices testam a propriedade do modelo que supõe discriminação constante
- Não testa o “ruído”
- Rasch não estabelece um valor absoluto para o parâmetro de discriminação. É um valor médio da discriminação dos itens ne teste. Se um idem tem *infit/outfit* alto significa que ele é muito diferente dos outros que estão no *set* de itens

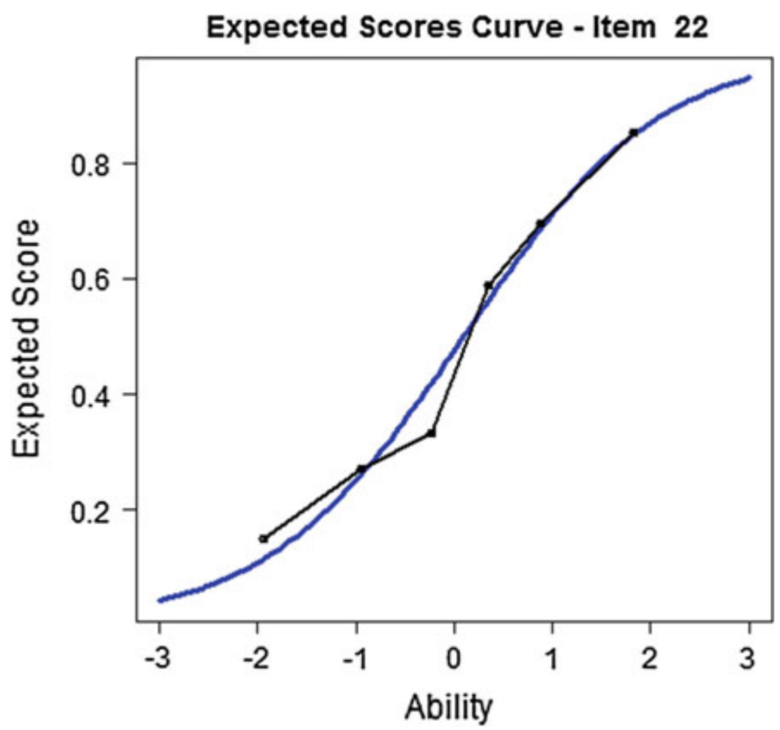


Fig. 8.4 Points of observed ICC are close to the theoretical ICC (infit MNSQ = 1.01)

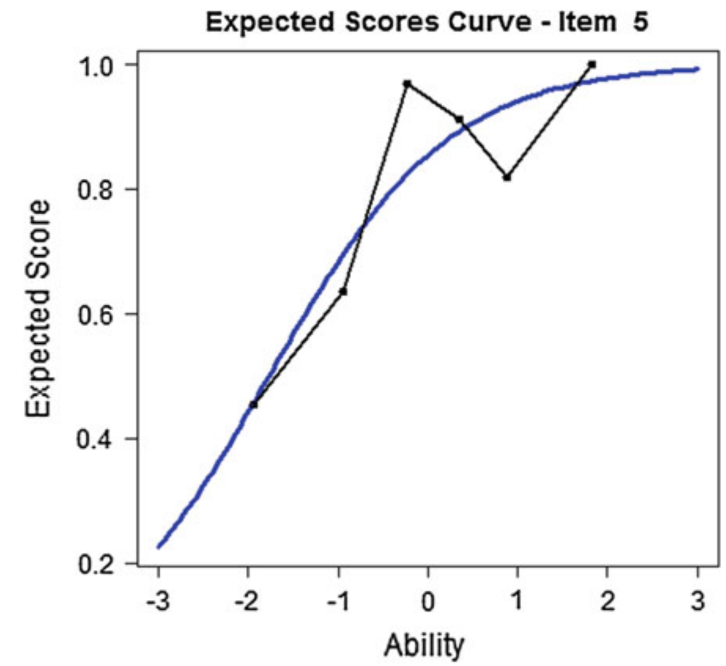


Fig. 8.5 Points of observed ICC "far away" from the theoretical ICC (infit MNSQ = 1.00)

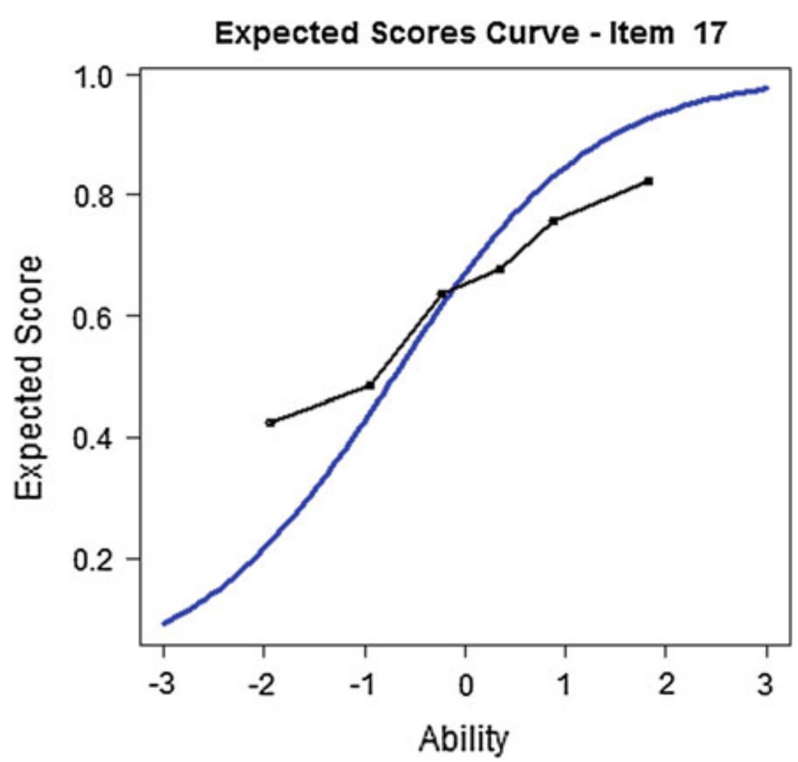


Fig. 8.2 Observed ICC is "flatter" than expected ICC (under-fit) (infit MNSQ = 1.27)

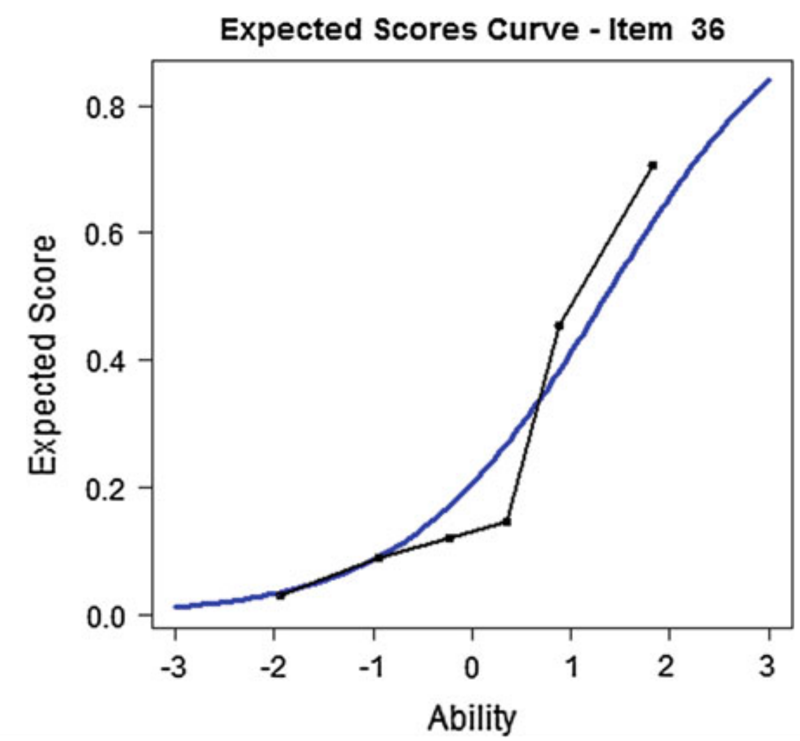


Fig. 8.3 Observed ICC is "steeper" than expected ICC (over-fit) (infit MNSQ = 0.90)

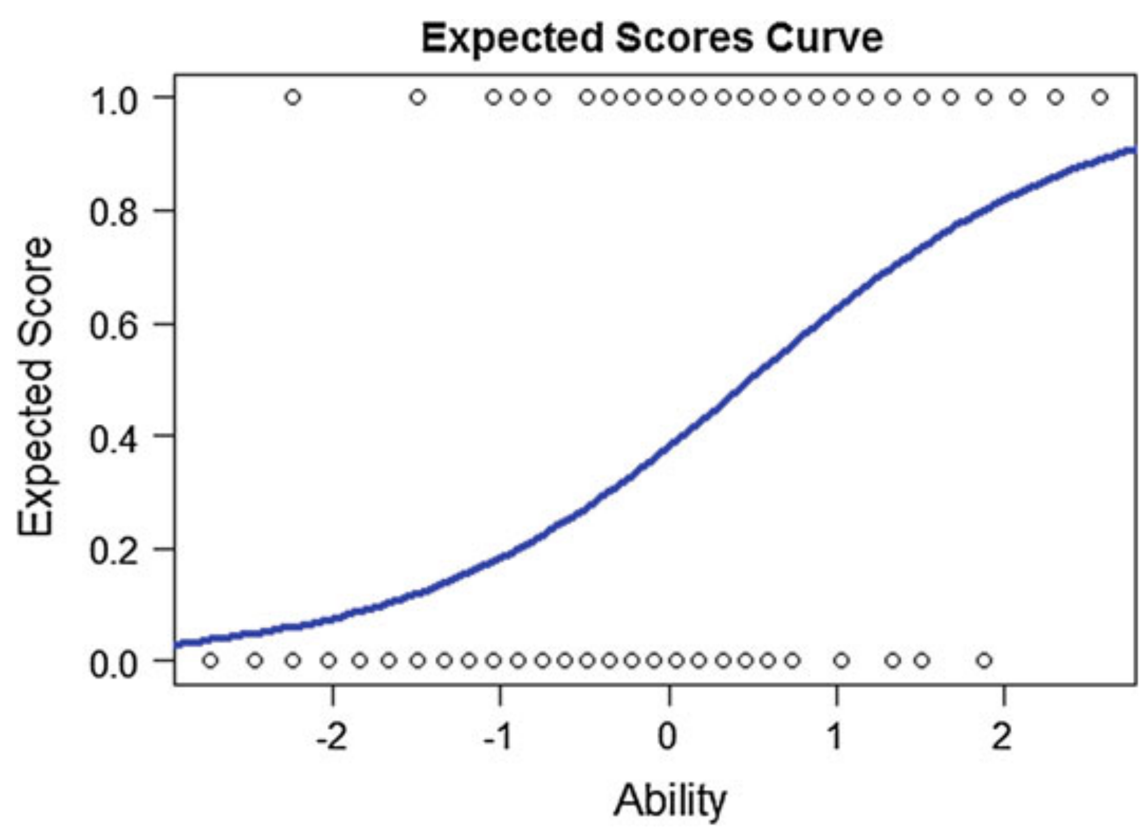


Fig. 8.6 Expected scores curve versus raw data

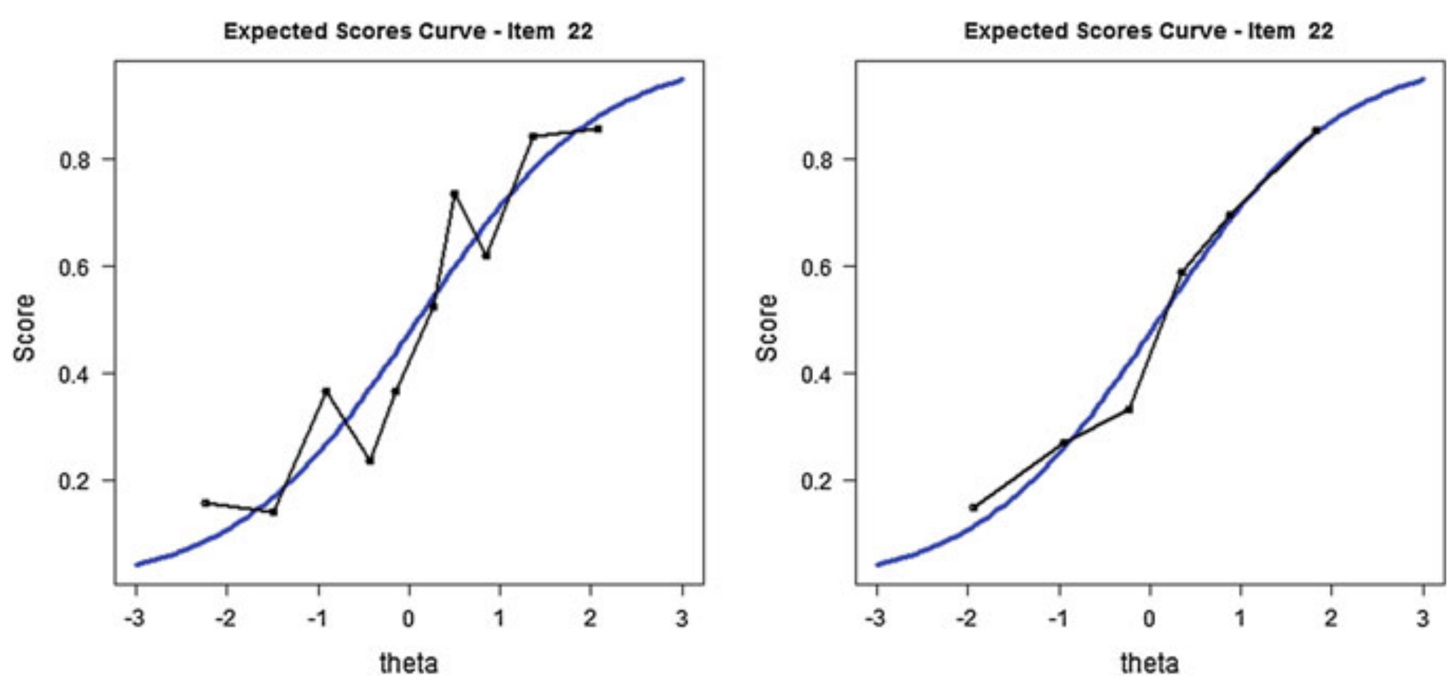


Fig. 8.7 ICCs of an item with 10 ability groups (*left*) and 6 ability groups (*right*)

- ajuste por meio da análise visual não necessariamente indica o ajuste
- diferença entre resíduos de regressão

Análise da distribuição dos índices de ajuste *infit* e *outfit*

- O que significa o "valor esperado = 1"
- Se ajustássemos o modelo de rasch a um conjunto de itens que seguem os pressupostos do modelo, como ficariam distribuídos os índices de ajuste ?
- M=1 Var=2/N

Additional Notes

The numerator in the unweighted fit mean-square statistic, $\sum_n z_{ni}^2$, is an observed value of the sum of squares of random variables Z_{ni} with mean 0 and standard deviation of 1. The random variable, Z_{ni}, has a discrete distribution, as the observed response can only take values 0 and 1 (in the dichotomous case). While Z_{ni} is not a standard normal random variable, when N is large, $\sum_n Z_{ni}^2$ can be regarded as having a chi-square distribution with N degrees of freedom (note that the sum of squares of independently distributed standard normal random variables has a chi-square distribution with N degrees of freedom.) The mean of a chi-square distribution with N degrees of freedom is N, and the variance is 2N. Consequently, the asymptotic variance of the unweighted fit mean-square is $Var\left(\frac{\sum_n Z_{ni}^2}{N}\right) = \frac{1}{N^2} \times 2N = \frac{2}{N}$

Figure 8.8 shows a fit map of 20 items administered to 100 students for a simulated data set. It can be seen that the fit mean-square values are generally between 0.8 and 1.2.

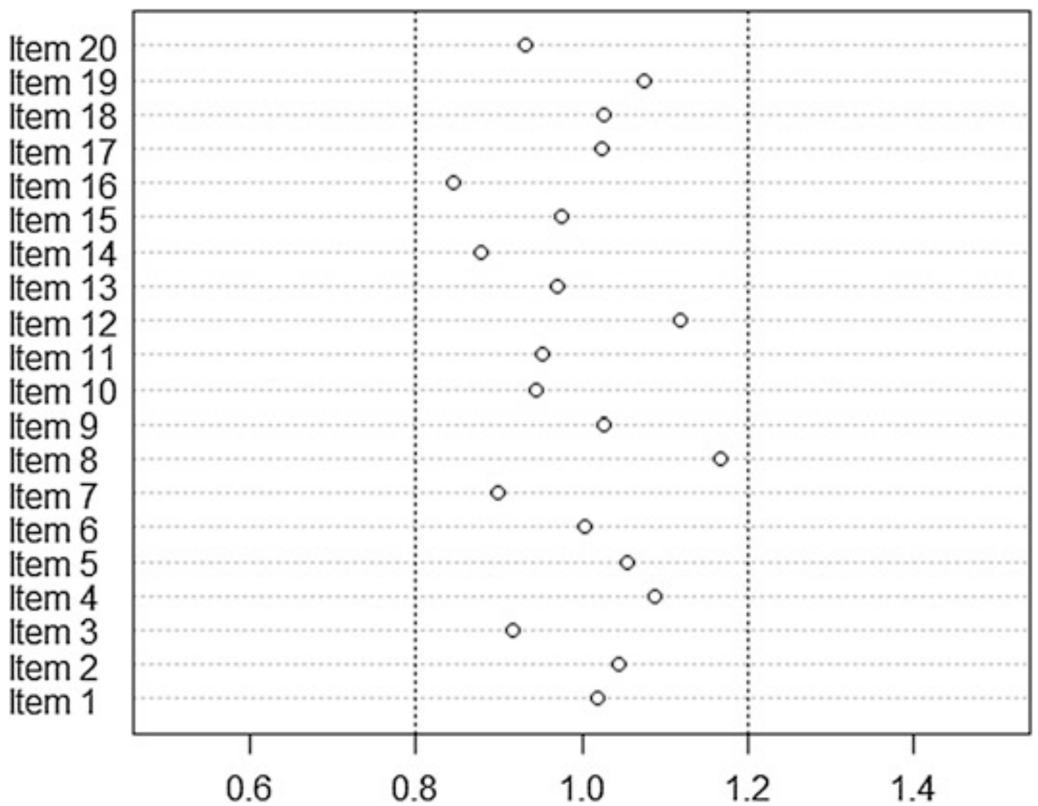


Fig. 8.8 Fit mean squares map when sample size = 100

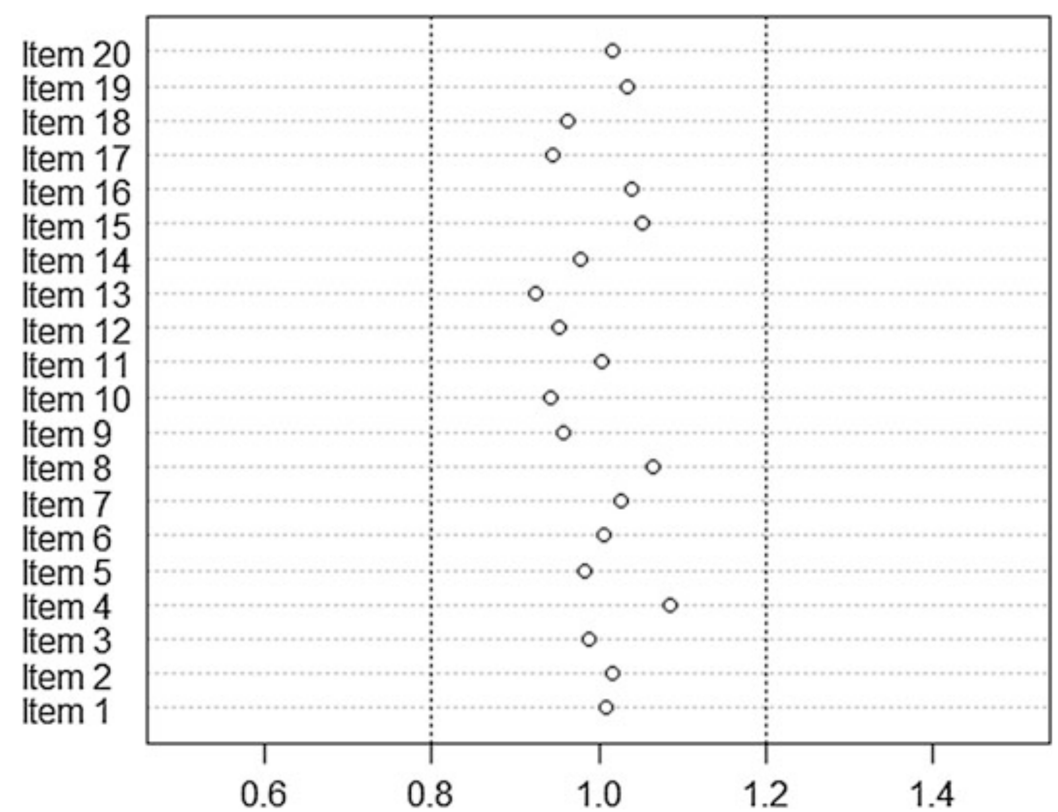


Fig. 8.9 Fit mean squares map when sample size = 500

The Fit t Statistic

The fit t statistic, however, does take sample size into account. Even though it is called a t statistic, the fit t statistic can be regarded as a normal deviate with a mean of zero and a standard deviation of one (i.e., a "z" score), as the sample is typically large enough to use the normal approximation. The fit t statistic is a transformation of the fit mean-square value, taking into account of the mean and variance of the fit mean-square statistic.

Additional Notes
To transform the fit mean-squares to a standardised normal statistic so that one can look up the level of significance easily, the Wilson-Hilferty transformation $t_{unwtt} = \left(Fit_{unwtt}^{1/3} - 1 + 2/(9N)\right)/(2/(9N))^{1/2}$ is often used, where *Fit* is the mean-square value.

An alternative transformation is given in Wright and Masters (1982) that uses a cube root transformation of the fit mean-square and its variance:

$$t_{unwtt} = \left[Fit_{unwt}^{1/3} - 1\right] \times \frac{3}{\sqrt{Var(Fit_{unwtt})}} + \frac{\sqrt{Var(Fit_{unwt})}}{3}$$

Since the fit t statistic can be regarded as a normal deviate, a t value outside the range of -2.0 to 2.0 (or -1.96 to 1.96, to be more precise) can be regarded as an indication of misfit, at the 95% confidence level.

On the surface, our problem regarding the lack of a stable frame of reference for the fit mean-square values seems to have been solved. Unfortunately, this is not the case.

The problem is, in real-life, no item fits the Rasch model perfectly. When items do not fit the Rasch model, any misfit, however small, can be detected when the sample size is large enough. This means that the fit t values will invariably show significance when the sample size is very large. In some sense, the t values are telling the "truth", that there are indeed differences between items, and the items do not tap into the same construct. However, some of these differences between items may be minute from a practical point of view.

	outfitItem	outfitItem_t	infitItem	infitItem_t
M1PTI1	1.0176412	0.1681252	0.9604863	-0.6563752
M1PTI2	0.8600060	-0.8820403	0.9132830	-1.4011322
M1PTI3	0.8747212	-0.5301774	0.9267337	-0.8653759
M1PTI6	0.8348506	-1.6447400	0.9073273	-2.1574356
M1PTI7	0.5103849	-2.8934127	0.7318945	-3.4068671
M1PTI11	0.7181897	-1.5328399	0.8410477	-2.1687957
M1PTI12	1.2043202	2.0302249	1.1212745	2.5738491
M1PTI14	1.4884776	4.9042209	1.2516368	5.6743577
M1PTI17	1.1407154	1.0991002	0.9672130	-0.5461018
M1PTI18	0.9293927	-0.6623501	0.9005372	-2.3313171
M1PTI19	0.6514076	-2.6648368	0.7768166	-3.6387173
M1PTI21	1.8352155	4.5175900	1.3942578	5.2328157
M1PTI22	1.2055754	1.1699414	1.0063787	0.1092650
M1PTI23	0.9120535	-0.6515327	0.9760430	-0.4367405

Fig. 8.10 Fit *t* values for a sample of 500 students

	outfitItem	outfitItem_t	infitItem	infitItem_t
M1PTI1	1.0617523	0.7379272	1.0214576	0.6765263
M1PTI2	0.7068081	-3.8128545	0.8672650	-4.3360205
M1PTI3	0.8524912	-1.3102948	0.9423666	-1.3888542
M1PTI6	0.8184879	-4.0480032	0.8627173	-6.9696935
M1PTI7	0.7469162	-2.9795967	0.8325752	-4.2399000
M1PTI11	0.7920345	-2.4043510	0.9034710	-2.9044195
M1PTI12	1.4255288	8.0500594	1.2021082	8.1901920
M1PTI14	1.3671890	7.4344970	1.2199722	9.3759386
M1PTI17	1.1176319	1.7561847	0.9639025	-1.1589724
M1PTI18	0.9377125	-1.2471293	0.9689911	-1.4644269
M1PTI19	0.5785473	-6.2854536	0.7602613	-7.1088021
M1PTI21	1.7399073	8.8961978	1.3602378	9.9630739
M1PTI22	0.9774774	-0.2471895	0.9211786	-2.1140544
M1PTI23	0.7957022	-3.3884918	0.9086099	-3.7092765

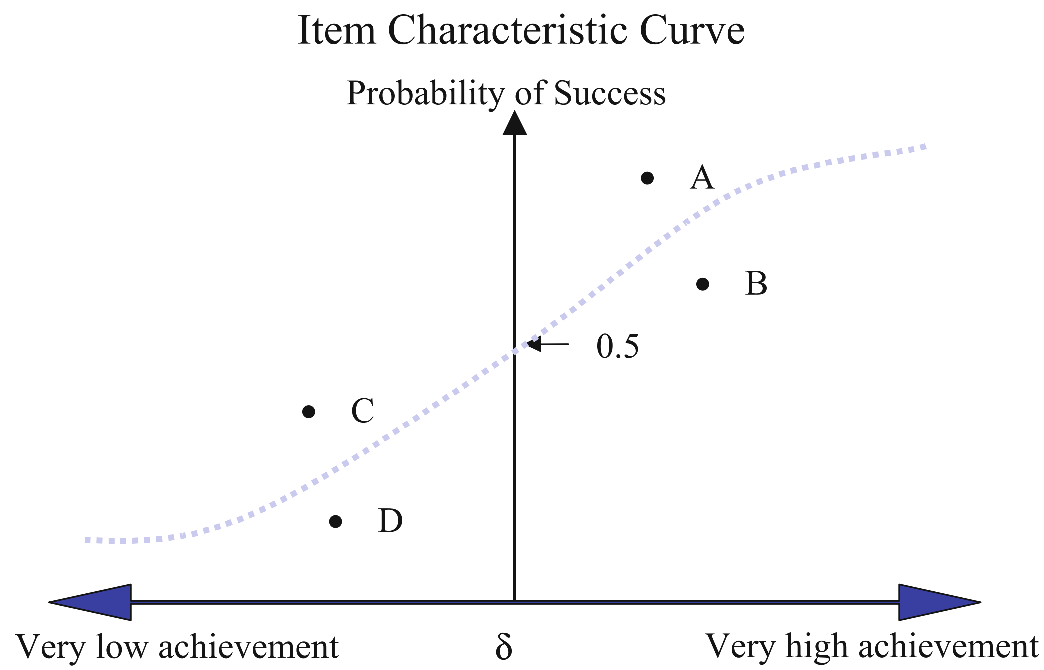
Fig. 8.11 Fit *t* values for a sample of 2000 students

From Figs. 8.10, 8.11, 8.12, it can be seen that as sample size increases, the fit *t* values became progressively far away from zero so that many items showed statistically significant misfit.

	outfitItem	outfitItem_t	infitItem	infitItem_t
M1PTI1	1.0772154	1.59714318	1.0182028	1.012313
M1PTI2	0.7127304	-7.01014880	0.8631072	-8.290753
M1PTI3	0.8548437	-2.31343680	0.9401009	-2.554359
M1PTI6	0.8511479	-5.89885727	0.8890838	-9.915885
M1PTI7	0.7157490	-5.61754107	0.8225674	-7.870286
M1PTI11	0.7980984	-4.14518615	0.8942556	-5.626692
M1PTI12	1.3649289	11.26997571	1.1814339	12.419531
M1PTI14	1.3721131	12.92338691	1.2133264	16.153633
M1PTI17	1.0984786	2.55868638	0.9547443	-2.676895
M1PTI18	1.0018899	0.07510419	0.9785618	-1.767476
M1PTI19	0.6063701	-9.90033565	0.7778577	-11.743530
M1PTI21	1.7837242	15.58244351	1.3510149	17.137759
M1PTI22	0.9783801	-0.41760535	0.9138654	-4.053861
M1PTI23	0.8246742	-5.15048395	0.9192772	-5.788562

Fig. 8.12 Fit *t* values for a sample of 6371 students

Fig. 8.14 Expected ICC and observed ICC points



Additional Notes

Figure 8.14 shows the theoretical, or expected, item characteristic curve for an item, with four points, A, B, C, and D denoting four regions where the observed ICC may fall. Point A denotes the region above the theoretical ICC, and to the right of the vertical line where $\theta = \delta$, the ability at which there is a 50% chance of obtaining the correct answer. Point B denotes the region below the theoretical ICC and to the right of the vertical line $\theta = \delta$. Point C denotes the region above the theoretical ICC but to the left of the $\theta = \delta$ line. Point D denotes the region below the theoretical ICC and to the left of the $\theta = \delta$ line. It can be shown mathematically that the contribution of observed points in the A and D region to the outfit mean-square, $z_{ni}^2 = \frac{(x_{ni} - E(X_{ni}))^2}{(Var(X_{ni}))}$, has an expectation less than one, while the expectation of z_{ni}^2 for points in the C and B regions is greater than one. It is clear then the fit mean-square value provides a test of whether the "slope" of the observed ICC is the same as the theoretical one. Given that the theoretical one can be regarded as an "average" of all items, the fit mean-square value tests whether the observed ICC for this item is the same as the slopes of the other items.

When residual-based fit statistics show that items fit the Rasch model, this is not sufficient to conclude that you have the best test instrument.

Benchmarks

- Critérios para Infit, Outfit:
  - >2,0: Degrada o sistema de mensuração,
  - 1,5 a 2,0: Não produtivo para medida
  - 0,5 a 1,5: Produtivo
  - <0,5 menor produtivo mas menos preocupante. Pode produzir artificialmente altas precisões
-

Reasonable Item Mean-square Ranges for INFIT and OUTFIT

Type of Test	Range
MCQ (High stakes)	0.8 - 1.2
MCQ (Run of the mill)	0.7 - 1.3
Rating scale (survey)	0.6 - 1.4
Clinical observation	0.5 - 1.7
Judged (agreement encouraged)	0.4 - 1.2

Análise de person misfit

Person Responses: Easy -- Items -- Hard	Diagnosis Pattern	OUTFIT Mean-square	INFIT Mean-square	Point-Measure Correlation	S.E. Inflator
111¦0110110100¦000	Modelled/Ideal	1.0	1.1	0.62	1.0
111¦1111100000¦000	Guttman/Deterministic	**0.3**	**0.5**	**0.87**	1.0
000¦0000011111¦111	Miscode	**12.6**	4.3	**-0.87**	3.5
011¦1111110000¦000	Carelessness Sleeping Slipping	**3.8**	1.0	0.65	1.9
111¦1111000000¦001	Lucky Guessing	**3.8**	1.0	0.65	1.9
101¦0101010101¦010	Response set/Miskey	**4.0**	**2.3**	**0.11**	2.0
111¦1000011110¦000	Special knowledge	0.9	**1.3**	0.43	1.1
111¦1010110010¦000	Imputed outliers *	**0.6**	1.0	0.62	>1.0*
111¦0101010101¦000	Low discrimination	**1.5**	**1.6**	0.46	1.3
111¦1110101000¦000	High discrimination	**0.5**	**0.7**	0.79	1.0
111¦1111010000¦000	Very high discrimination	**0.3**	**0.5**	**0.84**	1.0
Right¦Transition¦Wrong					
high - low - high	OUTFIT sensitive to outlying observations	>>1.0 unexpected outliers	>>1.0 disturbed pattern		
low - high - low	INFIT sensitive to pattern of inlying observations	<<1.0 overly predictable outliers	<<1.0 Guttman pattern		

* as when a tailored test is filled out by imputing all "right" response to easier items and all "wrong" to harder items. Increase S.E. based on number of observed response.

The exact details of these computations have been lost, but the items appear to be uniformly distributed about 0.4 logits apart.

Ajuste dos itens é relativo

- Os índices avaliam a discrepância entre a ICC esperada do item a ICC observada
- A ICC esperada terá a discriminação média dos itens presentes no teste
- Se retirarmos um item essa discriminação média mudará
- Procedimento:
  - deletar itens com valores muito maiores do que 1 (1.4). Manter itens com valores <1.
  - Recalcular os índices nos itens restantes

Para aprofundar …

- https://www.rasch.org/rmt/rmt82a.htm
- https://www.winsteps.com/winman/misfitdiagnosis.htm
- https://www.rasch.org/rmt/rmt83b.htm

- https://www.edmeasurementsurveys.com/TAM/Tutorials/index.htm

Exercício: calcule os índices de ajuste para os dados do exercício 2