Open Journal of Statistics, 2017, 7, 780-785
http://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

The Scaling Constant D in Item Response Theory

Gregory Camilli

Graduate School of Education, Rutgers University, New Brunswick, NJ, USA
Email: greg.camilli@gse.rutgers.edu

How to cite this paper: Camilli, G. (2017) The Scaling Constant D in Item Response Theory. *Open Journal of Statistics*, **7**, 780-785.
https://doi.org/10.4236/ojs.2017.75055

Received: August 28, 2017
Accepted: October 10, 2017
Published: October 13, 2017

Abstract

In item response theory (IRT), the scaling constant $D = 1.7$ is used to scale a discrimination coefficient a estimated with the logistic model to the normal metric. Empirical verification is provided that Savalei's [1] proposed a scaling constant of $D = 1.749$ based on Kullback-Leibler divergence appears to give the best empirical approximation. However, the understanding of this issue as one of the accuracy of the approximation is incorrect for two reasons. First, scaling does not affect the fit of the logistic model to the data. Second, the best scaling constant to the normal metric varies with item difficulty, and the constant $D = 1.749$ is best thought of as the average of scaling transformations across items. The reason why the traditional scaling with $D = 1.7$ is used is simply because it preserves historical interpretation of the metric of item discrimination parameters.

Keywords

Item Response Theoru, IRT, Scaling Constant, D

1. Introduction

Two common families of models used in item response theory (IRT) are the normal and logistic distribution functions. These two models are both used extensively. The logistic model is used more in ongoing assessment programs, while the normal model tends to be used more in research studies. It is thought, with some justification, that these two models obtain coefficients estimates that are practically indistinguishable after a simple multiplicative scaling. Below, this claim is explained in detail and investigated more thoroughly.

2. Item Response Theory (IRT)

Assume a set of test items $j = 1, \cdots, J$ for subjects $i = 1, \cdots, N$, where the items are dichotomously scored: correct responses are scored $Y_{ij}=1$, and incorrect responses $Y_{ij}=0$. Also, assume there is a single latent variable θ_i, which is known as examinee ability or proficiency, that accounts for an examinee's observed item responses. While θ can be multidimensional, only the unidimensional case is considered here.

With the two parameter normal ogive function (2PN), a correct response on item j presented to examinee i is modeled by

$$\begin{aligned} P_{ij}\left(Y_{ij}=1 \mid \eta_{ij}\right) &= G\left(\eta_{ij}\right) \\ \eta_{ij} &= a_j\left(\theta_i-b_j\right), \end{aligned} \tag{1}$$

where G is the cumulative normal distribution function defined as

$$G(\eta)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\eta}\exp\left(-t^2/2\right)\mathrm{d}t. \tag{2}$$

In this two-parameter IRT model, a_j is a guessing parameter, and b_j is an item difficulty parameter. The person parameter θ_i is defined above. In the two parameter logistic model (2PL), correct responses are modeled by

$$P_{ij}\left(Y_{ij}=1 \mid \eta_{ij}\right)=F\left(\eta_{ij}\right)=\frac{\exp\left(\eta_{ij}\right)}{1+\exp\left(\eta_{ij}\right)}, \tag{3}$$

where again $\eta_{ij}=a_j\left(\theta_i-b_j\right)$. Note that the parameterizations of both the 2PN and 2PL IRT models are in terms of a_j, b_j, and θ_i, and the interpretations of model coefficients as item discrimination, item difficulty, and person ability are identical. In short, both the 2PN and 2PL models for dichotomous items provide an estimate of the probability that an examinee will answer correctly, and both models are based on the same item and person parameters.

3. Scaling for Equivalence

In this context, it has been known for some time that the logistic distribution is very similar to the normal distribution because conditional on θ, 1) both distributions are determined by a location and a scale parameter, and 2) both distributions are bell shaped. At the same time, the logistic distribution has heavier tails than the normal distribution. The question then arises of how well $G(\eta)$ can be approximated with a scaled version of the logistic distribution $F\left(D\eta\right)=F\left(Da^*\left(\theta-b\right)\right)$, where D is known as the scaling constant. This produces the scaled discrimination $a^*=a/D$, which is purportedly interpretable in the normal metric. Note also that $D=a/a^*$, so that D can be conceptualized as the ratio of the logistic-metric discrimination to the normal-metric discrimination.

Several informal suggestions have included D = 1.814 [2] and D = 1.700 [3]. Haley [4] provided the constant predominately used today in IRT, D = 1.702 (usually rounded to 1.7), which is obtained by minimizing the maximum absolute difference over η[1]. Savalei [1] proposed a scaling constant based on Kullback-

[1] Camilli [5] also provided an illustration of how the minimax estimator is estimated; however, the formula for the normal ogive was incorrectly written Equation (3). A note by Camilli [6] did not clarify this issue. However, Equation (4) in Camilli [5] is correct, so the error did not affect the ultimate outcome.

Leibler (KL) information by minimizing the function

$$K(g,f)=\int \ln\left[g(x)/f(Dx)\right]g(x)\,\mathrm{d}x, \tag{4}$$

where g and f are the normal and logistic density functions, and integration is over $\mathbb{R}^1$. This results in the scaling constant D = 1.749. More recently, Pingel [7] considered a number of measures of accuracy of scaling the logistic to the normal: 1) a minimax estimator labeled as $\|\cdot\|_\infty$, and 2) the root mean square difference $\|\cdot\|_2$. He showed that relatively similar values of D are obtained with either $\|F-G\|$ or $\|f-g\|$, where f and g are the corresponding density functions. For example, $\|f-g\|_\infty$ obtains D = 1.618[2]. Minima of the expected value of other distance functions provide similar results. For example, minimizing the average absolute difference leads to D = 1.701. With this panoply of scaling options, the question arises "Which one is best?" The scaling constant D = 1.749 seems to work the best, but this answer is somewhat misleading as shown below.

4. Simulation Study

While it is theoretically established than the use of a scaling constant results in a close match between F and G, the current paper provides empirical verification, and a new result. For this purpose, a simulation study was designed to compare estimates obtained with a logistic model (2PN) for data generated with a normal model (2PL). In theory, the estimated 2PL parameters scaled with D should be very close to the known 2PN parameters. In addition, the ability estimates of θ obtained with the 2PL model should closely match those of the known 2PN model. The steps in the simulation were as follows:

1) For 100 items, generate item parameters with discrimination with a ~ lognormal (0.25, 0.25) and intercepts b ~ normal (0, 0.85). These generating distribution give adequate approximations to observed empirical distributions of item parameter estimates.

2) For 100,000 persons, generate ability parameters with θ ~ normal (0, 1). A large sample size is used to minimize the effects of estimation errors on the estimation of a scaling coefficient.

3) Estimate 2PL model parameters (a, b, θ) using the EM algorithm with 61 quadrature points, and

a) Compare $\hat{a}$ to the normal generating value of a for each item. If the scaling constant is accurate, it should be the case that $D \approx \hat{a}/a^*$ for all items.

b) Compare $\hat{b}$ to the normal generating value of b for each item

c) Compare $\hat{\theta}$ to the normal generating value of θ for each person

The idea here is to obtain the ratio of the estimate of the 2PL a to its normal generating parameter. This ratio is the empirical scaling value D. This process should reveal which of the proposed scaling values is most accurate. Note that the b and θ parameters do not need to be scaled; their metric is typically defined by the identification restriction θ ~ normal (0, 1), which is employed in many

[2] Pingel [7] also showed that the approximation error can be decreased by an order of magnitude when F is approximated by a t distribution.

IRT software packages.

For the purpose of this simulation, the R software was used to randomly generate item responses Y_{ij} by (a) computing $G(\eta_{ij})$ in Equation (1) from simulated parameters obtained from steps 1 and 2 above, (b) drawing a uniform random variate U[0, 1], and (c) setting $Y_{ij} = 1$ if $U < G(\eta_{ij})$ and $Y_{ij} = 0$ otherwise. Parameter estimates for the 2PL model were obtained using flexMIRT [8]. Person parameters for the 2PL model were estimated with the EAP method, which is the average of the posterior distribution of θ.

5. Results

To compare discriminations, the ratio was taken of the logistic estimate $\hat{a}$ to its normal generating parameter a^* for each item. In theory, this ratio should be close to D for all items. Across items, the median ratio was 1.751 and the mean ratio was 1.756. This is very close to Savalei's [1] ratio of D = 1.749. However, the minimum ratio across items was 1.666, and the maximum 2.031. The unexpected finding here is that the ratios vary across items, and this variation is not due to sampling error: the standard error of a is about 0.01 - 0.03 units across items for the current sample size. The ratio tends to increase slightly as the 2PN generating value of a increases, but also increases more noticeably for tail values of b as shown in **Figure 1**. In the IRT context, it is therefore more appropriate to think of D as an average scaling value rather than a scaling "constant."

The b parameter estimates on average differed from the 2PN generating values by 0.002 on a unit normal scale, with a minimum difference of −0.021 and a maximum difference of 0.062. This indicates the 2PL IRT model provides estimates that are empirically very similar to those of a 2PN model. To study the

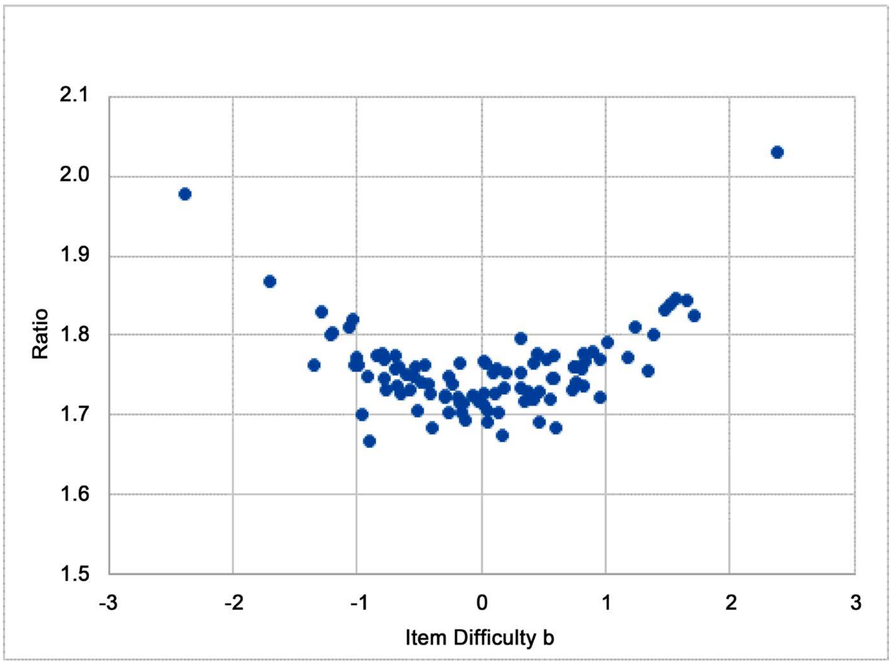

Figure 1. Plot of $\hat{a}/a$ against item difficulty b for simulation study.

ability parameter θ, estimated values were regressed on 2PN generating values. This resulted in an intercept of −0.003 and a slope of 0.981. A linear correlation of $r = 0.989$ was obtained. The plot (not shown) provided no evidence of nonlinearity.

The empirical conclusion based on these results is nearly the same as the theoretical expectation: there is little difference between the 2PN and the scaled 2PL estimates of item parameters for items having nonextreme values of *b*. In a simulation not shown, this result was also verified for IRT models for partial credit data, using the logistic model [9] to approximate the normal model [10]. However, the scaling becomes less accurate as |b| increases.

6. Discussion

So far, this paper has omitted consideration of the most fundamental question: *Why scale*? The term "accuracy" implies the normal metric is the correct one for obtaining IRT parameters, but this fundamental assumption is rarely recognized let alone tested. The logistic function due to its heavier tails may even be preferable in situations involving noisy data. Ironically, one could even argue that D^{-1} should be used to scale 2PN item parameters to the logistic metric. In short, there is no necessary relationship between scaling and the accuracy of the IRT model.

The best choice of scaling in logistic IRT models would be not to scale at all—an approach taken in some current IRT software packages such as flexMIRT. The sole rationale for scaling with D is to establish historical continuity in interpreting the magnitude of item parameter estimates. More than two generations have passed since Alan Birnbaum's suggested use of scaling [11]. During this time, the psychometric community and other interested parties have grown accustomed to discrimination parameters scaled with $D = 1.7$ through thousands of technical reports and journal articles, and years of day-to-day work activities in large-scale assessment programs. However, it should be recognized that scaling does not reproduce the normal metric in all cases, the bias is systematic, and scaling in increasingly popular multidimensional logistic models is awkward. Figure 1 also suggests that the use of $D = 1.749$ leads to an unacceptable degree of bias in the approximation if the difficulty of a test is not matched to examinee ability, *i.e.*, if item difficulty parameters are not centered over average ability. The time for traditional scaling with $D = 1.7$ may be drawing to a close.

References

[1] Savalei, V. (2006) Logistic Approximation to the Normal: The KL Rationale. *Psychometrika*, **71**, 763-767. https://doi.org/10.1007/s11336-004-1237-y

[2] Cox, D.R. (1970) The Analysis of Binary Data. Methuen, London.

[3] Johnson, N.J. and Kotz, S. (1970) Continuous Univariate Distributions-2. Houghton Mifflin, Boston.

[4] Haley, D.C. (1952) Estimation of the Dosage Mortality Relationship When the Dose

Is Subject to Error Technical Report No. 15 (Office of Naval Research Contract No. 25140, NR-342-022). Applied Mathematics and Statistics Laboratory, Stanford University.

[5] Camilli, G. (1994) Origin of the Scaling Constant in Item Response Theory. *Journal of Educational and Behavioral Statistics*, **19**, 293-295. https://doi.org/10.2307/1165298

[6] Camilli, G. (1995) Correction. *Journal of Educational and Behavioral Statistics*, **20**, np.

[7] Pingel, R. (2014) Some Approximations of the Logistic Distribution with Application to the Covariance Matrix of Logistic Regression. *Statistics and Probability Letters*, **85**, 63-68. https://doi.org/10.1016/j.spl.2013.11.007

[8] Houts, C.R. and Cai, L. (2013) flexMIRT Users Manual Version 2.0: Flexible Multilevel Item Factor Analysis and Test Scoring. Vector Psychometric Group, Seattle.

[9] Muraki, E. (1992) A Generalized Partial Credit Model: Application of an EM Algorithm. *Applied Psychological Measurement*, **16**, 159-176. https://doi.org/10.1177/014662169201600206

[10] Samejima, F. (1969) Estimation of Latent Ability Using a Response Pattern of Graded Scores. (Psychometrika Monograph, No. 17). Psychometric Society, Richmond. http://www.psychometrika.org/journal/online/MN17.pdf

[11] Birnbaum, A. (1968) Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In: Lord, F.M. and Novick, M.R., Eds., *Statistical Theories of Mental Test Scores*, Addison-Wesley, Reading, 397-479.