ED 268 148	TM 860 138
AUTHOR	Holland, Paul W.; Thayer, Dorothy T.
TITLE	An Alternate Definition of the ETS Delta Scale of Item Difficulty. Program Statistics Research.
INSTITUTION	Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO	ETS-RR-85-43; ETS-TR-85-64
PUB DATE	Oct 85
NOTE	17p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Difficulty Level; Error Patterns; Estimation (Mathematics); *Item Analysis; *Mathematical Models; Predictive Validity; *Scaling; *Test Items
IDENTIFIERS	*Delta Scale; *Logit Analysis

ABSTRACT

An alternative definition has been developed of the delta scale of item difficulty used at Educational Testing Service. The traditional delta scale uses an inverse normal transformation based on normal ogive models developed years ago. However, no use is made of this fact in typical uses of item deltas. It is simply one way to make the probability scale of item difficulty a more useful set of units that are not compressed near 0 or 1. The alternative scale uses a different function to achieve the same result. Both a logistic definition and a normal definition can be calculated. For item difficulty between .10 and .90, the difference between the standard definition of the delta scale and one based on the logistic distribution is negligible. The logistic definition scales very easy items as easier than the normal definition. This change in the delta scale, as compared with the traditional delta, would have little effect on the values of the statistics used. However, it offers the advantage of the use of logits. Also, differences in item deltas (e.g., for a comparison of two subpopulations' performance on an item) can be interpreted in terms of odds-ratios of the corresponding difficulty values. (GDC)

AN [illegible]
ETS [illegible]

Paul W. Holland [illegible]

TECHNICAL REPORT NO. [illegible]

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

AN ALTERNATE DEFINITION OF THE ETS DELTA SCALE OF ITEM DIFFICULTY

by

Paul W. Holland

and

Dorothy T. Thayer

Program Statistics Research
Technical Report No. 85-64

Research Report No. 85-43

October 1985

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

TABLE OF CONTENTS

Page

1. The Standard Definition of an Item's "Delta" 1

2. An Alternative Definition of $\Delta(p)$ 2

3. Comparative Values of Δ_N and Δ_L 3

4. Why Another Definition of the Delta Scale? 4

References 8

ABSTRACT

This note develops an alternative definition of the "delta scale" of item difficulty that is used at ETS. A comparison is given with the traditional definition of the delta scale that is based on the normal distribution. Some advantages of the alternative scale are mentioned.

1. THE STANDARD DEFINITION OF AN ITEM'S "DELTA"

Suppose p denotes the proportion of examinees in a given population of examinees who answer a particular item correctly. The value, p, is a population parameter that measures the easiness of the item, i.e., higher values of p denote easier items. At ETS, the difficulty of an item is measured by a transformation of p to the "delta scale." The transformation of p into Δ is given by the equation

$$\Delta(p) = 13 - 4Z_p$$

where Z_p is the usual "z-value" that corresponds to p. That is, the probability that a normal deviate is smaller than Z_p is p. Δ(p) may also be expressed as

$$\Delta(p) = 13 - 4\Phi^{-1}(p) \qquad (1)$$

where $\Phi^{-1}(p)$ denotes the inverse function of the normal cummulative distribution function, i.e.,

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du. \qquad (2)$$

The value of Δ(p) is a population measure of the difficulty of an item because higher values of Δ(p) denote more difficult items. The location and scale values of 13 and 4 in (1) are arbitrary, but they ensure that typical delta values range from about 5 to about 21. This avoids negative values and may have other practical advantages.

The use of the inverse normal transformation $\Phi^{-1}(p)$ in (1) is based on "normal ogive" types of models for item responses that were developed years ago. However, no use of this fact is made in typical uses of "item deltas." We regard the use of $\Phi^{-1}(p)$ as simply one way to stretch out the probability scale of p into a more useful set of units that are not seriously compressed near p=0 or p=1. The alternative scale given in section 2 uses a different function to alter the p-scale in a similar way.

Estimates of $\Delta(p)$ are used in practice. These are based on samples from given populations of examinees. Let $\hat{p}$ denote a sample proportion of examinees (out of n) who give the correct answer on the item in question. The sample delta value is

$$\hat{\Delta} = \Delta(\hat{p}) \tag{3}$$

where $\Delta(p)$ is the function defined in (1).

The standard error of $\hat{\Delta}$ can be obtained using the δ-method (see Bishop, Fienberg, and Holland, 1975). It is given by the asymptotic variance formula,

$$\mathrm{Var}(\hat{\Delta}) = 4^2 \, \frac{2\pi \; p(1-p)}{n} \exp((\Phi^{-1}(p))^2), \tag{4}$$

so that the standard error of $\hat{\Delta}$ can be estimated by

$$\mathrm{s.e.}(\hat{\Delta}) = 4\sqrt{\frac{2\pi \; \hat{p}(1-\hat{p})}{n}} \exp(\tfrac{1}{2}(\Phi^{-1}(\hat{p}))^2). \tag{5}$$

2. AN ALTERNATIVE DEFINITION OF $\Delta(p)$

Lord and Novick (1968, page 399) report that the normal cumulative $\Phi(x)$ and a suitably scaled logistic cumulative differ by no more than .01 for all x. For example, if

$$\Psi(x) = e^x/(1+e^x), \tag{6}$$

then

$$|\Phi(x) - \Psi(1.7x)| \leq .01 \quad \text{all } x.$$

Hence, we can approximate $\Phi(x)$ by the scaled logistic $\Psi(1.7x)$. This suggests approximating $\Phi^{-1}(p)$ by

$$\frac{1}{1.7}\,\Psi^{-1}(p), \tag{7}$$

where

$$\Psi^{-1}(p) = \ln(\frac{p}{1-p}), \qquad (8)$$

and ln(u) denote the natural log of u. Hence, the formula for $\Lambda(p)$ in (1) can be approximated by

$$13 - \frac{4}{1.7}\ln(\frac{p}{1-p}). \qquad (9)$$

We may create an alternative definition of $\Delta(p)$ by using (9) as its definition rather than (1). Some reasons for doing this will be mentioned in section 4.

We will denote by $\Delta_L(p)$ the <u>logistic</u> definition of the Δ-scale for p, i.e.,

$$\Delta_L(p) = 13 - \frac{4}{1.7}\ln(\frac{p}{1-p}). \qquad (10)$$

or

$$\Delta_L(p) \doteq 13 - 2.35\ln(\frac{p}{1-p}). \qquad (11)$$

and we will denote the <u>normal</u> definition of Δ by $\Delta_N(p)$, i.e.

$$\Delta_N(p) = 13 - 4\Phi^{-1}(p). \qquad (12)$$

3. COMPARATIVE VALUES OF Δ_N AND Δ_L

The approximation of $\Phi(x)$ by $\Psi(1.7x)$ is quite good for all values of x. However, when we go to the inverses of these two functions we have no guarantee of a similarly good approximation. This needs to be examined directly. Table 1 and Figure 1 give values of $\Delta_N(p) - \Delta_L(p)$ for values of p = .01, .02, ..., .99. From this we see that for p between .09 and .91 the difference between Δ_N and Δ_L never exceeds .11. As p approaches 0 and 1 the difference grows more rapidly. The difference exceeds 0.50 for $p \geq .97$ or $p \leq .03$, and at p=.99 or p=.01 it is 1.51.

In rough summary then, for $.10<p<.90$ the difference between the standard definition of the delta scale and one based on the logistic distribution is negligible for practical purposes. For values of p in excess of .90 the logistic definition of Δ always yields smaller values of Δ than does the normal definition. For values of p smaller than .10 the logistic definition of Δ always yield values for Δ that are greater than the normal definition. In many practical situations (e.g., multiple choice tests) values of p less than .1 are rarely encountered. In these situations the only real difference that one might notice between the two definitions of Δ is that the logistic definition will scale very easy items (i.e., $p\geq.95$) as easier (i.e., lower Δ values) than will the normal definition of Δ.

4. WHY ANOTHER DEFINITION OF THE DELTA SCALE?

Our purpose is not to argue strongly for a change in the delta scale that has been used for a long time at ETS and which is familiar to those who need to use it in test construction and analysis. Rather, we wisn to show that if such a change were made, it would have little effect on the values of the statistics that are used buc would have some advantages that may prove useful. At the very least, our analysis shows that useful results that apply to the logistic definition of the delta scale may be translated into results that almost hold for the normal definition of this scale.

Possibly the most important advantage of the $\Delta_L(p)$ over $\Delta_N(p)$ is that $\Delta_L(p)$ involves "logits". The logit of p is $\log(p/(1-p))$. This is a very well studied quantity in the statistical (especially biostatistical) literature. For example, it is known that a good estimator of $\Delta_L(p)$ is not the obvious $\Delta_L(\hat{p})$ but the smoother

$$\hat{\Delta}_L = 13 - \frac{4}{1.7} \ln\left(\frac{X+\frac{1}{2}}{n-X+\frac{1}{2}}\right), \qquad (13)$$

where $\hat{p} = X/n$ and X is the sample <u>number correct</u> ($\hat{p}$ is the sample <u>proportion correct</u>). The estimator in (13) is unbiased to order $O(n^{-2})$, unlike the more obvious estimate, $\Delta_L(\hat{p})$. The bias of $\Delta_L(\hat{p})$ is $O(n^{-1})$ so that while $\Delta_L(\hat{p})$ and $\hat{\Delta}_L$ from (13) both converge to the true population value $\Delta_L(p)$ as $n\to\infty$, $\Delta_L(\hat{p})$ does so at a slower rate that does $\hat{\Delta}_L$.

Formula (13) is derived from the Haldane-Anscombe estimator of the logit of p -- see Bedrick (1984).

In addition, the standard error of $\hat{\Delta}_L$ can be estimated well using the formula

$$\text{s.e.}(\hat{\Delta}_L) = \frac{4}{1.7}\sqrt{\frac{X+.1}{(X+.3)^2} + \frac{n-X+.1}{(n-X+.3)^2}} . \qquad (14)$$

The formula in (14) is derived from the work of Bedrick (1984) on estimators of the standard deviation of the Haldane-Anscombe estimate of the logit of p. Bedrick shows that the square of (14) provides an unbiased estimate of the variance of $\hat{\Delta}_L$ to order $O(n^{-3})$. Hence, (13) and (14) provide a rather complete package for estimating $\Delta_L(p)$ that has good statistical properties, even in relatively small samples. No such claim can be made for the corresponding formulas (12) and (5) to estimate $\Delta_N(p)$. They are only justified in large samples.

A second virtue of the logistic definition of Δ is that <u>differences</u> in item deltas -- say, in a comparison of the performance of two subpopulations of examinees on the same item -- can be interpreted in terms of odds-ratios of the corresponding p values. For example, suppose p_1 is the proportion in group 1 who got the item correct while p_2 is the corresponding proportion in group 2.

If we form the difference,

$$\Delta_L(p_1) - \Delta_L(p_2),$$

a bit of algebra reveals it to equal

$$-\frac{4}{1.7}\ln\left(\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2}\right) \tag{15}$$

which, except for the factor -4/1.7, is the log of the odds-ratio

$$\frac{p_1}{1-p_1} \Big/ \frac{p_2}{1-p_2} . \tag{16}$$

The odds-ratio is also the cross-product ratio for the following 2x2 table,

	Right	Wrong	Total
Group 1	p_1	$1-p_1$	1
Group 2	p_2	$1-p_2$	1

(17)

i.e., the cross-product ratio is

$$\frac{p_1(1-p_2)}{p_2(1-p_1)} . \tag{18}$$

The cross-product ratio and its natural log are widely regarded as useful, margin-free, measures of associations in 2x2 tables. By margin-free we mean that if the marginal distributions of the 2x2 table in (17) are modified by multiplying each row and column by factors then the cross-product ratio is unchanged. The margin-free nature of the cross-product ratio is quite important for test development use of the Δ-scale since it insures that changes in the overall correct answer rate of an item for a population will have a minimal effect on the comparison of item deltas for subgroups within the population. For example, differences in deltas found in one test administration will tend to hold up in other test administrations. Hence, the use of $\Delta_L(p)$ rather than $\Delta_N(p)$ brings the <u>comparison</u> of item difficulty indices into line with a wellestabl:shed statistical theory of dependence in 2x2 tables, e.g., Bishop, Fienberg, and Holland (1975, chapter 11).

REFERENCES

Bedrick, E. (1984) Estimating the variance of empirical logits and contrasts in empirical log probabilities. Biometrics, 40, 805-809.

Bishop, Y., Fienberg, S., and Holland,P. (1975) Discrete Multivariate Analysis: Theory and Practice. Cambridge, MA: MIT Press.

Lord, F. and Novick, M. (1968) Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

FIGURE 1. PLOT OF $\Delta_N(p) - \Delta_L(p)$ VERSUS p.

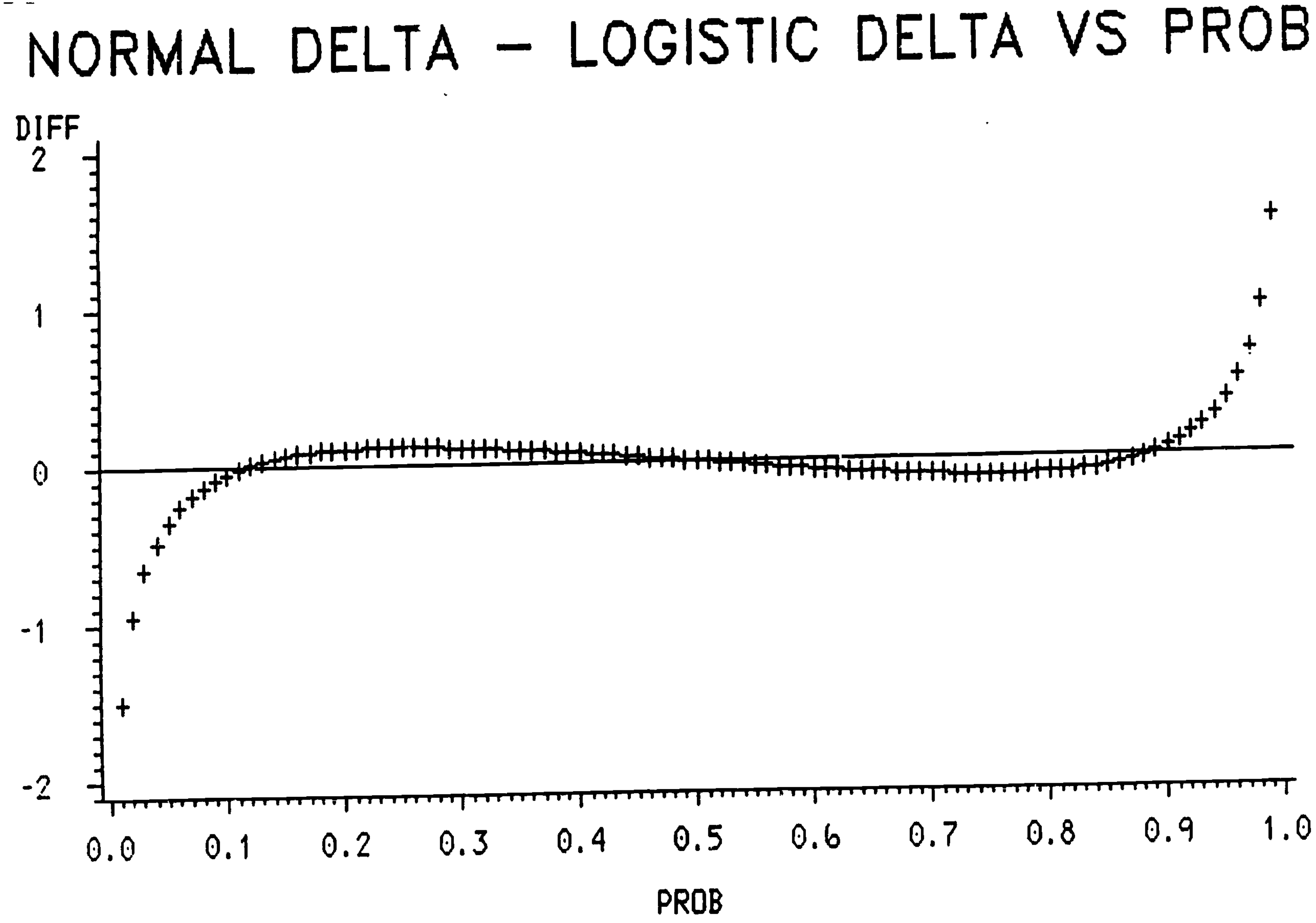

TABLE 1. VALUES OF p and $\Delta_N(p) - \Delta_L(p)$ for p=.01 (.01) .99

p	$\Delta_N(p)-\Delta_L(p)$	p	$\Delta_N(p)-\Delta_L(p)$
.01	-1.51	.51	-.01
.02	-.94	.52	-.01
.03	-.66	.53	-.02
.04	-.48	.54	-.02
.05	-.35	.55	-.03
.06	-.26	.56	-.04
.07	-.18	.57	-.04
.08	-.13	.58	-.05
.09	-.08	.59	-.05
.10	-.04	.60	-.06
.11	-.01	.61	-.06
.12	.01	.62	-.07
.13	.03	.63	-.08
.14	.05	.64	-.08
.15	.06	.65	-.08
.16	.08	.66	-.09
.17	.09	.67	-.09
.18	.09	.68	-.10
.19	.10	.69	-.10
.20	.10	.70	-.10
.21	.11	.71	-.11
.22	.11	.72	-.11
.23	.11	.73	-.11
.24	.11	.74	-.11
.25	.11	.75	-.11
.26	.11	.76	-.11
.27	.11	.77	-.11
.28	.11	.78	-.11
.29	.11	.79	-.11
.30	.10	.80	-.10
.31	.10	.81	-.10
.32	.10	.82	-.09
.33	.09	.83	-.09
.34	.09	.84	-.08
.35	.08	.85	-.06
.36	.08	.86	-.05
.37	.08	.87	-.03
.38	.07	.88	-.01
.39	.06	.89	.01
.40	.06	.90	.04
.41	.05	.91	.08
.42	.05	.92	.13
.43	.04	.93	.18
.44	.04	.94	.26
.45	.03	.95	.35
.46	.02	.96	.48
.47	.02	.97	.66
.48	.01	.98	.94
.49	.01	.99	1.51
.50	.00		