ED 305 385 TM 012 915

AUTHOR Wothke, Werner; Zimowski, Michele
TITLE Item Analysis of the Paper Folding Test (Wks. 622). Technical Report No. 1988-6.
INSTITUTION Johnson O'Connor Research Foundation, Chicago, IL. Human Engineering Lab.
PUB DATE Sep 88
NOTE 21p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; Factor Analysis; *Item Analysis; *Latent Trait Theory; *Spatial Ability; *Test Validity; *Visualization; *Work Sample Tests
IDENTIFIERS Johnson O Connor Aptitude Tests; *Paper Folding; Structural Visualization

ABSTRACT

Large-sample item response data for the 10-item Paper Folding worksample 622 (N=2,749) and for five new experimental paper folding items (N=2,514) are analyzed with the logistic item response model and with full-information item factor analysis. The main results of the unidimensional analysis are that: (1) item discrimination is heterogeneous, so that a Rasch model cannot be ascertained for the Paper Folding worksample; (2) item difficulty does not increase with presentation order; and (3) there are not enough items in the midrange of the scale. Item factor analysis identifies an additional minor factor attributable to during-the-test learning of specific solutions for items with similar stems. Based on these results, a modified 11-item version of the Paper Folding worksample is proposed. (Six graphs and five tables present the data.) (Author)

ITEM ANALYSIS OF THE PAPER FOLDING TEST (Wks. 622)

Werner Wothke
and
Michele Zimowski

JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.
HUMAN ENGINEERING LABORATORY, INC.

Technical Report 1988-6
September 1988

ITEM ANALYSIS OF THE PAPER FOLDING TEST (Wks. 622)

Werner Wothke and Michele Zimowski

Abstract

Large-sample item response data for the 10-item Paper Folding worksample 622 ($N = 2{,}749$) and for 5 new experimental paper folding items ($N = 2{,}514$) are analyzed with the logistic item response model and with full-information item factor analysis. The main results of the unidimensional analysis are that (a) item discrimination is heterogeneous, so that a Rasch model cannot be ascertained for the Paper Folding worksample, (b) item difficulty does not increase with presentation order, and (c) there are not enough items in the midrange of the scale. Item factor analysis identifies an additional minor factor attributable to during-the-test learning of specific solutions for items with similar stems. Based on these results, a modified 11-item version of the Paper Folding worksample is proposed.

Contents

Introduction 1

Data collection 1

Coding 1

Descriptive statistics 4

One-dimensional item response analysis 4

Full-information item factor analysis 9

Conclusion 13

References 15

List of Figures

1 Raw score distribution of the Paper Folding worksample 3

2 Item response curves for the Paper Folding worksample 6

3 Observed versus predicted item difficulties 8

4 Two-factor solution, varimax rotated 11

5 Test information and standard error of measurement curves for the 15-item Paper Folding worksample 12

6 Projected test information and standard error of measurement curves for the revised 11-item Paper Folding worksample 14

List of Tables

1 Descriptive statistics for the Paper Folding items 2

2 Test of fit: one- and two-parameter logistic models 5

3 Two-parameter logistic item parameter estimates 7

4 Item factor analysis: test of dimensionality 9

5 Two-factor solution, varimax rotated loadings 10

Introduction

The Paper Folding worksample was introduced into the Foundation's battery as a structural visualization test by Dean Trembly in the 1960s, and has since undergone several modifications based on contributions by David Ransom, Mark Daniel and Richard Smith. Previous studies (Kyllonen, Lohman, & Snow, 1984; Technical Reports 725 and 1986-1) indicate that paper folding items can often be solved through nonspatial strategies, a matter of concern in many so-called spatial tests.

The present report is part of an extensive validity study of the Foundation's spatial visualization tests. It summarizes our analyses of the internal psychometric characteristics of the current Paper Folding worksample with 5 experimental items included.

Data collection

Data were obtained from the 12 laboratories in Atlanta, Boston, Dallas, Denver, Los Angeles, New Orleans, New York, Philadelphia, San Diego, Seattle, Tulsa, and Washington, D.C. The experimental worksample comprised the 10 items from Wks. 622FB, augmented by the five experimental items proposed in Statistical Bulletins 1985-8 and 1985-10. The total sample size was 2,749; in all cases, the 10 original worksample items were administered, and responses for the 5 experimental items were collected fr. .ı 2,514 respondents.

Coding

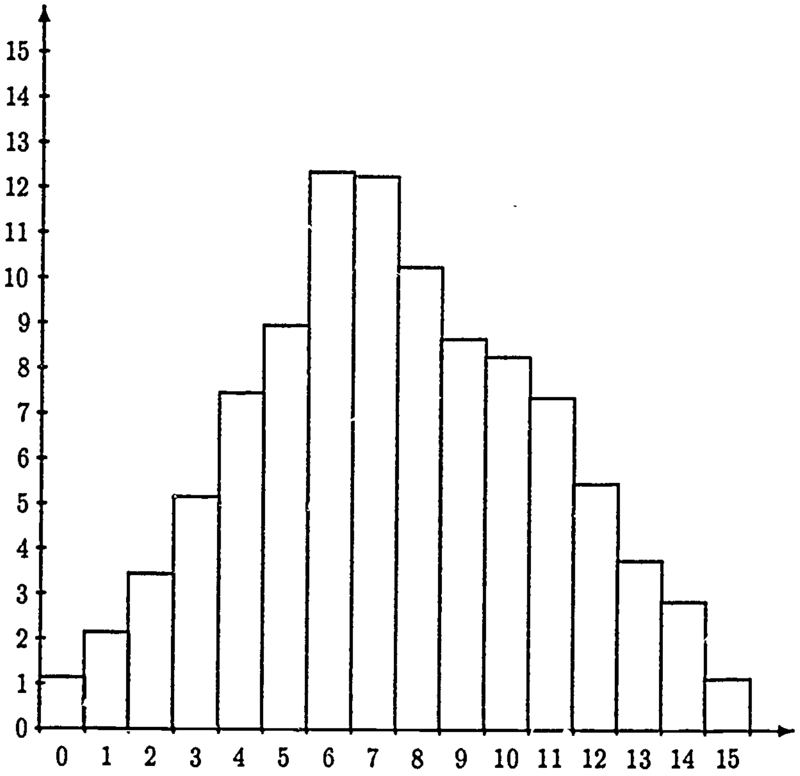
For purposes of item analysis, item responses were coded as right or wrong. Response time information was not examined.

Two attempts were allowed for each item, so that an individual item response could be *correct on the first trial*, *wrong on the first trial and correct on the the second*, or *wrong on both trials*. While the psychometric theory to analyze a graded response format of this type has been available for some time, the corresponding statistical software is still too underdeveloped for serious applications. To permit at least some momentarily feasible approach to multivariate item response analysis, answers were coded as binary right/wrong responses: right, when there was a correct response on either of the two trials, and wrong otherwise. We are well aware that this coding may appear somewhat arbitrary and have spent some effort on the question of whether one or two trials should be counted. Using the correct solutions from only the first trial essentially replicates the results stated in this report; however, the fit of the psychometric models is generally worse. It appears that respondents incorporate a "two-trial" allowance in their solution strategies, showing greater concentration on the second trial. This interpretation conforms to Mark Daniel's repeated

Table 1: Descriptive statistics for the Paper Folding items

Item	Percent correct	Item-Test Correlation		Comment
		Pearson	Biserial	
PF01	94.2	0.210	0.423	Experimental item
PF02	85.0	0.326	0.499	Old item 1
PF03	75.6	0.370	0.506	Old item 2
PF04	76.2	0.408	0.562	Old item 3
PF05	78.6	0.379	0.533	Old item 4
PF06	83.6	0.368	0.551	Experimental item
PF07	42.9	0.457	0.576	Old item 5
PF08	40.3	0.530	0.671	Old item 6
PF09	27.2	0.492	0.660	Experimental item
PF10	33.0	0.322	0.418	Experimental item
PF11	65.0	0.487	0.627	Experimental item
PF12	18.4	0.473	0.688	Old item 9
PF13	19.6	0.517	0.743	Old item 10
PF14	22.9	0.534	0.740	Old item 11
PF15	12.2	0.433	0.700	Old item 12

Figure 1: Raw score distribution of the Paper Folding worksample



findings (Technical Reports 856, 865) that Paper Folding scores are somewhat more reliable when the two trials are counted equally.

Descriptive statistics

Table 1 shows the percent correct and item-test correlation for each item. Item difficulties (percent incorrect) range between 6% and 88%, covering nearly the entire ability range of the sample. The earlier observation by Smith (Statistical Bulletin 1985-10) that items PF12 and PF15 were too difficult for New York and Chicago samples does not generalize to the national data. Nationwide, item difficulty levels of the Paper Folding worksample cover an appropriate range.

Item-test correlations are provided as indicators of item discrimination. Biserial correlations are acceptably large, ranging from 0.42 to 0.74; Pearson correlations are generally lower due to attenuation. Items 1 and 10 have the poorest discrimination, with biserial correlations of 0.42. Since it is well known that item-test correlations are sample dependent and drop at both ends of the difficulty scale, item evaluations should be primarily based on item response scaling analyses, which are relatively immune to this problem. The results from these analyses are presented below.

One-dimensional item response analysis

Unidimensional latent trait analyses were performed with both the one-parameter and the two-parameter logistic response models. The *one-parameter* logistic model, also known as the Rasch model, is based on the assumption that all items discriminate equally on a hypothesized latent ability scale. The probability of an individual item response would then depend only on the difficulty b_j of item j and on the ability level θ_i of person i. Under the *two-parameter* logistic model, items may also have different discriminations, expressed by an additional parameter a_j for the slopes. Both item response models are formally expressed in terms of the response probability function

$$P(X_{ij} = 1|a_j, b_j, \theta_i) = \frac{e^{1.7a_j(\theta_i - b_j)}}{1 + e^{1.7a_j(\theta_i - b_j)}}, \qquad (1)$$

describing the probability of the correct response in terms of an S-shaped probability ogive when plotted against ability levels θ_i for fixed a_j, b_j. The factor of 1.7 scales the item parameter estimates approximately into the metric of the Normal response function. In the Rasch model, all item slopes a_j are fixed at a common value.

The one-parameter logistic (Rasch) model is generally preferable since it greatly simplifies ability estimation. Experience has shown that fit of the Rasch model is acceptable with carefully constructed homogeneous tasks. For the majority of ability tests, however, items

Table 2: Test of fit: one- and two-parameter logistic models

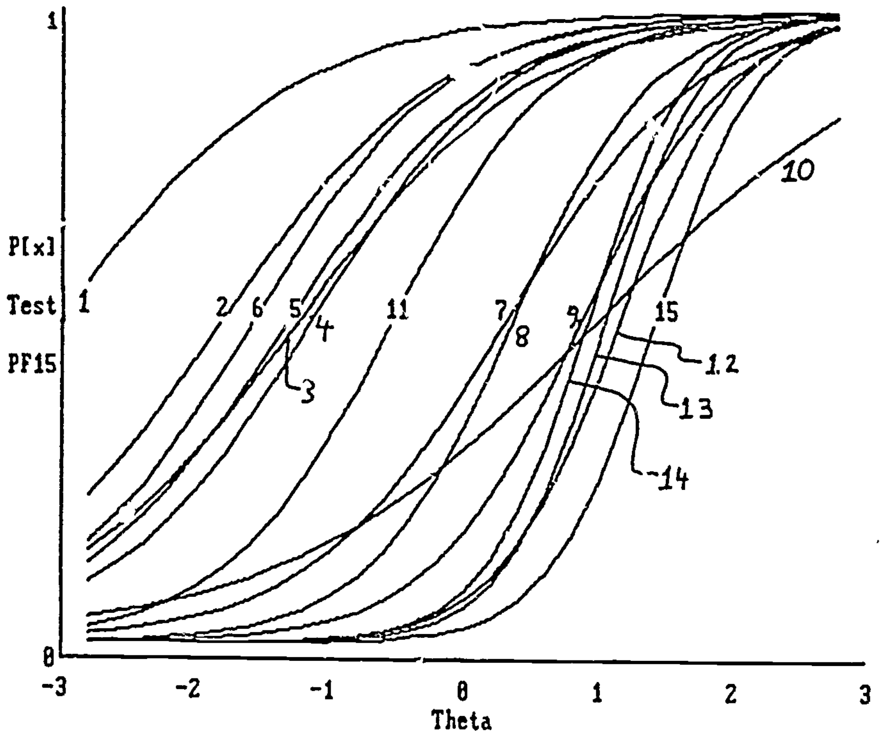
Model	-2log L	# item parameters	Diff χ^2	df
1-parm	35938.8	15	—	—
2-parm	35472.2	30	466.6	15

frequently differ in slope. The present analysis applies both models to the same item response data. The increase in fit from the one-parameter to the two-parameter solution can then be used as a test of the Rasch model. Item parameter estimation utilizes the marginal maximum likelihood method implemented in the BILOG program (Mislevy & Bock, 1984). This method provides a maximum likelihood χ^2 test for fit improvement, based on the difference in log-likelihoods from each model. The χ^2 statistic is generally inflated when a clustered sample design is employed. This is the case with the current data, where the Johnson O'Connor laboratories establish major clusters. The size of the design effect can be expected to be somewhat larger than those in more carefully controlled representative area cluster samples (*cf.* Frankel & McWilliams, 1981) and probably lies in the vicinity of 2.5.

Table 2 shows that the difference in model fit is highly significant and cannot be explained by sampling design effects. Therefore, item discrimination cannot be considered constant. The obvious culprit is easily identified in the trace line plot of Figure 2 as PF10, one of the experimental items proposed in Statistical Bulletin 1985-10. Slopes of the remaining items also show considerable variation, though this is not directly apparent from Figure 2; slope parameters in Table 3 are estimated with standard errors of 0.022 or less, and the last four items show twice the discrimination as the first seven.

There are two clear reversals in the order of item difficulty that are problematic. While, overall, the correlation between the present difficulty estimates and the earlier projections in Statistical Bulletin 1985-10 is very satisfactory at $r = 0.92$, experimental items [illegible] and 11 are much easier than the ones immediately preceding. It is apparent from Figure [illegible] hat items 1, 6, and 11 are much easier than predicted. On the latent scale of Paper Folding aptitude, item 6 is approximately 0.3 σ easier than the preceding three items, and item 11 is an entire standard deviation easier than the preceding four items. These difficulty reversals are substantial, especially considering that the standard error of the difficulty parameters is below 0.033 for all but the first two items. Since the Paper Folding worksample is designed to be a power test and since its administration is often terminated early, items 6 and 11 should be deleted or reordered in future versions of the worksample.

Figure 2: Item response curves for tn_ Paper Folding worksample



6

Table 3: Two-parameter logistic item parameter estimates

Item	Slope a_j	Difficulty b_j
PF01	0.644	-3.032
PF02	0.719	-1.806
PF03	0.673	-1.236
PF04	0.804	-1.143
PF05	0.764	-1.314
PF06	0.808	-1.574
PF07	0.818	0.295
PF08	1.113	0.351
PF09	1.069	0.861
PF10	0.502	0.972
PF11	0.959	-0.541
PF12	1.340	1.148
PF13	1.599	1.040
PF14	1.574	0.908
PF15	1.533	1.413

Figure 3: Observed versus predicted item difficulties

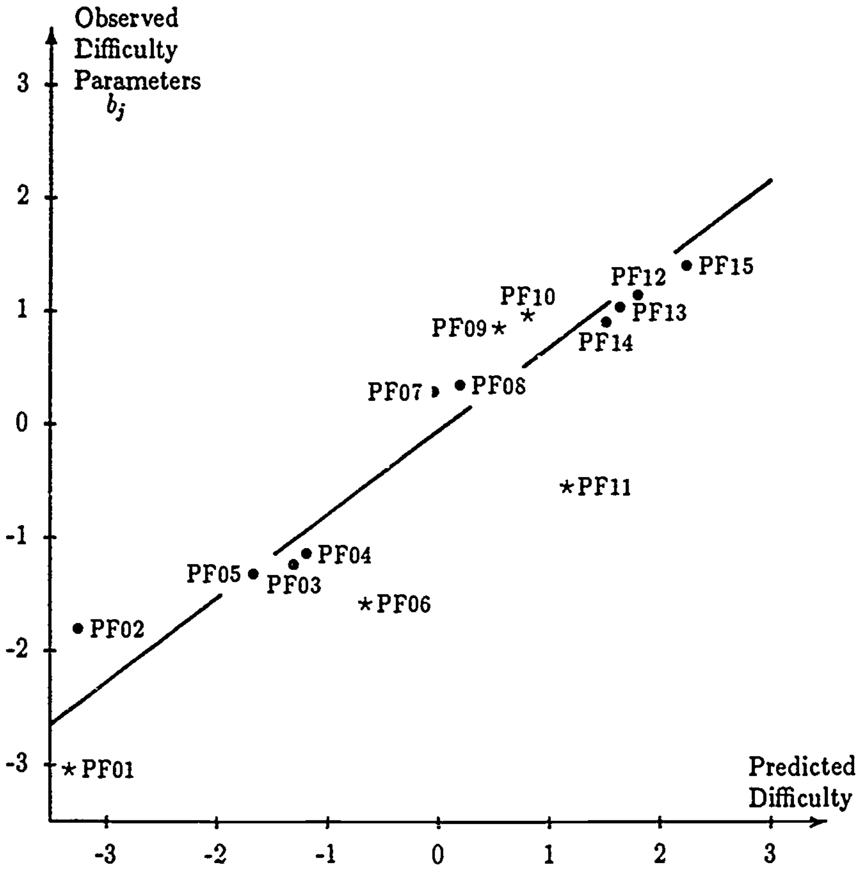


Table 4: Item factor analysis: test of dimensionality

# Factors	-2log$L$	# item parameters	Diff χ^2	df
1	35498.1	30	—	—
2	35357.5	44	140.6	14
3	35322.7	57	34.8	13

Full-information item factor analysis

The previous discussion was based on the tacit assumption that the items of the Paper Folding worksample are unidimensional, reflecting a single factor of ability differences. In the present section, this dimensionality assumption is tested using item factor analysis.

Item factor analyses were computed with the TESTFACT program (Wilson, Wood, & Gibbons, 1984), incorporating parameter estimation under a full-information approach. TESTFACT provides marginal maximum likelihood estimates of the multivariate *normal ogive* item response model:

$$P(x_{ij} = 1|\theta_i) = \Phi[Z_j(\theta_i)], \tag{2}$$

where Φ is the multivariate normal probability function with the argument

$$Z(\theta_i) = a_j + \sum_{k=1}^{m} b_{jk}\theta_{ik}. \tag{3}$$

The parameter a_j is the intercept of the item, expressing difficulty, and the b_{jk}'s are slopes on the m dimensions. The factor loadings are functions of the slopes, computed as

$$\lambda_{j\ell} = \frac{b_{j\ell}}{\sqrt{1 + \sum_{k=1}^{m} b_{jk}^2}}. \tag{4}$$

Since intercepts are estimated separately, and because full-information estimation is employed, the appearance of artificial difficulty factors is not a problem with TESTFACT solutions.

Similar to the one-dimensional latent trait analysis where the fit statistics of increasingly higher parameterized item response models are compared, factor models are evaluated relating the fit of incrementally higher-dimensioned factor models. The one-dimensional factor model corresponds closely to the two-parameter logistic model discussed in the preceding

Table 5: Two-factor solution, varimax rotated loadings

Item	Dim-1	Dim-2
PF01	0.127	0.519
PF02	0.371	0.416
PF03	0.355	0.436
PF04	0.399	0.488
PF05	0.335	0.524
PF06	0.244	0.646
PF07	0.545	0.340
PF08	0.591	0.461
PF09	0.575	0.474
PF10	0.346	0.316
PF11	0.472	0.539
PF12	0.776	0.273
PF13	0.804	0.332
PF14	0.825	0.297
PF15	0.738	0.362

section. The slight difference in the corresponding fit statistics shown in Tables 2 and 4 is due to negligible differences in the logistic and normal item response models. Table 4 indicates that a two-factor model fits the item responses significantly better than a unidimensional model, while the size of the Difference-χ^2 from two to three factors is marginal, approaching the expected design effect for the self-selected cluster sample.

The second factor, though having a significant contribution, is relatively minor and can be disregarded for many practical applications. The correlation of the oblique promax rotated factors is 0.79, large enough to combine the two factors. The two-factor solution does, however, suggest somc problems in the test construction, and future test revisions should incorporate these findings.

The varimax rotated item factor loadings in Figure 4 and Table 5 show two peculiar item clusters at the low and high ends of the first dimension. The cluster on the high end comprises items 12, 13, and 14. Items 12 and 14 are merely rotated versions of each other, and item 13 differs only by an added fold irrelevant to the solution. Similar item-design factors have been found in the DAT Space Relations test by Zimowski (1985) and should be avoided in the design of aptitude tests. The artificial nature of this cluster is corroborated by the difficulty levels of items 12, 13, and 14, which become successively *easier*, instead of becoming more difficult. Since item design is largely identical, the decrease in difficulty

Figure 4: Two-factor solution, varimax rotated

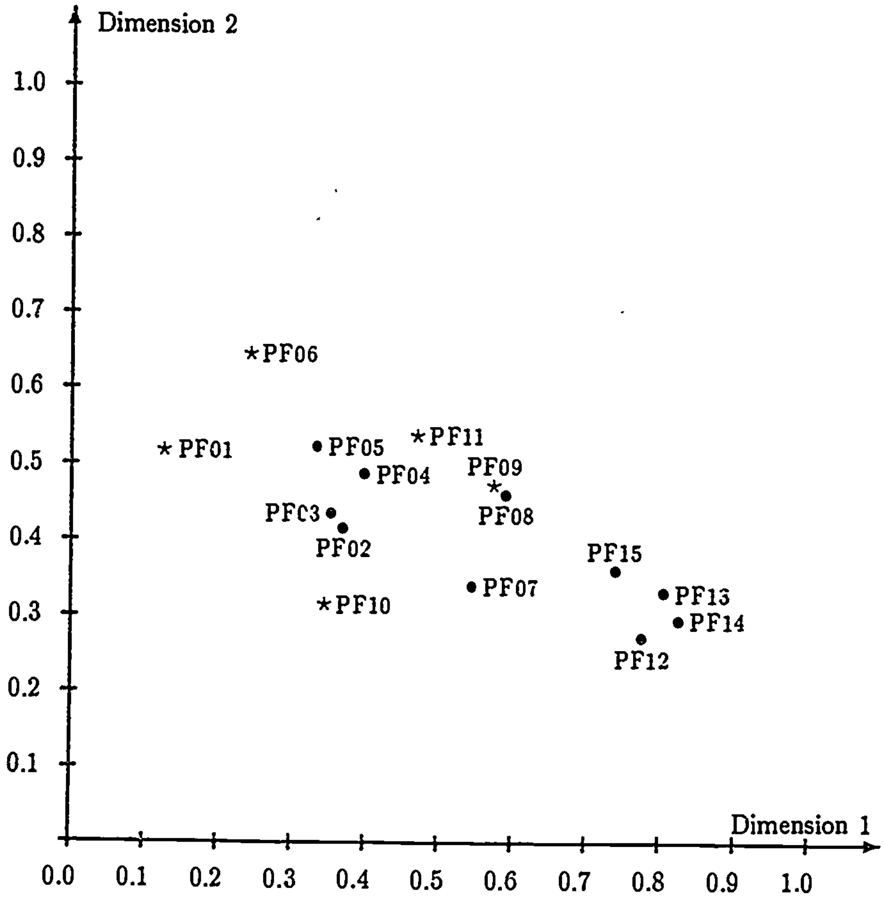
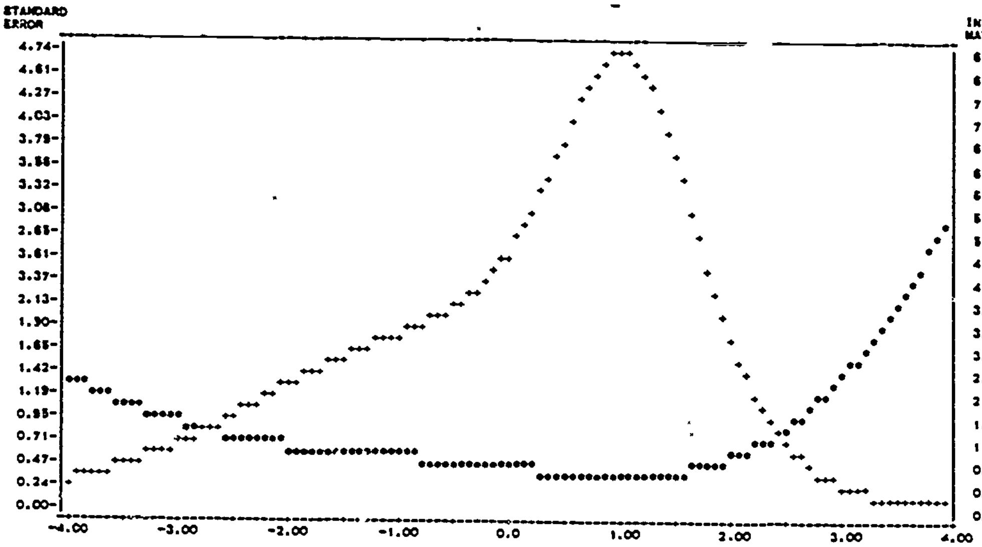


Figure 5: Test information and standard error of measurement curves for the 15-item Paper Folding worksample



indicates learning effects that occurred during test administration. Any such learning effects are intensified when feedback is provided after each response (as laid out in the worksample instructions given in Technical Report 844).

The low end of Dimension 1 shows a cluster made up of the two experimental items, PF01 and PF06. Both items show related folds not shared by the remaining 13 items. Only parallel horizontal folds are employed, so that the holes in the unfolded figure must lie within the same vertical column. Items 1 and 6 identify respondents who understand this specific solution principle.

The item factor solution is two-dimensional largely because these two item clusters were included in the worksample. Profile differences based on individual performance in the two clusters are relatively independent of the individual's average performance on the remaining Paper Folding problems. Since the item clusters identify learning effects and specific nonanalog solution strategies, their contribution is a function of deficient item design and should be considered a methodological artifact.

Conclusion

Overall, the 15-item experimental Paper Folding worksample shows acceptable psychometric characteristics: test reliability is high, and the items follow, by and large, a two-parameter logistic univariate response model. The test information curve and the (U-shaped) standard error of measurement curve are plotted against ability level in Figure 5. The standard error of measurement remains at or below 0.75 for most of the test range. The average expected reliability of EAP scores is 0.83, computed under the assumption that the aptitude is normally distributed with a mean of 0.0 and a standard deviation of 1.0.

The peculiar high peak of the test information curve at $\theta = 1.0$ is for the most part due to contributions from items PF12, PF13, and PF14. Since these items essentially duplicate each other, the height and steepness of the peak should be regarded as artifacts. Item duplication typically introduces learning components to the test and can thereby adversely affect the factorial structure of the worksample.

Two of the three items, PF12, PF13, and PF14, should be *deleted* from the worksample. It should be noted here that Daniel had already suggested replacing PF12 with an experimental item referred to as "C" (Statistical Bulletin 1981-1), based on item-test correlations. The folds of this experimental item are very similar to items PF12–14, so that the danger of learning effects would remain. Therefore, Daniel's item "C" should *not* be added to the worksample.

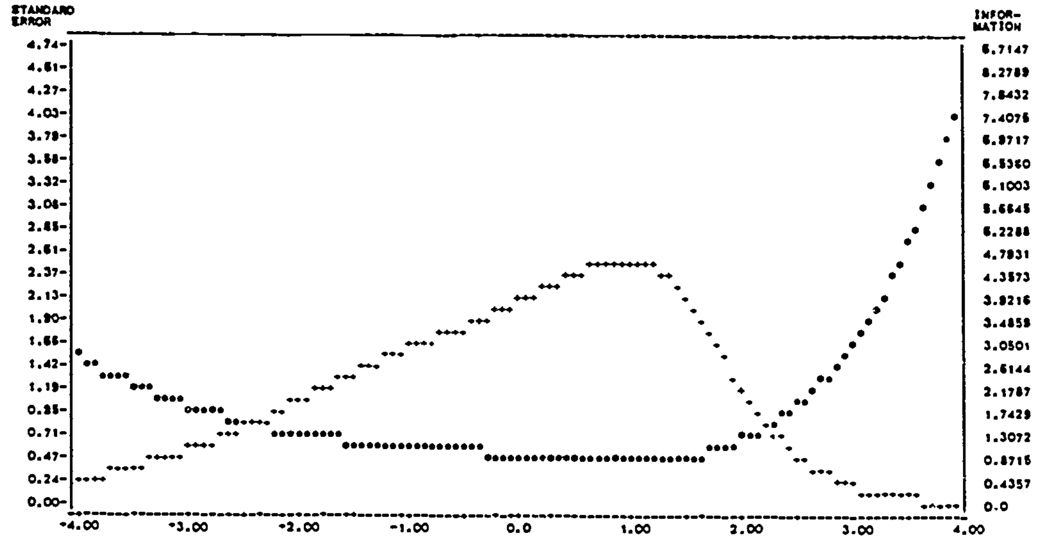
The new experimental items PF01 and PF09 appear to be well-fitting in their proposed difficulty ranges. These two items should be retained in future versions of the worksample.

Item PF11 fits well but is far too easy to be administered towards the end of the worksample. It should be presented immediately after PF05 in future administrations. New data are needed to judge whether rearranging the presentation order will affect the slope and difficulty of PF11. Effects of performance feedback, on the one hand, and of early test termination, on the other, make prediction uncertain.

Experimental items PF06 and PF10 should be *removed* from the worksample. PF10 discriminates poorly between high and low Paper Folding aptitude, while PF06 is too easy and clusters with PF01.

Figure 6 shows the projected test information and standard error of measurement curves for a resulting 11-item version of the Paper Folding worksample. Included are items PF01–PF05, PF07–PF09, PF11, PF12, and PF15. While the maximum test information is still found at $\theta = 1.0$ and does not coincide with the mean of the aptitude distribution, the information curve is now noticeably less peaked. However, deleting items PF06, PF10, PF13, and PF14 affects the measurement error only marginally: the average expected reliability

Figure 6: Projected test information and standard error of measurement curves for the revised 11-item Paper Folding worksample



for EAP scores is 0.78, a very sizable value for 11 dichotomous items.

A final problem with the Paper Folding test is the lack of items in the medium ability range between $\theta = -1.1$ and $\theta = 0.3$, *i.e.*, items with between 43% and 76% correct solutions. Judging by the item trace lines in Figure 2, the experimental items introduced in Statistical Bulletin 1985-10 have not resolved this problem. Item PF11 could alleviate the problem somewhat. To improve overall measurement performance of the Paper Folding worksample, we suggest the development and testing of additional experimental items for the midrange of the scale. Care should be exercised not to duplicate item stems. For example, experimental items "E" and "I" described in Statistical Bulletin 1981-1 are promising: they have difficulties within the desired range, the item-test correlations are substantial, and their content appears to be sufficiently different from the current items in Worksample 622F.

References

Frankel, M.R., & McWilliams, H. (1981). *The profile of American youth: Technical sampling report*. Chicago: National Opinion Research Center.

Kyllonen, P.C., Lohman, D.F., & Snow, R.E. (1984). Effects of aptitudes, strategy training, and task facets on spatial task performance. *Journal of Educational Psychology*, **76**, 130–145

Mislevy, R.J., & Bock, R.D. (1984). *BILOG II. Item analysis and test scoring with binary logistic models*. Mooresville, IN: Scientific Software

Statistical Bulletin 1981–1. *Revision of Paper Folding, Wks. 622 FA*. M. Daniel. Boston: Johnson O'Connor Research Foundation.

Statistical Bulletin 1985–8. *Results of component difficulty analysis for the Paper Folding Worksample 622 FB*. R. Smith & K. Green. Chicago: Johnson O'Connor Research Foundation.

Statistical Bulletin 1985–10. *Research proposal: Experimental items for Paper Folding Worksample 622FB*. R. Smith. Chicago: Johnson O'Connor Research Foundation.

Technical Report 725 (1969). *Paper Folding research, Worksample 622 Form A: Preliminary analysis leading to Form AA*. I.C. Shambaugh. Boston: Johnson O'Connor Research Foundation.

Technical Report 844 (1976). *Test description and procedure for administration of the individual board version of Paper Folding, Worksample 622 Form D*. D. Ransom. Boston: Johnson O'Connor Research Foundation.

Technical Report 856 (1977). *Preliminary analysis of the Paper Folding Worksample 622, Form D*. M. Daniel. Boston: Johnson O'Connor Research Foundation.

Technical Report 865 (1978). *Second analysis of the Paper Folding Worksample 622, Form D, and suggestions for revision*. M. Daniel. Boston: Johnson O'Connor Research Foundation.

Technical Report 1986–1. *The measurement of human variation in spatial visualizing ability: A process-oriented perspective*. M.F. Zimowski & W. Wothke. Chicago: Johnson O'Connor Research Foundation

Wilson, D.T., Wood, R., & Gibbons, R.T. (1984). *TESTFACT. Test scoring, item statistics, and item factor analysis*. Mooresville, IN: Scientific Software

Zimowski, M.F. (1985). *Attributes of spatial test items that influence cognitive processing*. Unpublished doctoral dissertation, University of Chicago, Chicago.