ED 479 923 TM 035 227

AUTHOR	Sadesky, Greg S.
TITLE	Exploring Boundary Conditions on the Invariance Property of Item Parameter Estimates of Item Response Models.
PUB DATE	2003-00-00
NOTE	16p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	Ability; Estimation (Mathematics); *Item Response Theory; *Robustness (Statistics); Simulation
IDENTIFIERS	*Invariance; *Item Parameters

ABSTRACT

In this study, the robustness of item parameter estimates with respect to the underlying distribution of abilities was explored. Using simulated datasets, item parameters were derived from a large sample representing the population versus samples representing a subset of ability and subsequently compared for possible mismatch. Comparisons were made under these conditions: (1) increasing differences between ability of examinees in the pilot sample and those of the population; and (2) the size of the pilot sample. All datasets were generated using DIFSUM, a computer program that permits the creation of dichotomous test data according to prespecified characteristics. In general, the study shows that item parameter estimates derived from a sample contain more errors when ability differences exist between the sample and its population. Fewer errors result when the size of the sample is increased. This finding highlights some limitations of the item invariance property claimed by item response theorists. Some of the limitations of the study are also discussed. (SLD)

RUNNING HEAD: Item Invariance

Exploring Boundary Conditions on the Invariance Property of Item Parameter Estimates of Item Response Models

Greg S. Sadesky

University of Alberta,

Edmonton, Alberta, Canada

Exploring Boundary Conditions on the Invariance Property of Item Parameter Estimates of Item Response Models

The property of item parameter invariance is a central feature of item response theory (IRT). This property implies that the accurate estimation of item parameters for a given population does not depend on the individuals in the sample. Rather, item parameters can be derived from separate groups when the examinees in those groups come from the same underlying population.

There are some caveats to this property, however. Hambleton, Swaminathan, and Rogers (1991) point out that strict item parameter invariance only holds when the fit between the model and data are exact. Since an item characteristic curve (ICC) is based on the *best* fit of a logistic model to the data, estimates of the probability of a correct response at a given ability level may differ from the observed probability. Reasons for an observed difference include insufficient sample size, violation of the assumptions of the model, or even that a logistic function is not an appropriate model of performance on a given item (e.g., Hambleton, Jones, & Rogers, 1993).

One procedure that depends on the item invariance property is the pilot or field-testing of items for inclusion in operational test forms. A population of items is administered to students and on the basis of the performance of those items, a test is constructed that best matches the test specifications. Item parameter estimates are a key source of evidence for evaluating the psychometric characteristics of the field test items. However, if the assumptions underlying the item invariance property do not hold, item parameter estimates may not be accurate and therefore may not conform to the specifications of the test (Gierl et al., 2001; Hambleton & Jones, 1994).

Piloting procedures may be susceptible to the above problems when the characteristics of the sample are not known at the time of testing. For example, the province of Alberta, in their field-testing procedure for province-wide exams, pilots items only to students in schools that volunteer. As a result of this reliance on volunteer participation, no control can be maintained over the characteristics of the sample --most notably, the ability level of the examinees. The possible consequence of this lack of control is poor estimates of item parameters and therefore, a test whose performance does not match test specifications.

Some research has been conducted to highlight the central role of between-groups ability differences in accurate item parameter estimation. Gotzmann (2001) demonstrated that item parameters derived from two distinct samples of examinees could differ when the mean ability of each sample differs. Specifically, she showed that when an item's parameters are estimated separately from two different samples, the probability of the estimates being significantly different increases with larger increases in ability. However, these differences are partially mitigated by larger sample sizes in each of the groups being compared. Gotzmann (2001) also suggested that the effect of large ability differences between groups (i.e., 1.5 standard deviations and greater) on the accuracy of item parameter estimation have not been well researched. Therefore, the boundary conditions on accurate estimation with respect to the above ability differences have not yet been established.

So that test designers select items to meet proposed test specifications, it is important to know how the accuracy of item parameters is affected when there exist ability differences between the pilot sample and the population. Specifically, the relationship between the size of the pilot sample, the magnitude of ability differences between pilot and population, and the resulting effects on parameter estimation should be examined. Test designers, to determine the desired number of examinees to compose the pilot sample could use this information. In this case, partial information about ability differences perhaps combined with the perceived consequences of inaccurate parameters could be used to design a piloting program that would be less susceptible to error and thus contribute to the assembly of a test that matches intended specifications.

In order to determine the accuracy of item parameter estimation under conditions similar to those involved in piloting items, a methodology for detecting differences in parameters in required. One such methodology is that used to study differential item functioning, or DIF. DIF is said to occur when examinees from separate groups have different probabilities of answering an item correctly after examinees in those groups have been matched on ability. Though typically used to study between-group item bias (e.g., between males and females, White and African-American examinees), the presence of DIF could signal errors in item parameter estimation between a pilot sample and the population. In this case, the 'true' item parameters would be derived from a large sample whose mean ability matches that of the population. Mismatch between these parameters and those derived from a smaller sample whose mean ability differed significantly from the mean of the population could be examined. Since the characteristics of the pilot sample deviate either from the known characteristics of the population, optimal sample size required from accurate estimates, or both, observed mismatch could be attributed to item parameter estimation errors from the pilot sample. Using simulated data, this is the methodology to be used in the present study.

In order to examine DIF, the Simultaneous Item Bias Test (SIBTEST, Shealy and Stout, 1993) will be used. SIBTEST is a computer program that detects DIF between two groups of examinees by comparing differences in their respective item response curves (IRCs). These comparisons are made between examinees that have been matched on the total number of items answered correctly. In effect SIBTEST examines the area between the IRCs derived from the *focal* and *reference* groups, respectively the pilot sample and the population. In the present investigation, the presence of items exhibiting DIF will be interpreted as evidence for item parameter estimation errors derived from the pilot sample.

Thus, in the following study, the robustness of item parameter estimates with respect to the underlying distribution of abilities of the sample is explored. Using simulated datasets, item parameters will be derived from a large sample representing the population versus samples representing a subset of ability and subsequently compared for possible mismatch. Comparisons will be made under the following conditions: (a) increasing differences between the ability of examinees in the pilot sample and those of the population, and (b) the size of the pilot sample. The intention of this study is to shed light on both the robustness of the item invariance property and the practical consequences of piloting items using examinees of a limited range of ability and limited sample size.

Method

Data Simulations

All datasets were generated using DIFSIM (Stout Research Lab). DIFSIM is a computer program that permits the creation of dichotomous test data according to pre-specified characteristics. As the name implies, the program was designed to assist in the analysis of dataset characteristics underlying the presence of DIF, and therefore each run generates data for both a reference and a focal group. All items in each group were conceived as uni-dimensional and were generated following the 3 parameter logistic (3PL) model. Item parameters used to create the items were obtained from the 1994 administration of the Alberta Grade 6 Provincial Achievement Test in science and are displayed in Table 1. A total of twenty simulations were conducted, two for each of the 10 conditions outlined below.

Ability Differences

Five ability levels of the pilot sample (θ_{ps}) were examined: 0, -.5, -1.0, -1.5, and -2.0. At each level of the pilot sample, the mean ability of the population was set at 0. Standard deviations for both groups were set at 1. The negative values reflect lower performance in the pilot sample as compared with the population. Although differences between the two groups could conceivably go in either direction depending on the characteristics of the examinees in the pilot sample, anecdotal evidence suggests that negative differences have occurred more frequently in the Alberta Provincial Achievement Testing program (personal communication, Mark J. Gierl, April 2003). The range of ability differences was intended to capture the possible range encountered in practical testing situations.

Sample Size

Two sizes for the pilot sample were chosen: 400, 800. These two values were selected because of their relationship to current testing practice in Alberta. Four hundred examinees represent the maximum number of students piloted for most PATs. For the Alberta Grade 12 departmental exams, a high-stakes test worth 50% of students final grade for each of the core subjects, approximately 800 students compose the pilot sample (personal communication, Mark J. Gierl, Apr. 2003). The size of the population was kept constant across all conditions at 4000. This value was considered sufficient to ensure the stability of item parameters derived from the population.

Determination of DIF

In order to determine whether a given item exhibits DIF, criteria from Nandakumar (1993) were adopted. Under these criteria, items are considered to exhibit negligible or A-level DIF when the SIBTEST test statistic $\beta_{UNI} < .05$, moderate or B-level DIF when $\beta_{UNI} \geq .05$ and $< .1$, and large or C-level DIF when $\beta_{UNI} \geq .1$. For the purposes of this study, the presence of DIF was implied when an item exhibited either B- or C-level DIF.

Item parameter estimation accuracy for each of the pilot samples was defined as follows. First, a sample was considered to have acceptable levels of estimation accuracy if two or less items on the test exhibited DIF. This corresponds to 5% of the test items and thus is comparable to a nominal type I error rate of .05. A medium level of DIF on a test was defined as having between three and six DIF items, corresponding to a type I error rate of .075 to .15. Last, a high level of DIF was defined as 7 or more items out of 40 exhibiting DIF, or a type I error rate of .175 or higher.

Results

The results for each of the conditions are displayed in Table 2. For the no ability difference condition, the average number of items exhibiting DIF did not exceed 2 in either of the two sample size conditions, $\underline{M}_{0,400} = 2$, $\underline{M}_{0,800} = 0$, indicating a low level of DIF. Similarly, when the ability level of the pilot sample was -0.5, the level of DIF again was low, $\underline{M}_{-0.5,400} = 0$, $\underline{M}_{-0.5,800} = 1.5$. However, at a mean difference of -1.0, a medium level of DIF was observed in the n=400 group, as opposed to low in the n=800 group, $\underline{M}_{-1.0,400} = 4$, $\underline{M}_{-1.0,800} = 0$. This difference carried through to the next largest ability difference, -1.5, with the small sample group showing a high level of DIF, ($\underline{M}_{-1.5,400} =$

7.5) while the larger sample groups showed a medium level of DIF ($\underline{M}_{-1.5,\underline{800}}$ = 3.5). Last, for the largest ability difference, -2.0, both groups showed a high level of DIF, $\underline{M}_{-2.0,\underline{400}}$ = 16, $\underline{M}_{-2.0,\underline{800}}$ = 10.5.

To summarize, for the two smallest ability differences between pilot sample and population, $\theta_{ps} = 0$, and $\theta_{ps} = -0.5$, no differences in the amount of DIF was observed between the two sample size groups, n = 400 and n = 800. Under these conditions, the amount of DIF for both groups was low. At the next largest ability differences, differences in the amount of DIF was observed between the two sample sizes, medium and low for n = 400 and n = 800, respectively at the $\theta_{ps} = -1.0$ condition; high and medium for the two sample sizes in the $\theta_{ps} = -1.5$ condition. Last, both sample size groups had high levels of DIF at the largest ability difference, $\theta_{ps} = -2.0$.

Discussion

The purpose of this study was to examine the conditions that support the robustness of item parameter estimates. The motivation for this examination was twofold. First, this study could shed light on the boundary conditions of the item invariance property of item response models. Second, the piloting of test items for inclusion on operational tests depends directly on the accuracy of item parameter estimates derived from the pilot sample. Thus, a determination of the conditions that most likely lead to accurate estimates would be useful.

Implications for Item Invariance

In general, this study showed that item parameter estimates derived from a sample contain more errors when ability differences exist between the sample and its population. Fewer errors result when the size of the sample is increased. On the surface, this finding appears to contradict the item invariance property claimed by item response theorists.

The property implies that item parameter estimates for the same items should be the same, regardless of the sample from which they are derived. Thus, the present finding highlights some limitations of this property.

Given the characteristics of the groups of examinees that underlie the parameter estimates, deviation from strict item invariance can be attributed to sampling factors. For example, the conditions under which these estimates were derived are not consistent with the sample size recommendations for the 3PL model. The two sample sizes in this study fall well short of the recommended size of 1500 (Mark, do you have a reference for this?). However, the erosion of the accuracy of parameter estimates also appears to depend on the overlap of ability distributions between the two samples from which parameters are derived.

Several possibilities exist to account for the increase in errors. First, the SIBTEST procedure compares the IRCs from the two groups after having matched the respective examinees on total score. When one of groups has few examinees at a given score level, the precision of the points comprising the IRC will be subject to greater standard error. This could lead to greater levels of observed DIF. If the above account is true, increasing the variance of the pilot sample could have a similar effect to increasing sample size. That is, if the whole ability spectrum were represented in the pilot sample despite differences in mean ability, more examinees would be comparable at each score point. Provided this still allowed for enough examinees at each score point, this could lead to increased precision in the SIBTEST procedure. Of course, this is an empirically testable hypothesis, accomplished with the simulation methodology employed in the present study.

A second related possibility is that less variance exists in examinee responses to particular *items* resulting from a mismatch between their level of difficulty and the ability level of the sample of examinees. In the current study, such a mismatch might exist when an item is difficult for the population as a whole and the ability level of the sample is low. In this case, DIF might arise not because the two groups are from different populations, but because errors in item parameter estimation may result from item difficulty / examinee ability mismatch. In this case, items showing this mismatch would be flagged for DIF more often than those that do not.

In order to determine the feasibility of this hypothesis, all items in the present study were examined with respect to the frequency across all conditions and replications that an item was flagged for DIF. A correlation was conducted between this frequency and the item difficulty parameters listed in Table 1. The resulting correlation was modest, but positive, $r = .48$, indicating that the more difficult the item, the more likely it was to be flagged for DIF. Although this result provides tentative support for the above hypothesis, a larger scale investigation is needed to determine its tenability.

Piloting Items

With respect to the piloting test items, this research has practical advice for anticipating potential problems. First, it is clear that piloting items on a sample based on a limited subset of abilities can lead to item estimation errors. When sample sizes are small (i.e., $n \sim 400$), these errors can emerge with differences as little as 1 standard deviation. Differences greater than this can result in errors even with larger sample sizes. In general, however, item estimation errors are less prevalent with larger sample sizes.

Recommendations for conducting pilots similar to those in the province of Alberta include gathering information about the ability level of the sample. For example, data may be collected on the historical performance on the test of the schools participating in the pilot. When the sample as a whole seems to be systematically above or below the mean, a larger sample size is advised. More effective, however, is to choose which schools will participate in a pilot. These schools could also be chosen based on their historical performance on these tests, the goal being to produce a representative sample of schools with respect to the ability level of their students.

Limitations

One obvious limitation of the present study is the number of simulations conducted in each condition. Though a trend is evident in the data, not enough replications have been considered in order to have precise estimates regarding the impact of each of sample size and ability difference on item parameter estimation errors. Other limitations include the examination of a limited number of sample size conditions. It would be informative for test designers to know the sample sizes required to steer clear of any potential parameter estimation problems and therefore a more extensive manipulation of this variable would be of use. Future research could be undertaken to address these shortcomings.

References

Gierl, M. J., Henderson, D., Jodoin, M., & Klinger, D. (2001). Minimizing the influence of item parameter estimation errors is test development: A comparison of three selection procedures. *Journal of Experimental Education, 69*, 261-279.

Gotzmann, A. J. (2001). *The effect of large ability differences on type I error and power rates using SIBTEST and TESTGRAF DIF detection procedures.* Unpublished master's thesis, University of Alberta, Edmonton, Alberta, Canada.

Hambleton, R. K., & Jones, R. W. (1994). Item parameter estimation errors and their influence on test information functions. *Applied Measurement in Education, 2*, 297-312.

Hambleton, R. K., Jones, R. W., & Rogers, H. J. (1993). Influence of item parameter estimation errors in test development. *Journal of Educational Measurement, 30*, 143-155.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 16*, 159-176.

Shealy, R. & Stout, W. F. (1996). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Author Note

I would like to thank Dr. Mark J. Gierl for his encouragement and guidance in this project, particularly with respect to methodological issues and how this research fits within current research in IRT and DIF.

Table 1

Parameters for All Items.

Item	*a*	*b*	*c*
1	0.6	-0.3	0.4
2	0.7	0	0.3
3	0.6	-1.6	0.3
4	0.5	-0.7	0.3
5	0.3	0.1	0.2
6	0.7	-0.5	0.3
7	0.6	1.8	0.2
8	0.8	-0.1	0.2
9	0.7	-0.8	0.1
10	0.7	0.4	0.2
11	0.9	0.9	0.3
12	0.8	0.2	0.4
13	0.6	-1	0.2
14	0.3	-1.6	0.2
15	0.7	-1.8	0.2
16	0.8	-1.5	0.2
17	0.7	-1.7	0.2
18	0.3	-0.7	0.3
19	1.1	-0.2	0.5
20	0.8	0	0.3
21	0.5	-2	0.2
22	0.6	-0.7	0.1
23	0.3	1.1	0.3
24	0.4	-0.8	0.2
25	0.7	-1.5	0.2
26	1.1	-1.1	0.2
27	0.6	-0.8	0.2
28	0.7	-1.1	0.3
29	0.5	0.1	0.3
30	0.3	2.2	0.2
31	1.3	-0.2	0.2
32	0.6	-1.5	0.2
33	1.1	-1.6	0.2
34	0.5	-1	0.2
35	0.5	0.4	0.1
36	0.9	-0.8	0.4
37	0.6	-0.1	0.2
38	0.8	-0.4	0.1
39	0.8	0.1	0.3
40	0.9	-1	0.2

Note. *a, b,* and *c* correspond to the discrimination, difficulty, and pseudo-guessing parameters.

Table 2. Average number of items (total=40) showing DIF compared across all levels of ability and sample size.

	n=400	*n=800*
0	2	0
-0.5	0	1.5
-1.0	4	0
-1.5	7.5	3.5
-2.0	16	10.5

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC
Educational Resources Information Center

REPRODUCTION RELEASE

(Specific Document)

TM035227

I. DOCUMENT IDENTIFICATION:

Title: Exploring Boundary Conditions on the Invariance Property of Item Parameter Estimates of Item Response Models

Author(s): Sadesky, Gregory S.

Corporate Source:

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☒	□	□
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please →

Signature: G. Sl

Printed Name/Position/Title: Greg Sadesky

Organization/Address: University of Alberta
6-110 Education North, University of Alberta,
Edmonton, Alberta, Canada T6G 2G5

Telephone: 780-492-5927

FAX: 780-492-0001

E-Mail Address: gsadesky@ualberta.ca

Date: Aug 11/2003

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20742-5701
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org