דו"ח מרכז

325

The Fairness and Validity of the Higher Education Selection System for Students with Disabilities

Carmel Oren
Ariela Even

August 2005

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

דוח מרכז 325

ISBN:965-502-113-0

The Fairness and Validity of the Higher Education Selection System for Students with Disabilities

NCME Annual Meeting, Montreal, April 2005

Carmel Oren*

(Corresponding author: carmel@nite.org.il)

Ariela Even*

* National Institute for Testing & Evaluation, Jerusalem, Israel

Abstract

This study investigates into two aspects of the selection of university applicants who requested test accommodations to compensate for various disabilities. The first is the fairness of the selection system, and the second is its predictive accuracy. The study groups consist of students at Israeli universities, selected by a combination of PET (Israeli SAT-like Psychometric Entrance Test) and Bagrut (high-school plus matriculation scores), who commenced their studies between the years 1992-1997. The focal groups comprise impaired applicants who were granted various accommodations to suit their disabilities, and a minority of applicants who were not found eligible for accommodations, and were therefore tested regularly. The regular students in the same departments of study served as the reference groups. Two approaches to defining prediction bias were employed. If we prefer a conservative point of view and adopt the Boundary Condition approach which requires that the two reverse regressions (criterion-on-predictor and predictor-on-criterion) show consistent results, we reach the conclusion that the whole selection system has not been proven biased against persons tested with accommodations. The second approach's requirement for an unbiased selection system is that the mean difference between a focal group and the reference group be equal for both the predictor and the criterion. In this view, the results of our study show – with respect to the non-eligible Learning Disability group and the Hearing Impaired group – that PET slightly under-predicts First Year Average (FYA). The English subtest of PET seems to be responsible for under-prediction of the Hearing Impaired and the non-eligible Learning Disability group, while the Quantitative subtest is over-predictive for most groups, suggesting over-compensation where the impaired skills are mainly verbal. Regarding the accuracy aspect of prediction: The selection system seems to be less predictive for both the Learning Disability and non-eligible Learning Disability groups than it is for the reference Regular Students group. The resemblance of validity profiles between PET and Bagrut suggests that the criterion's reliability might account, at least partly, for the decreased prediction accuracy in groups of students with disabilities. The non-eligible Learning Disability group's FYA grades appear slightly under-predicted, and less accurately predicted, by non-accommodated PET. Since no information was available about accommodations that these examinees may have enjoyed during their university studies, only a further study, that facilitates such control, might elucidate to what extent the criteria according to which LD candidates are found eligible for test accommodations need to be reconsidered.

1. Introduction

Special testing accommodations for examinees with disabilities have become common practice in most large-scale, high-stakes testing programs. This evolving process has facilitated an increase in the proportion of learning-disabled students in higher education. However, the issue of special accommodations raises a variety of questions, such as: what is the proper way to determine which - and whose - disabilities warrant special treatment? What is the proper and fair accommodation for any given disability? What measures will be used to determine this, and how can the scores be interpreted in a manner that avoids under- or over-prediction of academic success, and thus an unfair admissions procedure (Pitoniak & Royer, 2001)? It should be remembered that the disability itself might hamper the applicant's chances of succeeding in the exam, regardless of the knowledge or ability that is being measured (Willingham et al., 1988).

Accommodations are changes in the standard evaluation measures that aim to level the "playing field" for learning-disabled students, by reducing the variability resulting from the disability. Grades obtained on the basis of valid accommodations purpose to measure the same trait in disabled examinees as standard exams do in non-disabled examinees (Tindal & Fuchs, 1999).

Examinees with disabilities are not a homogenous group, but are characterized by a wide spectrum of sensory, motor and cognitive disabilities. Today, learning disabilities are the most common grounds for requesting test accommodations (Camara, 1998; Wightman, 1993), raising questions regarding changes in the validity of test results (Philips, 1994).

Despite concerns about test validity and interpretation, little research has been done on the influence of accommodations on test results. The difficulty in conducting such research stems mainly from the fact that groups of examinees with disabilities that are large enough to comprise meaningful units of analysis are hard to find.

To date, research conducted on the effects of test accommodations has explored two different characteristics of scores. The first characteristic is criterion validity (i.e., how well test scores are associated with another criterion or measure). The second is the comparability of scores (i.e., examinees with disabilities scoring as well as examinees without disabilities when provided with testing accommodations).

Research into criterion validity has shown that generally, scores obtained with accommodations have lower correlations with criterion measures (Laing & Farmer, 1984; Braun, Ragosta & Kaplan, 1986; Ziomek & Andrews, 1996). The criterion with which scores are usually compared is the student's grade point average.

From reviewing the results of the research studies done to date, it appears that scores from entrance examinations are less valid as predictors of postsecondary education grade point average for students with disabilities who received test accommodations, than for those who did not. The entrance examination scores predicted higher grade point averages than students with disabilities actually obtained.

Research into the comparability of scores has shown that when examinees with disabilities are provided with accommodations, their scores as a group are similar to those of examinees without disabilities taking the test under standard administration conditions. There is some research evidence that testing accommodations can assist individuals with disabilities in achieving scores comparable to the scores of individuals without disabilities (Centra, 1986; Bennet, Rock, & Jirele, 1986; Willingham et al., 1988; Ziomek, 1996). The existing, yet scarce, empirical research has mainly considered the question of time extension and its influence on test scores. Findings point to low correlations between scores on time-extended tests and grade point averages (Braun, Ragosta, & Kaplan, 1986; Ziomek & Andrews, 1996; Zurcher & Pedrotty Bryant, 2001).

Disabilities that require accommodations

In the past, accommodations were given primarily to examinees who had physical or sensory disabilities. However, in recent years, with growing consciousness and understanding of learning disabilities, greater emphasis has been put on the need to provide accommodations for examinees with cognitive disabilities. Indeed, the learning-disability group is now the largest group applying for accommodations (Camara, 1998; Wightman, 1993), and this fact repeatedly raises the same questions concerning changes in the validity of test scores (Philips, 1994).

Learning disabilities

Learning disabilities have been defined as "disorder(s) in one or more of the basic psychological processes involved in understanding or using language…manifest in imperfect ability to listen, think, speak, read, write, spell, or do mathematical calculations" [U.S Department of Education Guidelines, 1996, cited in Pitoniak & Royer, 2001, p.58]. The diagnosis of a learning disability is usually made by comparing the level of obtained achievements with the level predicted on the basis of intelligence. In this sense the discrepancy between the two levels is unexpected (Gresham, MacMillan, & Bocian, 1996).

The discrepancy view has raised some conceptual and practical problems, such as an exclusionary rather than symptom-based definition and the absence of guidelines regarding how large this discrepancy has to be in order to qualify as a learning disability.
A new method of identifying learning disabilities is related to the efficacy of instructional practices (Aaron, 1997; Cisero, Royer, Merchant, & Jackson, 1997; Fuchs & Fuchs, 1998). This means that a student with a learning disability fails to profit from an instructional practice and is in need of an alternative that is especially designed to compensate for his/her disability. The accommodations often approved for students with learning disabilities are extra time, readers, transcribers, and/or the use of word processors. The provision of these accommodations is intended to compensate for the information-processing and writing skills deficits of students with learning disabilities (MacArthur, 1996).

Sensory and physical disabilities

As noted above, applicants with learning disabilities account for the greatest proportion of examinees who request accommodations. The definition of a learning disability is far more problematic than the definition of other disabilities. Examinees with a specific sensory or physical deficit also differ in their cognitive functions and their disability projects on their performance and has bearing the nature of accommodations needed. It is also important to consider the wide spectrum of the disabilities, both in nature and degree (Willingham et al., 1988).
This heterogeneous group can be divided into 3 main sub-groups which, second only to the learning disabilities group, comprise the largest proportion of examinees applying for testing accommodations:

- Hearing impairment – deafness or some degree of hearing difficulty
- Visual impairment – ranges between substantial visual impairment and blindness
- Physical disability – a wide variety of motor and neurological impairments

NITE's testing accommodations unit

The higher education admissions process in Israel is based primarily on two indicators: Bagrut – the national matriculation exams, which reflect high-school achievements, and PET – the Psychometric Entrance Test, a battery quite similar in nature to the SAT. PET, with various adjustments and accommodations, has been offered for many years now to applicants with disabilities who were carefully diagnosed, screened and found eligible.

The testing accommodations unit, at the Israeli National Institute for Testing and Evaluation (NITE) deals with providing accommodations on PET for people with various disabilities and impairments. It also handles the screening and sorting of examinee applications for test accommodations.

There are various accommodations provided for examinees:

- Modification of the medium in which the instructions and questions are presented. For example, examinees with visual impairments may require enlarged print, a reader, or the option of using a tape recorder.
- Modification of the medium used for responding to questions. Examinees who are not capable of using the regular answer sheet may be assisted by a person who records their answers.
- Extra time – difficulties in information processing, the use of a reader or a tape recorder, or reading in large print fonts often takes much more time. Extra time is therefore allocated for taking the test. It is worth noting that there is not much data addressing the issue of the exactly how much time should be allocated for various disabilities.
- Visually and hearing impaired individuals sometimes require modifications in test content. For instance, sometimes the exam involves abilities, knowledge or concepts that are dependent on visual ability. This can be accommodated by replacing such items with non-visually dependent ones. Yet, given the possibility that replacement items differ in other characteristics from the problematic items, new items require careful screening.
- Other modifications include: Changes in the examination venue if it is not suitable for handicapped persons; adapted tables and chairs and other accommodations pertaining to the physical test-taking conditions.

2. Objectives of the study

The present study investigates two aspects of the predictive validity of accommodated tests:

- Prediction bias, or fairness – Do similar test scores predict similar academic achievements for regular students and students with disabilities who are given accommodations?

- Prediction accuracy – Is the selection process as effective, or accurate, in predicting the academic success of regular students as it is for the academic success of students with disabilities who were tested with accommodations?

3. Research design

Independent variables (predictors):

- ***Bagrut*** (B) – (Matriculation) average score – based on a weighted combination of high school grades and the scores of the national matriculation exams. Scale: school grades, ranging from 40 to 120 (100 plus various bonuses for enhanced test levels).
- ***PET*** total (P) – A total score in which PET's three sub-tests' relative weights are: Verbal Reasoning 40%, Quantitative Reasoning 40%, and English 20%. These sub-tests are described below. PET is a multiple-choice battery, designed to measure various aspects of developed scholastic abilities and skills and aimed to predict future academic performance in higher education. Scale: 200 to 800, historic mean of 500, SD=100.

  The three sub-tests which comprise the PET are:
- ***Verbal Reasoning*** (V) – This section includes 60 items that focus on the verbal skills and abilities manifested in academic performance: analysis and comprehension of complex written material, systematic and logical thinking, drawing fine distinctions between the meanings of words and concepts. The Verbal section includes items such as synonyms and antonyms, analogies, sentence completions, logic and reading comprehension.
- ***Quantitative Reasoning*** (Q) – Includes 50 items that focus on the use of numbers and mathematical concepts (algebraic and geometrical) in solving quantitative problems and analyzing information presented in graphs, tables, and charts. In this sub-test, only a basic level of mathematics is involved – that which is acquired in the ninth or tenth grades in most high schools in Israel. Formulae and explanations of mathematical terms that may be needed in the course of the test are offered in the test booklet.
- ***English*** as a foreign language (E) – Includes 54-58 items designed to test mastery of reading and comprehension of academic-level texts in English. This section includes three types of items: sentence completion, restatements, and reading

comprehension. This sub-test serves a dual purpose: it is a component of the PET total score, and it is also used for placement of students in remedial English classes.

Scale of all three subtests: 50-150, historic mean of 100, SD=20.

- ***Composite score (C) – Bagrut and PET with equal weights. Scale: mean=50, SD=10.***

Dependent variable (criterion):

- ***University first year average score*** – (FYA). Scale: school grades, ranging up to 100.

Population

Students who began their first year of studies at one of the six Israeli universities during the years 1992-1997, took PET during these years (but before commencing their studies), and for whom first year grades were obtainable.

Groups

The source data set of 64,731 students is broken down as follows:

- 63,291 regular students (RS) who serve as the reference group;
- The focal groups consisting of students with disabilities who were given test accommodations: 821 students with learning disabilities (LD), 159 physically handicapped students (PH), 90 visually impaired students (VI), 19 blind students (BL), and 28 hearing impaired students (HI).
- Another group of interest includes students who had applied for accommodations but were **not found eligible** for special conditions, 276 of whom declared learning disabilities (LD-n), and 47 of whom declared physical problems (PH-n).

(It should be noted that some of the above groups are very small, limiting the level of generalizability their data affords. They are presented to provide a general impression.)

Simple statistics of all raw variables pooled across groups are presented in Table 3 (Means) and Table 4 (SD's) in Appendix 1.

Data Analyses

Unit of analysis

The basic unit of analysis was department of study (class), per institution per year, that included at least one student with disability. All variables (both predictors and criterion) were standardized (0, 1) within the unit of analysis. The main purpose was to achieve a comparable FYA scale across all classes and to avoid variance between classes in student ability (measured by the predictors), resulting from differential admissions policies and/or different levels of prestige.

Computed statistics

Four main statistics were calculated – the first three pertain to the question of prediction bias, and the third facilitates comparing the accuracy of prediction between groups:

1. D-values: Standardized differences between groups (focal minus RS, the latter serving as reference group) – for both predictors and criterion. Since all study variables were normally standardized (0, 1) within units of analysis, and the proportions of the focal groups' n's are negligible in comparison to the majority RS groups, average group D-values are practically equal to their average Z scores (mean Z scores of the RS groups practically equal zero). A negative group D-value of a variable, for instance, represents the magnitude of gap, the focal group stands below the RS group on that variable, in standard scores (Z) terms.
2. D'-values: Differences between each predictor's D-value, and the criterion D-value. A negative D'-value for a predictor would imply under-prediction, while a positive D'-value would indicate over-prediction. In the spirit of Thorndike (1971) and Darlington's (1971) second definition of bias ("equal representation"): A predictor is not biased regarding a focal group only if $D_{predictor} = D_{criterion}$
3. Residual Scores: The difference between the actual, and the predicted FYA was calculated using a linear regression, and averaged within groups. The prediction parameters are practically based on the vast majority (RS) group within each unit of analysis. A negative residual score for a predictor would imply that the actual performance of a group on the criterion is lower than the predictor predicted. This is classically referred to as over-prediction (Cleary, 1968). This regression model for prediction bias is criticized as inherently tending to show over-prediction for the lower-scoring groups, merely because of statistical artifacts, such as those resulting from the effects of low reliabilities and exclusion of relevant predictors

from the regression(Linn, 1990). We therefore adopted Linn's (1984) boundary conditions approach, which is based on Birnbaum (1979, 1981), and we used two linear regression analyses for each predictor – one regressing the criterion on the predictor, and the "inverse" regression – regressing the predictor on the criterion. Using both regressions makes it possible to draw a conclusive inference regarding prediction bias, thus avoiding the abovementioned biased results, which might stem merely from statistical artifacts. Only if the inverse regression residual score, whether positive or negative, is inverse in sign to the regular regression residual (meaning the same direction of bias), can we infer that there is conclusive evidence of bias. The residual scores presented (in Table 1) are those of the regression of the criterion on predictors, and the values marked with asterisks are the ones for which we can, with a considerable level of confidence, conclude bias.

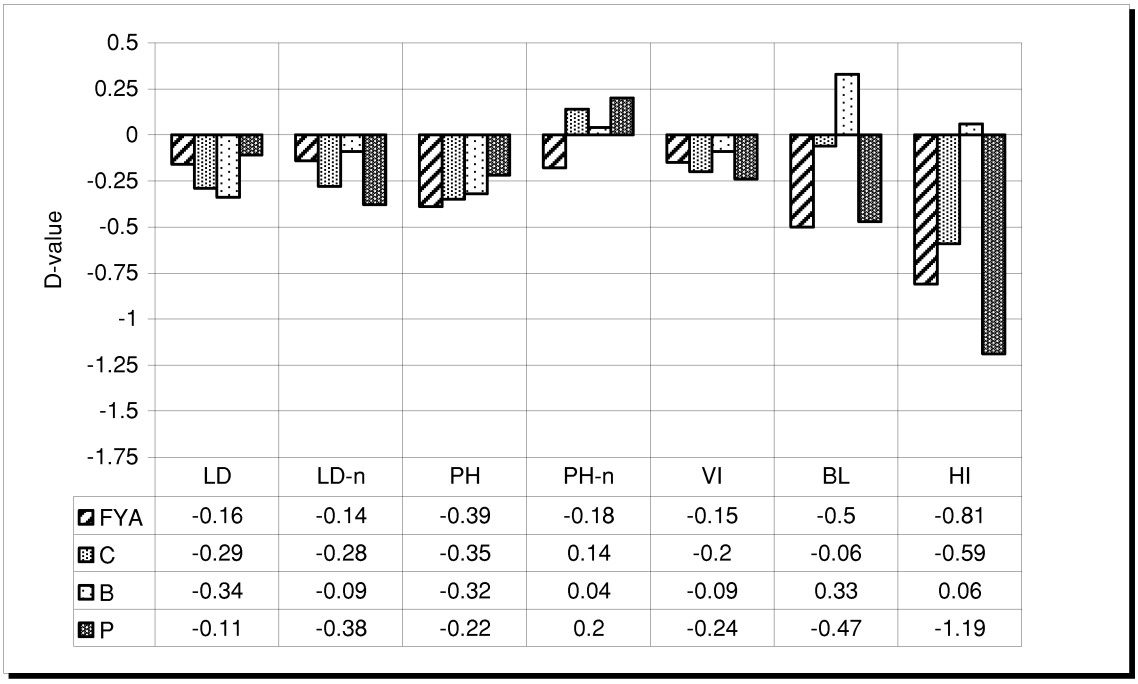
4. Pearson Correlations. Computed within all study groups between standardized-within-units predictors and FYA.

4. Results

Prediction bias

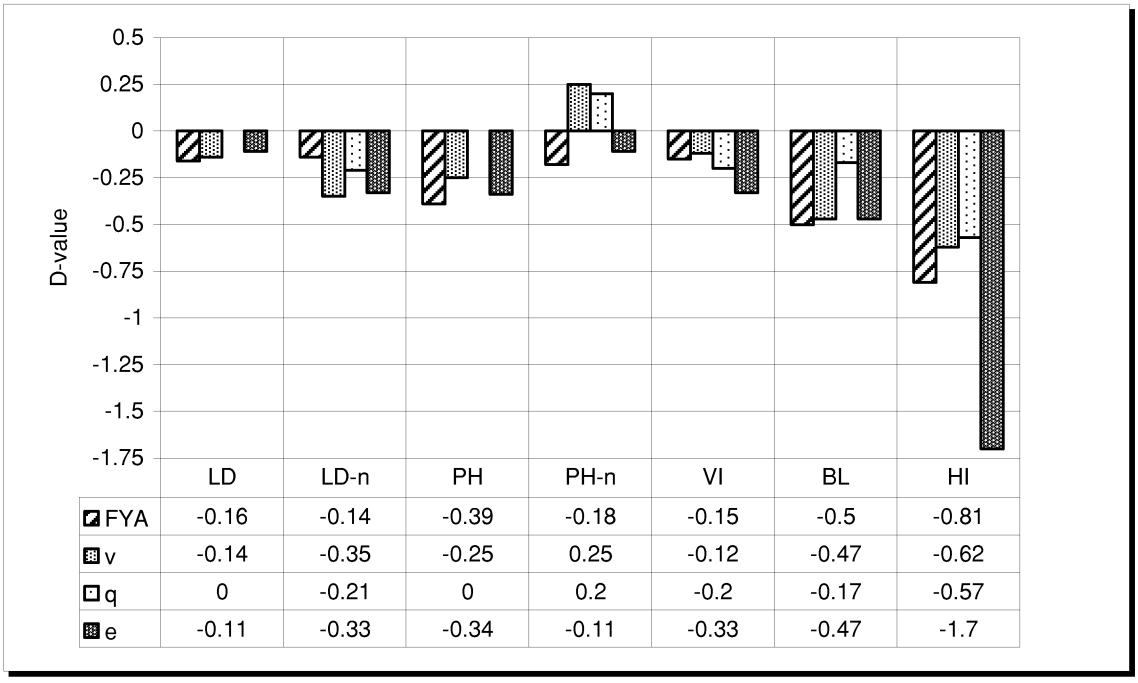
D-values of all predictors and criterion are presented in the following two figures (1 and 2) for all study groups.

Figure 1: D-values of main predictors and criterion



	LD	LD-n	PH	PH-n	VI	BL	HI
FYA	-0.16	-0.14	-0.39	-0.18	-0.15	-0.5	-0.81
C	-0.29	-0.28	-0.35	0.14	-0.2	-0.06	-0.59
B	-0.34	-0.09	-0.32	0.04	-0.09	0.33	0.06
P	-0.11	-0.38	-0.22	0.2	-0.24	-0.47	-1.19

Figure 2: D-values of PET subtests and criterion

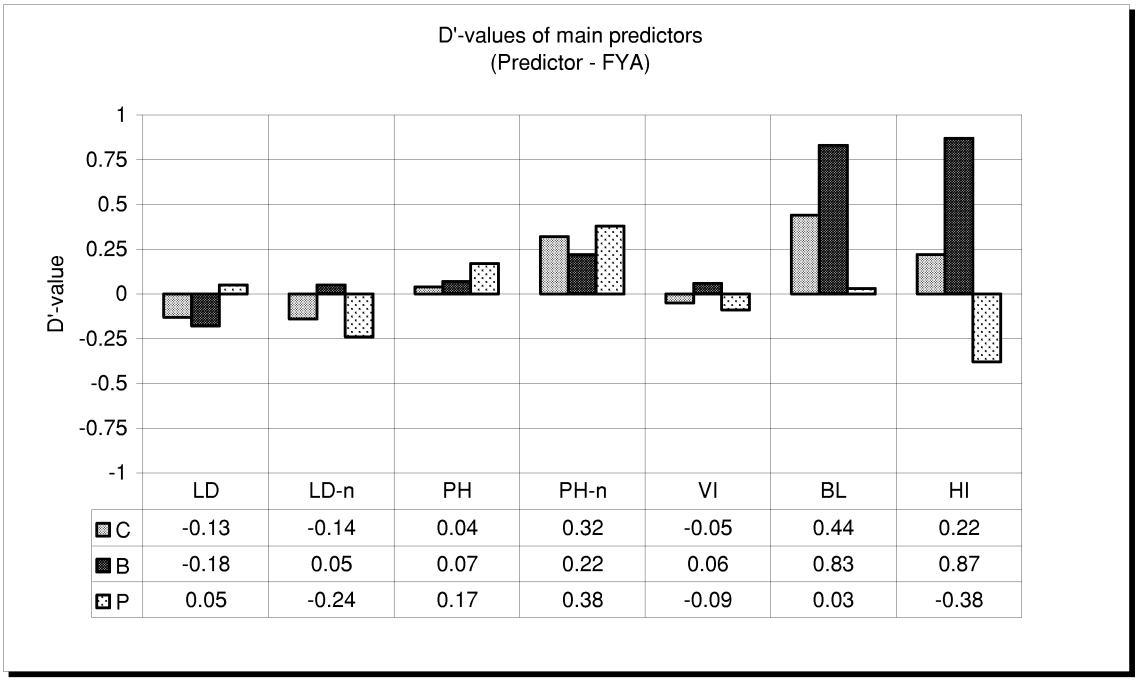


	LD	LD-n	PH	PH-n	VI	BL	HI
FYA	-0.16	-0.14	-0.39	-0.18	-0.15	-0.5	-0.81
v	-0.14	-0.35	-0.25	0.25	-0.12	-0.47	-0.62
q	0	-0.21	0	0.2	-0.2	-0.17	-0.57
e	-0.11	-0.33	-0.34	-0.11	-0.33	-0.47	-1.7

D-value analysis shows that examinees applying for accommodations (whether eligible or non-eligible) score, on average, lower than RS examinees, both on predictors and on FYA. (The PH-n students, whose predictors averages are slightly above RS's average, are an exception, as are the BL and HI groups, whose Bagrut is above RS's average – considerably so for the BL, and slightly so for the HI.) It seems that the differences in PET scores are mainly due to differences in V and E scores. This fact is in accordance with the greater proportion of LD students among those receiving accommodations, because their difficulties are mainly manifested in the verbal components.

The comparison of the predictors' deviation from the reference group (RS) with that of the criterion is made easier by the following two figures (3 and 4), where D'-values of all predictors are presented for all study groups:

Figure 3: D'-values of main predictors



	LD	LD-n	PH	PH-n	VI	BL	HI
C	-0.13	-0.14	0.04	0.32	-0.05	0.44	0.22
B	-0.18	0.05	0.07	0.22	0.06	0.83	0.87
P	0.05	-0.24	0.17	0.38	-0.09	0.03	-0.38

The D' (differences between deviations) analysis supplies indications for both the direction and the magnitude* of prediction bias:

*Effect size is verbally evaluated according to Cohen's (1988) rule of thumb: small - 0.2, medium - 0.5, large - 0.8.

LD – no bias for PET, slight under-prediction for Bagrut, resulting in slight under-prediction for the Composite score (0.13 SD).

LD-n – low under-prediction for PET (0.24 SD), no bias for Bagrut, resulting in slight under-prediction for the Composite score (0.14 SD).

PH – slight over-prediction for PET, no bias for Bagrut and Composite score.

PH-n – low-to-moderate over-prediction for all three predictors (0.22-0.38 SD).

And, reiterating the reservation about the very small n's, the results for following groups are:

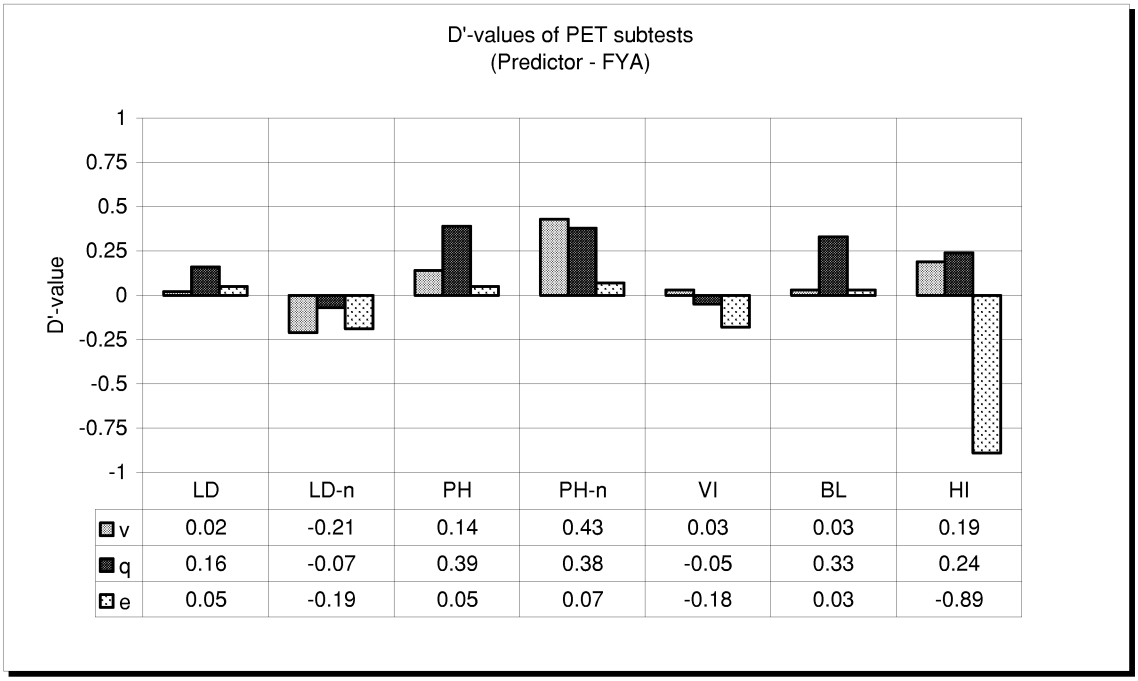
VI – no bias.

BL – no bias for PET, high over-prediction for Bagrut, resulting in a moderately over-predictive Composite score (0.44 SD).

HI – moderate under-prediction for PET, high over-prediction for Bagrut, resulting in a weakly over-predictive Composite score (0.22 SD).

A thorough look into PET's components is presented in Figure 4.

Figure 4: D'-values of PET subtests



	LD	LD-n	PH	PH-n	VI	BL	HI
v	0.02	-0.21	0.14	0.43	0.03	0.03	0.19
q	0.16	-0.07	0.39	0.38	-0.05	0.33	0.24
e	0.05	-0.19	0.05	0.07	-0.18	0.03	-0.89

It seems that the main reasons for these cases in which PET was found under-predictive, are the English section (LD-n, VI, HI), and also the Verbal section in LD-n. The

Quantitative section, in most cases, is slightly over-predictive, except in LD-n and VI, where it is not biased.

As described above, not everyone would agree with the definition of bias incorporated in the D-value analysis. Therefore, a more conservative treatment of the issue is the following two-regression analysis. Table 1 presents the average residual scores resulting from linear regressions of criterion on predictors, by focal group. All variables were standardized (0, 1) within units of analysis, and the averages are based on the reference group's (RS) prediction line calculated within units of analysis. (See Appendix 2 for graphical representations of the mean residuals and residual scores SD's).

Table 1: Mean residual (actual-predicted) scores*

	N	C	B	P	v	q	e
LD	821	-0.05	-0.05	-0.13	-0.13	**-0.16***	-0.15
LD-n	276	-0.03	-0.10	-0.04	-0.06	-0.07	-0.10
PH	159	-0.28	-0.30	-0.33	-0.32	**-0.39***	-0.35
PH-n	47	**-0.19***	**-0.16***	**-0.20***	**-0.20***	**-0.19***	-0.16
VI	90	-0.05	-0.10	-0.05	-0.09	-0.07	-0.08
BL	19	**-0.51***	**-0.62***	-0.41	-0.40	-0.49	-0.45
HI	28	-0.59	**-0.81***	-0.54	-0.69	-0.67	-0.63

The residual score analysis based on criterion-on-predictor regression shows that the direction of bias is seemingly in favor of all the groups. In other words, their actual first-year academic performance tends to be slightly to moderately over-predicted by all predictors. Since a tendency to suggest over-prediction for low-performing minorities is inherent to the regular (criterion on predictor) regression model, we consider only the results that satisfy the boundary condition method, namely, when the reverse (predictor-on-criterion) regression yields opposite-sign results evidence to non-artifact bias. (Opposite-sign results of reverse regressions are consistent!). This approach suggests that non-artifact bias in favor of the study groups (over-prediction) exists only in the following cases:

- All predictors in the PH-n group – a result of negligible importance, since no accommodations were supplied to these examinees, and also both the effect size and the n are small.

*residual (actual minus predicted) scores based on regular (criterion on predictors) regressions, where the inverse regressions (predictor on criterion) showed inverse-sign results, thus yielding consistent conclusions.

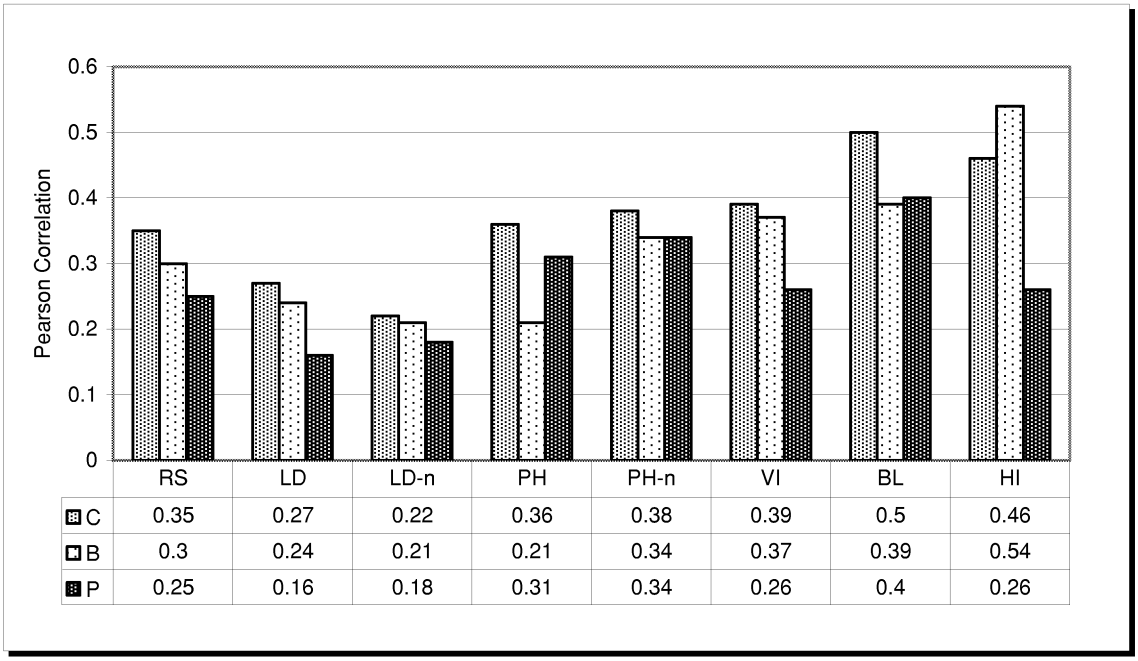
- Bagrut for groups BL and HI, (and consequently the Composite score for the BL).
- The Quantitative subtest for LD and PH groups. It seems that these groups suffer mainly from deterioration in their verbal abilities, and the compensation supplied by the accommodations is slightly excessive with respect to the Quantitative subtest.
- For the major groups of interest – LD, LD-n and PH – there is no conclusive evidence of predictive bias.

Prediction Accuracy

In order to examine the question of prediction accuracy, we calculated Pearson correlations between the predictors and the criterion. These correlations are not our estimates of the true predictive validities of the predictors: the latter are normally treated in a meta-analytic manner, which means – calculated within content-based homogeneous units of analysis, and corrected for possible statistical artifacts such as restriction of predictors' range, variation in criterion reliability, and sampling errors.

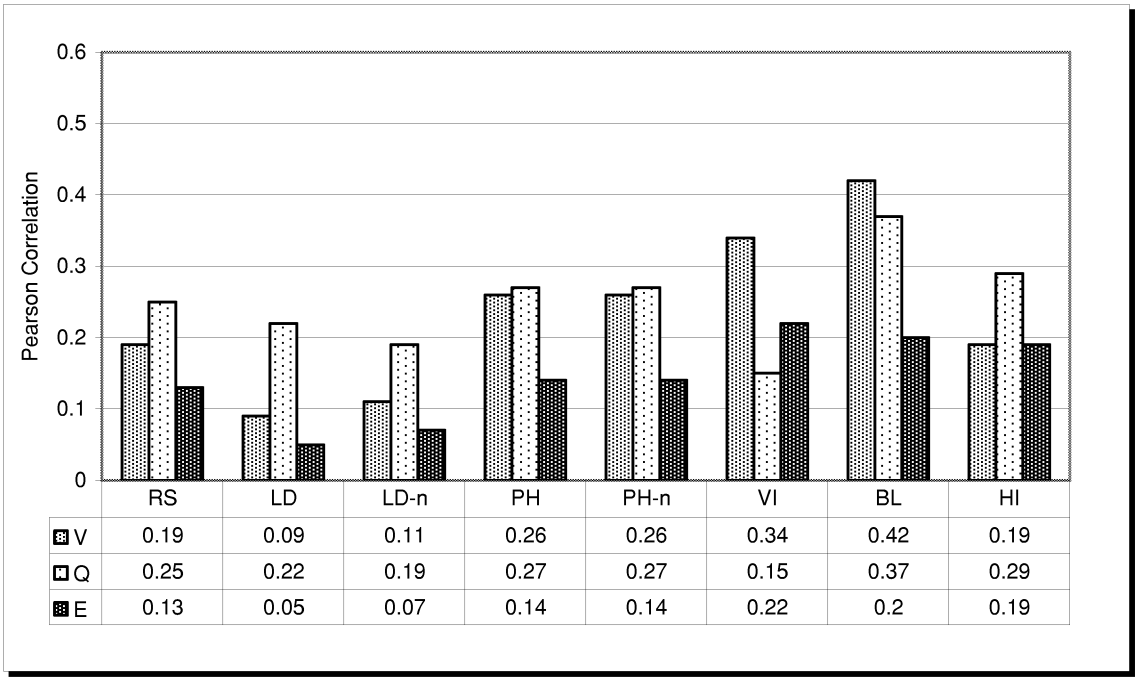
The limitations of the present study do not facilitate such treatment. Since, in each unit of analysis we can find only very few (sometimes one) students who were tested with accommodations, we had to pool examinees from various departments and calculate the correlations across study groups. Therefore, the presented correlations should be referred to only in a comparative manner – as relative to those of the RS group (calculated under same conditions). These correlations are presented in the next two figures.

Figure 6 – Correlations of main predictors with FYA among groups



	RS	LD	LD-n	PH	PH-n	VI	BL	HI
C	0.35	0.27	0.22	0.36	0.38	0.39	0.5	0.46
B	0.3	0.24	0.21	0.21	0.34	0.37	0.39	0.54
P	0.25	0.16	0.18	0.31	0.34	0.26	0.4	0.26

Figure 7 – Correlations of PET subtests with FYA among groups



	RS	LD	LD-n	PH	PH-n	VI	BL	HI
V	0.19	0.09	0.11	0.26	0.26	0.34	0.42	0.19
Q	0.25	0.22	0.19	0.27	0.27	0.15	0.37	0.29
E	0.13	0.05	0.07	0.14	0.14	0.22	0.2	0.19

The predictive accuracy of all the predictors (except for the Quantitative subtest in LD and LD-n examinees) is somewhat lower for the main focal groups than it is for RS examinees. The decline is manifest especially in the Verbal and English subtests. Among the other groups of disabilities (including non-eligible PH examinees – see Figure 6) the predictive validity is surprisingly high. Once again, however, it should be noted that these groups are too small to draw general conclusions. This whole pattern of relative correlations is not only seen in PET, but also in Bagrut. Beyond the possibility that these specific groups are too small and not representative enough, this may suggest that the difficulty in measuring disabled populations has to do with the criterion as well (problematic reliability), and not only the predictors. Among all groups (except HI), the validity of the Composite score is higher than that of each of its components separately (PET and Bagrut), meaning that each of them bears some differential additive value.

5. Conclusion

In Table 2 we summarize and compare the results of the two methods: "D" denoting the D-value analyses, and "BC" – the boundary conditions approach. For each cell (group*predictor) we indicate the direction and effect size of bias found by each method. Null cell means no-bias, plus sign indicates over-prediction (favoring the focal group), minus sign indicates bias against the focal group. The verbal quantification of effect-size is, again, derived from Cohen's (1988) rule of thumb: "small ~ 0.2, medium ~ 0.5, large ~ 0.8". The highlighted cells represent the cases in which both methods show bias.

Table 2: Two-method summary of bias results

Predictor	C		B		P		v		q		e	
Group	D	BC	D	BC	D	BC	D	BC	D	BC	D	BC
LD	-		-						+	+		
LD-n	-				-		-				-	
PH					+		+		++	+		
PH-n	+	+	+	+	+	+	++	+	++	+		
VI											-	
BL	++	++	+++	++					+			
HI	+		+++	+++	-		+		+		---	

Over-prediction: low"+", medium"++", large"+++"

Under-prediction: low"-", medium"--", large"---"

Do the results of the two analyses (D and BC) correspond? The answer is that there are no contradicting results, and the difference between the definitions lies mainly in the measure of leniency by which we recognize a difference between groups as bias. Whenever bias is found by BC, it is in agreement with D, both regarding the direction and the size of effect. These are the highlighted cells. In few extra cases D indicates bias which is not recognized as such by BC.

If we opt for methodological conservatism (BC), we can conclude that the whole selection system is *not* proved to be biased *against* persons tested with accommodations. In the case of two groups the selection system is slightly biased, in their favor: PH-n and BL. But the former were not tested with accommodations, and the latter is too small a group to generalize. Regardless, in the case of severely disabled people like Blind applicants, the system would prefer the positive direction of bias, namely in their favor.

If we adopt an "equal representation" definition of fairness in selection compatible with Thorndike (1971) and Darlington's 2nd definition (1971) of the term (D), which requires that in an unbiased selection system, the D-values between a focal group and the reference group be equal for both the predictor and the criterion (a hard-to-satisfy condition in itself, statistically speaking) – the results of our study differ a little. For the LD-n group, and also for the HI group – PET slightly under-predicts FYA. For the PH-n group, both approaches would agree that the selection system tends to be over-predictive, but again, since no accommodations were given to them – at least not in PET – this result is of little importance.

As to the accuracy of prediction aspect: The selection system seems to be less predictive for the both LD and LD-n groups than it is for the reference RS group. For the former – it is in accordance with previous findings in the literature, in spite of the accommodations, and for the latter – *because* (although causality is yet to be proven) of the denied accommodations. No decline in predictive accuracy is found for the rest of the groups.

A last remark should call attention to the LD-n group, whose FYA might be considered under-, and less accurately, predicted by non-accommodated PET. A possible implication of this study would be a re-consideration of the strictness of how LD eligibility is determined. But this should be stated with the following reservation: No information was available in this study about the accommodations this group might have enjoyed during their university studies. Therefore, the aforementioned effect might the result of potential gain in their criterion (FYA) level. Only control of this effect may imply whether and to what extent the criteria according to which LD candidates are found eligible for test accommodations need to be re-considered.

Summary of conclusions

Fairness of the prediction

- Both BC and D approaches – the whole selection system is *not* biased *against* persons tested with accommodations. Bagrut is over-predictive in BL, HI, PH-n groups.
- D approach – PET slightly *under-predictive* in LD-n and HI groups.

- The English subtest of PET seems to be responsible for under-prediction in the HI and the LD-n groups, while the Quantitative subtest is over-predictive for most groups, suggesting over-compensation where the impaired skills are mainly verbal.

Accuracy of the prediction

- The selection system is somewhat less predictive for both the Learning Disability and non-eligible Learning Disability groups than it is for the reference Regular Students group.
- The decline is manifested especially in the Verbal and English subtests, less so in the Quantitative
- Among other groups of disabilities, the predictive validity was surprisingly high, but these groups were relatively small in size.
- The resemblance of validity profiles between PET and Bagrut suggests that the criterion's reliability might account, at least partly, for the decreased prediction accuracy in groups of students with disabilities
- Among all groups, the validity of the Composite score is higher than that of each of its components taken separately (psychometric score and matriculation average), meaning that each has differential additive value.

General implications

- In a future research, after controlling for possible accommodations that the LD-n group might have been granted during FY studies, the strictness and criteria according to which LD candidates are found eligible for test accommodations should be reconsidered. (The LD-n group appears slightly under-predicted, and less accurately predicted, by non-accommodated PET.)
- Consider some fine tuning regarding the PET adaptations provided: more generosity with the English section, less with the Quantitative.

Appendix 1

Simple statistics for raw data across population.

Table 3: Means of study variables by groups

	N	FYA	C	B	P	v	q	e
RS	*63,291*	81	53.6	93.1	603.1	118.6	117.3	118.7
LD	*821*	79.2	51.4	90.7	598.4	116.7	117.5	117.7
LD-n	*276*	79	50.6	91.9	576.7	113.9	114	113
PH	*159*	77.9	52.3	92	598.7	116.8	118.6	115.7
PH-n	*47*	79.5	55.7	94.7	623.1	122.5	123	117.2
VI	*90*	81.6	51.9	93.6	588.5	118	113.2	115.2
BL	*19*	79.2	50.6	94.3	549.9	109.7	108.8	106.2
HI	*28*	75.9	51.6	96.2	563.8	114	114.9	98

Table 4: SD's of study variables by groups

	FYA	C	B	P	v	q	e
RS	9.4	10	9.3	84.8	15.8	17.9	19.5
LD	10.4	9.3	8.9	82	16.1	17.9	18.4
LD-n	10.7	9.5	8.2	82.5	15.8	17.9	20
PH	10.5	10.4	8.9	91	17	19.5	19.6
PH-n	8.4	11.5	10.3	79.1	15.8	17	16.8
VI	7.9	10.9	8.8	99	17.3	20.8	18.6
BL	12.7	9.8	9.3	102	16.5	19.9	22.8
HI	7.2	11.6	11	99.7	19	21.4	21.3

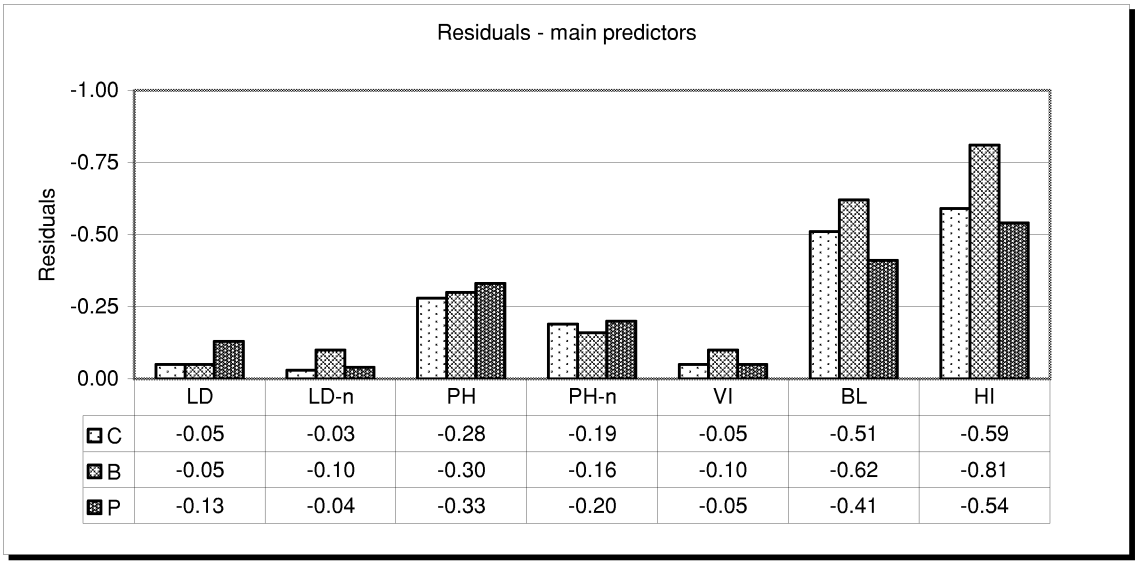
Appendix 2

Table 5: Std. Dev. of residuals

	C	B	P	v	q	e
LD	1	1.1	1.1	1.1	1.1	1.1
LD-n	1	1	1	1	1	1
PH	1.1	1.2	1.2	1.2	1.2	1.2
PH-n	0.9	0.9	0.9	0.9	0.9	0.9
VI	0.9	0.9	0.9	0.9	0.9	0.9
BL	1.4	1.6	1.5	1.5	1.5	1.6
HI	1	0.9	1.1	1.1	1	1.1

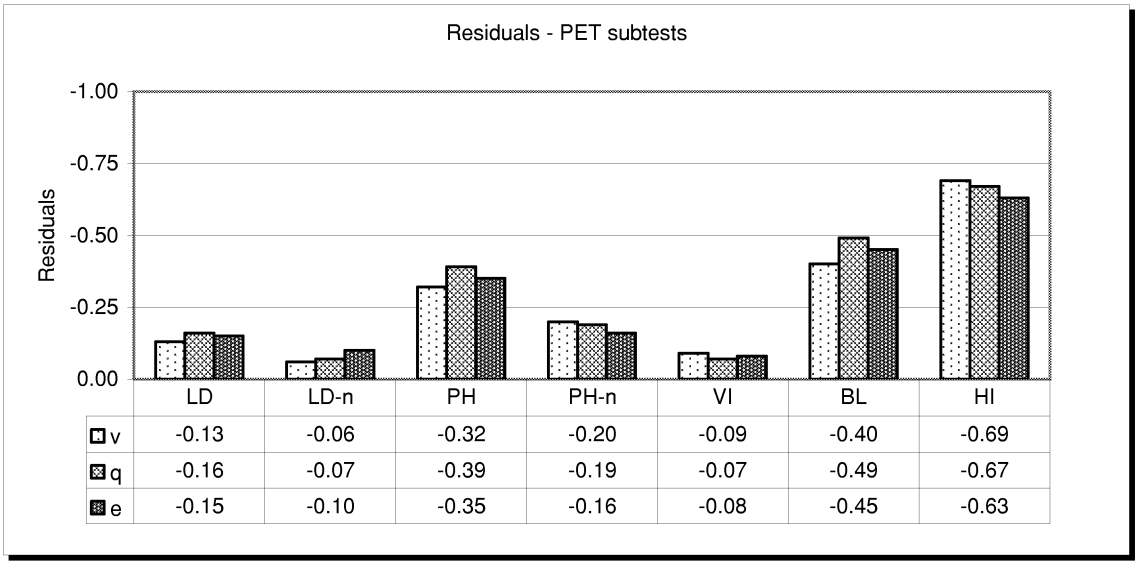
The values of **Table 1** are presented in the next two figures. Notice that the vertical axis (mean residuals, in Z terms) is inverse, since all values are negative

Figure 8: Mean residual scores of main predictors



	LD	LD-n	PH	PH-n	VI	BL	HI
C	-0.05	-0.03	-0.28	-0.19	-0.05	-0.51	-0.59
B	-0.05	-0.10	-0.30	-0.16	-0.10	-0.62	-0.81
P	-0.13	-0.04	-0.33	-0.20	-0.05	-0.41	-0.54

Figure 9: Mean residual scores of PET subtests



	LD	LD-n	PH	PH-n	VI	BL	HI
v	-0.13	-0.06	-0.32	-0.20	-0.09	-0.40	-0.69
q	-0.16	-0.07	-0.39	-0.19	-0.07	-0.49	-0.67
e	-0.15	-0.10	-0.35	-0.16	-0.08	-0.45	-0.63

6. References

Aaron, P.G. (1997). The impending demise of the discrepancy formula. *Review of Educational Research, 67*, 461-502.

Bennet, R.E., Rock, D.A., & Jirele, T. (1986). *The psychometric characteristics of the GRE General Test for three handicapped groups* (ETS Research Report RR-86-6). Princeton, NJ: Educational Testing Service.

Birnbaum, M.H. (1979). Procedures for detection and correction of salary inequity. In T.R. Pezzullo & B.F. Birtingham (Eds.), *Salary equity* (pp. 121-144). Lexington, MA: Lexington Books.

Birnbaum, M.H. (1981). Reply to McLaughlin: Proper path models for theoretical partialling. *American Psychologist, 36*, 1193-1195.

Braun, H., Ragosta, M., & Kaplan, B. (1986). *The predictive validity of the Scholastic Aptitude Test for disabled students* (Research Report 86-38). New York: College Entrance Examination Board.

Camara, W.F. (1998). *Effects of Extended time on score growth for students with learning disabilities*. New York: College Board.

Centra, J.A. (1986). Handicapped student performance on the Scholastic Aptitude Test. Journal *of Learning Disabilities, 19*, 324-327.

Cisero, C.A., Royer, J.M., Merchant, H.G. & Jackson, S.J. (1997). Can the Computer-based Academic Assessment System (CAAS) be used to diagnose reading disability in college students? *Journal of Educational Psychology, 89*, 599-620.

Cleary, A. (1968). Test bias: Prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (pp. 24-25).* New York: Lawrence Erlbaum associates.

Darlington, R.B., (1971). Another Look at "Culture Fairness". *Journal of Educational Measurement, 8*, pp. 71-81.

Fuchs, L.S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice*, 13, 204-219.

Gersham, F.M., MacMillan, D.L., & Bocian, K.M. (1996). Learning disabilities, low achievement, and mild mental retardation: More alike than different? *Journal of Learning Disabilities, 29*, 570-581.

Laing, J., & Farmer, M. (1984). Use of the ACT assessment by examinees with disabilities (Research Report No. 84). Iowa City, IA: American College Testing.

Linn, R. L. (1984). Selection bias: Multiple meanings. *Journal of Educational Measurement, 21*, 33-47.

Linn, R. L. (1990). Admissions testing: recommended uses, validity, differential prediction, and coaching. *Applied Measurement in Education*, *3(4)*, 297-318.

MacArthur, C.A. (1996). Using technology to enhance the writing processes of students with learning disabilities. *Journal of Learning Disabilities, 29,* 344-354.

Phillips, S.E. (1994). High stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7*, 93-120.

Pitoniak, M.J., & Royer, J.M. (2001). Testing accommodations for examinees with disabilities: A review of psychometric, legal and social policy issues. *Review of Educational Research, 71*(1), 53-104.

Thorndike, R. L (1971). Concepts of culture fairness. *Journal of Educational Measurement, 8*, 63-70.

Tindal, G., & Fuchs, L. (1999). *A summary of research of test accommodations: What we know so far*. Mid-South Regional Resource Center, University of Kentucky.

Wightman, L.F. (1993). *Test takers with disabilities: A summary of data from special administrations of the LSAT* (Research Report 93-03). Newtown, PA: Law Admission Council.

Willingham, W.W., Ragosta, M., Elliot Bennett, R., Braun, H., Rock, D.A., & Powers, D.E. (1988). *Testing Handicapped People*. Allyn and Bacon, MA.

Ziomek, R.L. (1996). *Achievement score gains of special-needs tested students we tested at least twice on the ACT assessment*. Paper presented at the meeting of the National Association for College Admission Counseling, Minneapolis, MN.

Ziomek, R.L., & Andrews, K.M. (1996). *Predicting the college grade point averages of special-tested students from their ACT assessment scores and high school grades* (ACT Research Report Series, 96-7). Iowa City, IA: American College Testing.

Zurcher, R., & Pedrotty Bryant, D. (2001). The validity and Comparability of Entrance Examination Scores after Accommodations are made for Students with LD. *Journal of Learning Disabilities*, 34(5), 462-471.