Examining the Fairness of Higher Education Admissions for Applicants Who Request Test Accommodations

Dvir Kleper, Elliot Turvall, Tamar Kennet-Cohen, and Carmel Oren
dvir@nite.org.il; elliot@nite.org.il;tami@nite.org.il;carmel@nite.org.il

National Institute for Testing and Evaluation

Abstract

The present study examines the fairness of higher education admissions practices with respect to three different groups of people granted accommodations for the Psychometric Entrance Test (PET) in Israel: those with Learning Disabilities (LD, N=958), those with Attention Deficit Hyperactivity Disorder (ADHD, N=187), and those with Physical Disabilities (PH, N=1,096). In addition, we examined the fairness of the admissions process with respect to individuals whose requests for accommodations were denied, either on technical grounds (MD, N=299) or because of professional considerations (DP, N=1,458).

Since the focal groups were very small relative to the reference group to which they were being compared (RS, N=120,503), we first used propensity scores with respect to the covariates of gender and age to match individuals from the reference group to individuals in the various focal groups.

Fairness can be measured in two ways: the first is bias in selection, and the second is predictive accuracy. We applied Cleary's model in order to detect possible biases and used Pearson correlation coefficients between the predictor and the criterion to determine predictive accuracy.

The results showed no bias with respect to the LD, ADHD and the DP groups, a small bias in favor of the PH group (on all predictors), and a very small under-prediction for the MD group (on all predictors except the Bagrut). Comparing the validity coefficients of the focal group with that of the reference group, we see that the results are quite similar for the PET and somewhat higher for the reference groups on the Bagrut and the Composite Score (except for the MD group). This is especially true for the LD and ADHD groups.

Key Words: Fairness, Cleary, Propensity, Accommodations

Introduction

The present study investigates the fairness of higher education admissions practices for applicants who request test accommodations in order to compensate for various disabilities. In particular, this study relates not only to those who were granted accommodations, but also to those who were denied accommodations, either on technical grounds or because of professional considerations.

Special accommodations for examinees with disabilities have become common practice in most large-scale, high-stakes testing programs (Fletcher, Francis, O'Malley, Copeland, Mehta, Caldwell, Kalinowski, Young & Vaughn, 2009; Gregg, Coleman, Davis & Chalk, 2007; Lai & Berkeley, 2012; Lindstrom, 2007; Solórzano, 2008). This evolving process has resulted in an increase in the proportion of learning disabled students in higher education.

Accommodations are modifications to standard evaluation measures, aimed at "leveling the playing field" for students with disabilities by reducing irrelevant variability. Scores obtained by disabled examinees on the basis of valid accommodations should be comparable to those obtained by non-disabled examinees under standard conditions (Tindal & Fuchs, 1999).

Previous research studies have examined the fairness of higher education admissions. Turvall, Bronner, Kennet-Cohen & Oren (2008), for example, examined whether the higher education admissions process in Israel discriminates against the Arab population. They found that even though the Arab sector scored much lower on the Psychometric Entrance Test (PET), the test over-predicted first-year averages (FYA) in most academic departments.

Another study (Oren & Even, 2005) assessed the fairness of selection with regard to learning disabled and physically disabled examinees and denied accommodations for learning disabilities. The study found that the PET is fair with respect to the learning disabled group, but yields a slight over-prediction for the physically disabled group and a slight under-prediction for the denied group. However, this research into the effects of accommodations on fairness was hindered by the fact that it was difficult to locate sufficiently large groups of disabled examinees to comprise meaningful units of analysis. In the present study, the propensity score matching technique (Rosenbaum & Rubin, 1983) was employed as a means of compensating for this limitation. This technique calculates a single scalar number, called the propensity score, for each individual using age and gender covariates. An algorithm is then applied to match the calculated propensity scores of the group of examinees who requested accommodations with those of a subgroup of regular examinees (who did not apply for accommodations).

Fairness in selection is an issue wherein measurement and value-dependent considerations converge (Camilli, 2006). In keeping with the prevailing view that any assessment of fairness should take into account the criterion that the test is designed to predict (in the context of university admissions, the first-year GPA), we applied Cleary's (1968) approach to detecting possible selection biases. Cleary's model of test bias, which is recommended by professional associations (AERA/APA/NCME, 1999), is by far the most common method of defining and detecting bias in selection (Young, 2001). According to Cleary's model, a test is biased if criterion scores predicted from a common regression line tend to be too high or too low for a particular group.

Another aspect of criterion prediction is the validity coefficient, i.e., how well test scores are associated with the criterion measure. We used this coefficient to compare validity between the reference group and the focal group, referring to the result as differential validity. Research on differential validity has shown that scores obtained with accommodations generally have lower correlations with the criterion being measured than those obtained without accommodations (Cahalan, Mandinach & Camara, 2002; Zurcher & Bryant, 2001; Ziomek & Andrews, 1996; Braun, Ragosta & Kaplan, 1986; Laing & Farmer, 1984).

The present study examines the fairness of the admissions process used by Israeli universities, which is based on the Composite Score, a combination of the Psychometric Entrance Test (PET) score and the high school matriculation exam scores (Bagrut).

Method

Sample

The study population consists of students who began their studies at one of six Israeli universities in the academic years 2002/03 through 2008/09, and who took the PET in Hebrew, after July 2000.

Accommodations were given to examinees with physical or sensory disabilities (PH), as well as to examinees with learning disabilities. We further divided the learning disabilities group into two distinct subgroups: those with pure Attention Deficit Hyperactivity Disorder (ADHD) and those with other learning disabilities (reading, writing or mathematical disabilities – LD).

In addition to those who were granted accommodations, it is also interesting to examine fairness with regard to those examinees that applied for, but did not receive, accommodations. We divided the examinees whose applications for accommodations were rejected into two subgroups: those who were denied for technical reasons (i.e., failure to provide all necessary documentation or to meet application deadlines – MD), and those whom the professional staff deemed ineligible (DP).

Eligibility for PET accommodations is determined by the Special Test Accommodations Unit at the National Institute for Testing and Evaluation. The unit's professional staff consists of experts in learning disabilities who review each application and decide whether test accommodations are warranted. Accommodations are granted to individuals with a documented primary learning disability (in reading, writing, math, or attention deficit disorder) or a documented physical disability. The staff aims to minimize the effect of the disability, while preserving the general accuracy and validity of PET scores. Possible accommodations include time extensions, periodic rest breaks, reading or writing facilitators, etc.

The original data set, which comprised 124,501 records of first-year students in 2,036 academic departments at six Israeli universities, was, for the purposes of this study, divided into six categories: RS (Regular Students), LD (Learning Disabilities), ADHD, PH (Physical Disabilities), MD (Missing Documents), and DP (Denied Professional).

The number of examinees in each group is presented in Table 1.

Table 1
Sample Sizes

Group	Frequency	Percentage	
RS	120,503	96.79	**did not request accommodations**
LD	958	0.77	**requested and received accommodations**
ADHD	187	0.15	
PH	1,096	0.88	
MD	299	0.24	**requested accommodations and were denied**
DP	1,458	1.17	

Predictors

***Bagrut - High school matriculation certificate (B) score*.** In Israel, most high school graduates receive a matriculation certificate, with grades for various general high school subjects. These grades are based on a combination of high school grades and scores on national tests. The Bagrut score is a weighted average of the subject scores. Scale: 40 to 120 (100 plus various bonuses for enhanced test levels).

***Psychometric Entrance Test (PET) total score*.** The PET is designed to measure various cognitive and scholastic abilities with the goal of serving as a good predictor of success in future studies. It includes three multiple-choice subtests – verbal reasoning, quantitative reasoning, and English as a foreign language. Scale: 200 to 800, historic mean of 500, SD=100.

Verbal Reasoning (V). This section of the PET includes 60 items that focus on the verbal skills and abilities required for academic success: analysis and comprehension of complex written material, systematic and logical thinking, and the ability to draw fine distinctions between the meaning of words and concepts. The Verbal section includes analogies, critical reading and inference questions, and reading comprehension.

Quantitative Reasoning (Q). This section of the PET includes 50 items that focus on the use of numbers and mathematical concepts (algebraic and geometric) in solving quantitative problems and analyzing information presented in graphs, tables, and charts. The level of mathematics is basic – equivalent to that acquired in the ninth or tenth grades in most Israeli high schools. Formulas and explanations of mathematical terms that may be needed for the test are provided.

English as a foreign language (E). This section of the PET consists of 54-58 items designed to assess ability to comprehend academic-level texts in English. This section includes three types of items: sentence completions, restatements, and reading comprehension questions. This subtest serves a dual purpose: it is a component of the PET total score, and it is also used by each institution of higher education to place students in remedial English classes.

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

Composite Score (C). This is generally an equally weighted average of the PET score and Bagrut score. Scale: mean=50, SD=10.

Criterion

First-Year Grade Average (FYA). Grade point average from the first year of university studies. Scale: 40 to 100.

Data Analysis

Each of the five groups (LD, ADHD, PH, MD and DP), was deemed as a focal group and compared separately to the reference group (RS).

For each of the five comparisons, a subgroup of the RS group was chosen on the basis of propensity score matching, which matches each individual in the focal group to a "similar" individual from his/her unit of analysis (defined as an academic department within an academic institution and academic year), according to age and gender covariates. We chose these variables because they were available for virtually everyone in the sample. Other background variables, such as parents' education or socioeconomic status had too many missing values and were hence not feasible.

Propensity Score Matching Method (Adapted from D'Agostino, 1998)

In a randomized experiment, the randomization of units (that is, participants) to different treatment conditions guarantees that, on average, there should be no systematic differences in observed or unobserved covariates (that is, bias) between units assigned to the different treatments. However, in a non-randomized observational study in which investigators have no control over the treatment assignment and direct comparison of outcomes from the treatment groups may be misleading. This difficulty can be mitigated to some degree if information on

measured covariates is incorporated into the study design. Traditional methods of adjustment are often limited since they can use only a limited number of covariates for adjustment. However, propensity scores, which provide a scalar summary of the covariate information, do not have this restriction.

Intuitively, the propensity score is a measure of the likelihood that a participant would have been treated using only their covariate scores.

Formally, the propensity score (Rosenbaum & Rubin, 1983) for subject i is the conditional probability of assignment to a particular treatment ($T_i = 1$) versus control ($T_i = 0$), given a vector of observed covariates, $\vec{x}_i$:

$$propen(\vec{x}_i) = Prob\left(T_i = 1 \mid \vec{X}_i = \vec{x}_i\right) \quad (1)$$

where it is assumed that, given the X's, the T_i's are independent:

$$Prob(T_1 = t_1, \ldots, T_N = t_N \mid \vec{X}_1 = \vec{x}_1, \ldots, \vec{X}_N = \vec{x}_N) = \prod_{i=1}^{N} propen(\vec{x}_i)^{t_i} \cdot (1 - propen(\vec{x}_i))^{1-t_i} \quad (2)$$

This definition implies that the *T* and *X* are conditionally independent given $propen(x)$. Thus, individuals in treatment and control group with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distribution on their background covariates.

When covariates contain no missing data, the propensity scores can be estimated using logistic regression.

The results of the process of matching using the propensity scores are presented in the following table.

Table 3

Comparison of Covariates Before and After the Matching

Groups	Before matching*					After matching				
	N	Gender		Age		N	Gender		Age	
		Female Percent	Male Percent	Mean	STD		Female Percent	Male Percent	Mean	STD
RS	59,283**	60	40	21.62	2.53	913	51	49	22.51	2.09
LD	913	51	49	22.58	2.07	913	51	49	22.58	2.07
RS	23,571**	60	40	21.64	2.59	182	39	61	22.86	2.16
ADHD	182	39	61	22.91	2.32	182	39	61	22.91	2.32
RS	64,930**	59	41	21.51	2.51	1,046	47	53	21.96	2.07
PH	1,046	47	53	22.02	2.32	1,046	47	53	22.02	2.32
RS	32,586**	63	37	21.63	2.67	286	57	43	21.91	2.59
MD	286	58	42	21.92	2.36	286	58	42	21.92	2.36
RS	70,615**	59	41	21.56	2.49	1,398	57	43	22.03	2.08
DP	1,398	57	43	22.06	2.05	1,398	57	43	22.06	2.05

* Only participants with complete data for the variables (gender and age) are presented

** Only participants from units of analysis including at least one participant with a "disability" found in the focal group are presented

The above data demonstrate that the procedure resulted in successful matching with respect to the covariates.

We conducted the following analyses for each of the six predictors.

Cleary's Model

Cleary's model was applied to each pair of matched groups.

According to this model there is bias in selection when the use of a common regression line for predicting the criterion by the predictor results in over-prediction or under-prediction for the criterion.

In applying this model, a single regression equation (i.e. across both groups) for predicting the criterion was estimated. For each observation, the earned FYA was subtracted from the predicted FYA. The residuals were averaged. Positive (residual) values indicate over-prediction for the focal group (bias in favor), and negative values indicate under-prediction (bias against).

The averaged residuals were computed separately for each unit of analysis, and then averaged across units of analysis by weighting the number of students in each unit of analysis.

Differential Validity

Pearson correlation coefficients between the predictor variables and the criterion variable (FYA) were calculated separately for each of the two matched groups.

Since, for the focal group, each unit of analysis had very few students (typically only one), we had to pool students together and hence, contrary to Cleary's measure of bias, this coefficient was calculated across all units of analysis. Therefore, before calculating the coefficient, all the predictors and the criterion were standardized within each unit of analysis.

It is important to emphasize that no correction for range restriction was made, mainly because of an absence of shrinkage data on the focal groups. This examination serves to compare the groups on predictive validity and not to assess the "true" validity of the selection system.

Results

Descriptive Statistics

Means and standard deviations of the criterion and the predictors were computed for each group (raw variables). Table 2 relates to the original setup described in Table 1.

Table 2
Means and Standard Deviations of the Criterion and the Predictors

Group	Variable													
	FYA		B		PET		C		V		Q		E	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RS	82.2	9.6	99.0	8.8	630.5	85.0	59.1	9.4	122.4	16.4	122.6	17.0	124.6	19.0
LD	80.7	9.8	94.2	7.9	593.2	83.8	54.1	8.7	116.8	15.7	116.6	20.4	114.7	21.0
ADHD	80.8	10.6	94.3	7.3	613.7	83.0	55.2	8.5	120.1	14.8	119.5	18.1	120.7	20.4
PH	80.0	10.7	98.6	8.3	632.2	87.0	58.8	9.4	123.6	16.1	122.9	17.6	122.9	20.6
MD	81.8	9.3	94.3	8.9	547.9	97.3	51.4	10.2	107.0	18.3	109.4	19.7	109.1	22.9
DP	81.8	9.8	96.1	8.6	585.9	87.6	54.6	9.4	114.4	16.7	115.5	17.7	115.3	20.9

Bias in Selection

Table 4 presents Cleary's measure of selection bias.

Table 4
Cleary's Measure of Selection Bias[*] with Respect to the Focal Group

	Variable					
Group	B	PET	C	V	Q	E
LD	0.03	0.02	0.01	0.04	0.05	0.04
ADHD	0.02	0.01	0.00	0.03	0.02	0.03
PH	0.15	0.14	0.14	0.16	0.14	0.14
MD	-0.03	-0.10	-0.09	-0.08	-0.06	-0.05
DP	0.03	-0.02	-0.02	0.00	0.01	0.02

[*] In terms of standard deviations of FYA; a positive value indicates bias in favor of the focal group and a negative value indicates bias against it.

Predictive Accuracy

Table 5 presents validity coefficients.

TABLE 5
Validity Coefficient[*]

Group	Variable											
	B		PET		C		V		Q		E	
	f[**]	r[**]	f	r	f	r	f	r	f	r	f	r
LD	0.23	0.38	0.30	0.27	0.33	0.40	0.28	0.23	0.25	0.24	0.08	0.20
ADHD	0.24	0.43	0.29	0.33	0.33	0.48	0.14	0.29	0.26	0.36	0.27	0.12
PH	0.30	0.34	0.27	0.23	0.37	0.37	0.23	0.18	0.28	0.21	0.12	0.14
MD	0.33	0.27	0.26	0.28	0.36	0.34	0.25	0.28	0.23	0.23	0.12	0.15
DP	0.28	0.33	0.25	0.25	0.35	0.37	0.23	0.20	0.24	0.25	0.09	0.15

[*] No correction for range restriction was made.
[**] f=focal, r=reference

Discussion

Prediction Bias

We analyzed the results in Table 4 according to Cohen's rule of thumb: small ~0.2, medium ~0.5, large ~0.8 (Cohen, 1998). This table shows that no bias was found with respect to the LD, ADHD and DP groups, while there was a small bias in favor of the PH group (on all predictors) and a very small bias against the MD group (on all main predictors except the Bagrut).

The PH group for which the results show a small positive bias consists of individuals with problems like diabetes, hyperhidrosis and other conditions that do not necessarily affect test performance itself. For these individuals, accommodations probably over-compensate for their disabilities, and hence lead to over-prediction. However, this group also includes individuals with severe physical disabilities, for whom the higher education selection system would prefer the favorable bias.

The professional staff that reviews applications for accommodations requests additional documentation only from applicants who have learning disabilities. Therefore, it is reasonable to assume that the members of the MD group, for whom the results show a very small negative bias, would probably have received accommodations, had they submitted all necessary documentation on time. This bias is hence anticipated.

Finally, with regard to the MD and the DP groups, no information was available concerning what accommodations they might have received during their university studies. Therefore, the aforementioned effect may be the result of potential gain in their criterion (FYA) level. Only by controlling for this effect could one determine whether, and to what extent, the criteria for establishing eligibility for accommodations should be reconsidered.

Prediction Validity

It is worth noting once again that in each unit of analysis, we could find very few students (sometime only one) who were tested with accommodations. Therefore, we had to pool examinees from various departments and calculate the correlations across study groups. This limitation prevented us from correcting for statistical artifacts such as restriction of predictor range, variation in criterion reliability, and sampling errors. Hence, the presented correlations should be considered only in relation to those of the RS group (which were calculated under the same conditions).

For the ease of the interpretation, Figures 1 and 2 present the results from Table 5 in graph form.

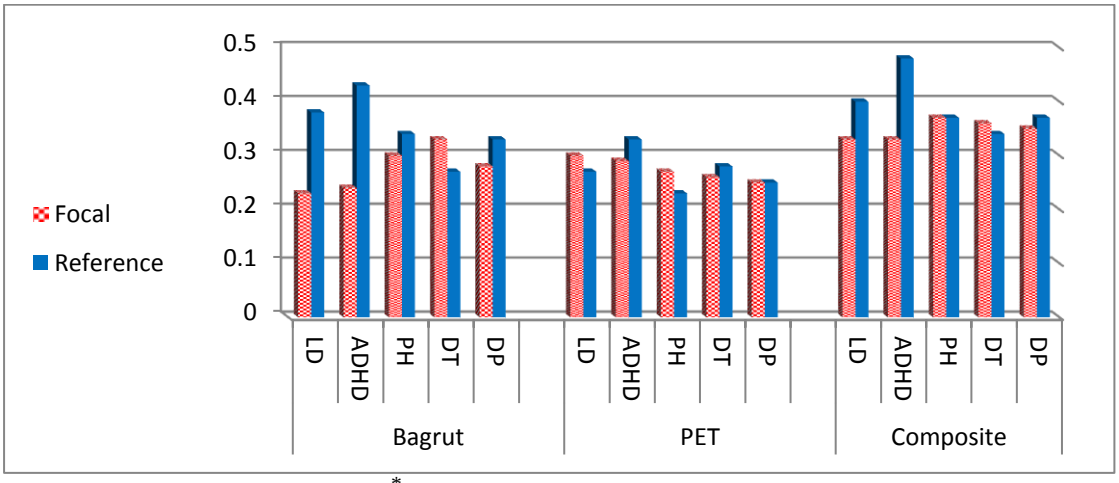

Figure 1. *Validity Coefficients*[*] *- Primary Predictors*
[*] No correction for range restriction was made.

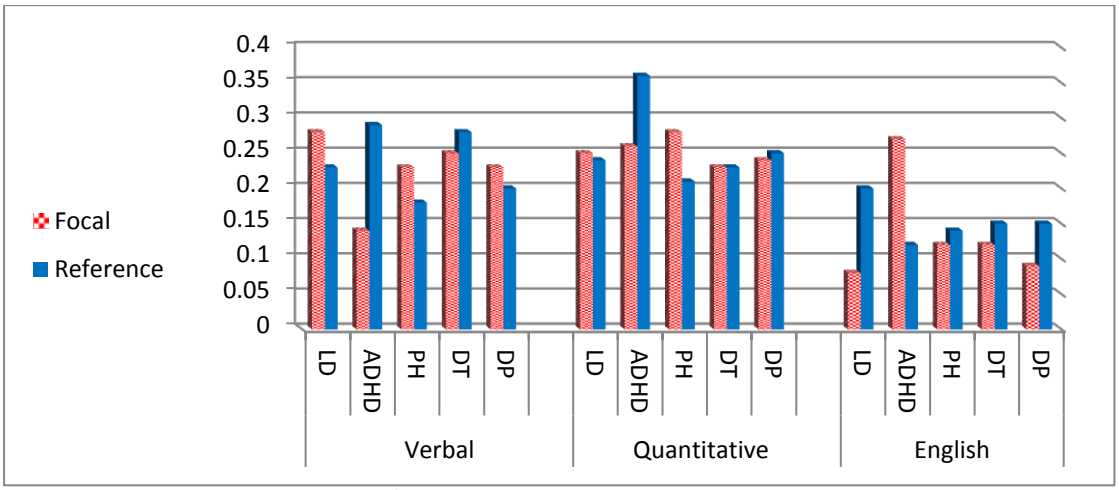

Figure 2. *Validity Coefficients*[*] *- Secondary Predictors*
[*] No correction for range restriction was made.

As expected, Figure 1 demonstrates that the validity of the Composite Score is higher than that of each of its individual components (PET and Bagrut), indicating that each of these components bears some differential additive value (which is the case for both the focal group and the reference group). As expected, this is also the case for PET and its three components (Verbal, Quantitative and English).

Comparison of the validity of the focal group with that of the reference group shows that the predictive validity of the Bagrut and the Composite Score is somewhat higher for the reference groups (except for the MD group). This is manifest especially in the LD and ADHD groups. The picture is less consistent for the components of the PET. For some groups and predictors (LD and PH on Verbal and Quantity subtests), the results are higher for the focal group, while for others (ADHD on Verbal and Quantitative), there is a slight advantage for the reference group.

The results show that the English subset bears the lowest predictive validity compared to the other predictors and, with the exception of the ADHD focal group, the reference group always has higher predictive validity.

In general, the pattern of results obtained in this study with respect to the question of fairness toward applicants who requested test accommodations resembles the findings reported in a previous study conducted by the National Institute for Testing and Evaluation, which relied on the traditional method (Oren & Even, 2005).

Acknowledgments

The authors wish to thank Naomi Gafni and Noa Saka from the National Institute for Testing and Evaluation for their assistance with the data analyses and for their many insightful comments and suggestions during the preparation of this manuscript.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Braun, H., Ragosta, M. & Kaplan B. (1986). *The Predictive Validity of the Scholastic Aptitude Test for Disabled Students* (ETS Research Report No. 86-38). Princeton, N.J.: Educational Testing Service.

Cahalan, C., Mandinach, E. B. & Camara, W. J. (2002). *Predictive Validity of SAT I: Reasoning Test for Test Takers with Learning Disabilities and Extended Time Accommodations* (College Board Research Report No. 2002-5). New York: College Entrance Examination Board.

Camilli, J. (2006). Test Fairness. In R. L. Brennan (ed.), *Educational Measurement* (4th ed., pp. 221-256). Westport: American Council on Education & Praeger.

Cleary, T. A. (1968). Test Bias: Prediction of Grades of Negro and White Students in Integrated Colleges. *Journal of Educational Measurement,* 5, 115-124.

Cohen, J. (1998). *Statistical Power Analysis for the Behavioral Sciences* (pp. 24-25). New York: Lawrence Erlbaum Associates.

D'Agostino, R. B. (1998). Tutorial in Biostatistics Propensity Score Methods for Bias

Reduction in the Comparison of a Treatment to a Non-Randomized Control Group. *Statist. Med* 17, 2265-2281.

Fletcher, J. M., Francis, D.J., O'Malley, K., Copeland, K., Mehta, P., Caldwell, C., Kalinowski, S., Young, V., & Vaughn, S.R. (2009). Effects of a Bundled Accommodations Package on High Stakes Testing for Middle School Students with Reading Disabilities. *Exceptional Children,* 75, 412-428.

Gregg, N., Coleman, C., Davis, M., &Chalk, J. (2007). Timed Essay Writing-Implications for High Stakes Tests. *Journal of Learning Disabilities,* 40, 306-318.

Lai, A. S. & Berkeley, S. (2012). High-Stakes Test Accommodations: Research and Practice, *Learning Disability Quarterly,* 35(3), August 2012, 158-169.

Laing, J. & Farmer, M. (1984). *Use of the ACT Assessment by Examinees with Disabilities* (Research Report No. 84). Iowa City, IA: American College Testing.

Lindstrom, J. H. (2007). Determining Appropriate Accommodations for Postsecondary Students with Reading and Written Expression Disorders. *Special Issue of Learning Disabilities Research & Practice: Postsecondary Learning Disabilities,* 22(4), 229-236.

Oren, C. & Even, A. (2005). *The Fairness and Validity of the Higher Education Selection System for Students with Disabilities* (Report No. 325). Jerusalem: National Institute for Testing and Evaluation.

Rosenbaum, P. R., & Rubin, D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika,*70, 41-55.

Solórzano, R. W. (2008). High Stakes Testing: Issues, Implications, and Remedies for English Language Learners. *Review of Educational Research,* 78(2), 260-329.

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.

Turvall, E., Bronner, S., Kennet-Cohen, T. & Oren, C. (2008). *Fairness in the Higher Education Admissions Procedure: The Psychometric Entrance Test in Arabic*. (Report No. 349). Jerusalem: National Institute for Testing and Evaluation.

Young, J. W. (2001). *Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis* (College Board Research Report No. 2001-6). New York: College Board.

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. & Bryant, D. P. (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.