Research Report

Examining Test Items for Differential Distractor Functioning Among Students With Learning Disabilities

Kyndra Middleton

Cara Cahalan Laitusis

Research & Development

November 2007
RR-07-43

Examining Test Items for Differential Distractor Functioning Among Students With Learning Disabilities

Kyndra Middleton and Cara Cahalan Laitusis
ETS, Princeton, NJ

November 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study examined whether distractor choices functioned differently for students without learning disabilities than they functioned for students with learning disabilities who received no accommodation, students with learning disabilities who received a read-aloud accommodation, and students with learning disabilities who received some form of accommodation other than read-aloud. The study's purpose was twofold: (a) to examine the results of the DDF analysis to determine whether the distractors functioned differently for the various groups of students and (b) to aid in determining whether the test may be modified for those students with learning disabilities by removing a distractor while maintaining adequate test validity and information.

Key words: Differential distractor functioning, read-aloud accommodation, learning disability

Table of Contents

Introduction

Prior to 1997, it was common practice to exempt many students with disabilities from statewide achievement tests. This practice changed with the reauthorization of the Individuals with Disabilities Education Act of 1997 (IDEA, 1997), which mandated that students with disabilities be included in standardized assessments and that accommodations be made where appropriate to allow their inclusion. This mandate not only required schools to include students with disabilities in standardized test administrations but also required these schools to include their scores in the reporting of student progress and achievement for accountability purposes. The responsibility for determining appropriate testing accommodations was placed on the child study team, which develops an individualized education plan (IEP) for each student with a disability; however, each state could develop guidelines to be used when determining reasonable and appropriate accommodations.

The No Child Left Behind Act of 2001(NCLB) redefined the role of the federal government in K-12 education. Along with mandating annual student testing in grades 3-8, the act stipulates that assessments provide accommodations for students with disabilities as defined in the Individuals with Disabilities Education Act (IDEA) of 1991 and of 1997. It also mandates the reporting of assessment results and state progress by student groups based on poverty, race and ethnicity, disability, and limited English proficiency. Since the passage of NCLB in 2001, states have increased the participation of students with disabilities in accountability assessments. In addition to this increased participation, a greater percentage of students who receive special education services are taking part in standards-based instruction and achieving proficiency on state standards-based assessments (Thompson, Thurlow, Johnstone, & Altman, 2005). This practice of inclusion raises important questions, however, concerning the appropriateness of common performance standards for students with disabilities, the appropriate accommodations to use, the effects of testing accommodations on the validity of the assessment, and the reporting of scores when accommodations have been used (Pitoniak & Royer, 2001).

To address these concerns, several studies have examined the psychometric properties of current state assessments for students with disabilities. These studies have focused primarily on determining whether tests have the same factor structure for two populations (e.g., students with disabilities and students without disabilities) or examining whether test items perform similarly for students with and without disabilities (i.e., differential item functioning). This paper will

review the research on differential item functioning (DIF) and differential distractor functioning (DDF) with a primary focus on students with disabilities and extend this research by examining DDF comparing students with and without learning disabilities on a state English language arts assessment.

Differential Item Functioning

When assessing test validity, researchers can examine the test itself to determine whether test scores have the same meaning for different groups, or they can choose to examine the individual test items to determine whether an item is performing differently for different groups through a process called differential item functioning (DIF). DIF occurs when examinees with the same estimated latent trait level have differing probabilities of answering the same item correctly (Lord, 1980). Thus far, most DIF analyses have been conducted between groups that differed on gender, ethnicity/race, and socioeconomic status. Until recently, there has not been a large enough population of students with learning disabilities to provide sufficient sample size for DIF analyses. However, with the increase in the number of students identified as having learning disabilities and the implications of NCLB, which requires students with disabilities to take statewide assessments, researchers can now obtain large enough samples to compare students without disabilities to those with disabilities. DIF analyses have become a vital part of the test assembly process, since knowing how each item on a test functions among different groups is highly important to test developers. Thus, researchers are beginning to closely examine item differences between students with and without learning disabilities.

Differential Distractor Functioning

Another method for examining items individually is through the use of differential distractor functioning (DDF) analyses. DDF analyses are used to determine whether different distractors, or incorrect option choices, attract various groups differentially (Green, Crone, & Folk, 1989). Since most standardized test items are dichotomously scored as either correct or incorrect, the distractors chosen by those answering the item incorrectly were not the focus of much concern in the past. However, distractors are becoming more important, because they can provide an additional source of information about student performance, particularly as states begin to consider the elimination of distractors as one way of developing modified alternate assessments (Barton, 2006). The results of DDF analyses will allow test developers to determine

which items may need further observation, since items exhibiting DDF have distractors that are functioning differently for different groups. Thus the distractors may provide a possible reason why a particular item exhibits DIF.

DIF and DDF Research

Since there is a vast body of literature related to DIF in various subject areas and between various groups, this paper will present a more narrow review of DIF and DDF studies. More specifically, the DIF studies discussed will be those related to reading assessments and gender as well as those related to reading assessments and race. Additionally, a more detailed discussion of DIF involving students with and without learning disabilities in math and reading will be provided. Because DDF has only recently become an area of interest, few previous studies have examined this topic. As such, the discussion related to DDF will include cultural differences in distractor choice on a reading assessment, racial differences in distractor choice on the verbal portion of the SAT®, and gender differences in distractor choice on a mathematics assessment.

DIF research. A study by Stoneberg (2004) used data from between 8,500 and 10,000 students from 4th, 8th, and 10thgrade who took the Idaho Standards Achievement Tests (ISAT) in spring 2003. DIF was assessed between males and females and between Hispanic and White test-takers using the simultaneous item bias test (SIBTEST) and the Mantel-Haenszel (MH) procedure, which are two methods commonly used to assess DIF (Holland & Thayer, 1988; Shealy & Stout, 1993). The 4th- and 8th-grade tests comprised 42 reading, language usage, and mathematics items each. The 10th-grade test comprised 55 reading, 56 language usage, and 60 mathematics items. Only the DIF results for the reading items are reported here. Stoneberg found that in 4th-grade reading items assessed for DIF by gender, 36% of the items were identified as showing DIF using SIBTEST, 31% were identified using MH, and 29% were identified using both methods. In the 8th-grade test, 43% of the items were identified as showing DIF using SIBTEST, 50% were identified using MH, and 43% were identified using both methods. Among 10th-grade items, 42% of the items were identified as having DIF using SIBTEST, 44% using MH, and 36% using both methods. Overall, the percent agreements between the two procedures were 86%, 92%, and 85% for 4th, 8th, and 10th grades, respectively.

When the 4th-grade items were examined for DIF based on ethnicity/race, 12% of the items were identified using SIBTEST, 21% were identified using MH, and 12% were identified using both methods. For the 8th-grade test, 12% were identified using SIBTEST, 12% were

identified using MH, and 10% were identified using both procedures. For the 10th-grade items, 15% were identified using SIBTEST, 16% were identified using MH, and 15% were identified using both procedures. For ethnicity/race, the percent agreements between the two procedures were 71%, 80%, and 94% for 4th, 8th, and 10th grades, respectively.

The literature related to students with learning disabilities and DIF is still being developed, since this area has only recently begun to gather attention. Bolt (2004) conducted a study in which she compared several groups of students with different disabilities to students without a disability. The results of interest for her study were the comparisons among students without a disability, those with disabilities who did not receive an accommodation, and those with disabilities who received a read-aloud accommodation. Data were collected from three different states, with two of the states using reading and math assessments and the other state using communication arts and math assessments. The sample sizes for math ranged from 5,219 on one state's assessment to 58,051 on another state's assessment. The reading assessment sample sizes ranged from 2,556 to 44, 330.

The study found that on State 1's 30-item math assessment, 17% of the items showed DIF when the nonaccommodated elementary students with disabilities were compared to the nonaccommodated students without disabilities, while 77% of the items showed DIF when the read-aloud accommodated students with disabilities were compared to the nonaccommodated students without disabilities. On the 30-item reading assessment, 30% of the items for the nonaccommodated students with disabilities and 73% of the items for the read aloud-accommodated students with disabilities showed DIF when compared to the nonaccommodated students without disabilities. On State 2's 30-item elementary math assessment, 13% of the items exhibited DIF when the nonaccommodated students with disabilities were compared to the nonaccommodated students without disabilities, and 41% of the items showed DIF when the read-aloud accommodated students with disabilities were compared to the nonaccommodated students without disabilities.

On the 41-item communication arts assessment, 24% of the items exhibited DIF when the nonaccommodated students with disabilities were compared to the nonaccommodated students without disabilities, and 59% of the items exhibited DIF when the read-aloud accommodated students with disabilities were compared to the nonaccommodated students without disabilities. In State 3's high school assessments, 31% of the items on the 39-item math assessment showed

DIF when the nonaccommodated students with disabilities were compared to the nonaccommodated students without disabilities, and 38% showed DIF when the read-aloud accommodated students with disabilities were compared to the nonaccommodated students without disabilities. Twelve percent of the 25 reading items showed DIF when the nonaccommodated students with disabilities were compared to the nonaccommodated students without disabilities, and 40%, showed DIF when the read-aloud accommodated students with disabilities were compared to the nonaccommodated students without disabilities.

Bolt and Diao (2005) examined whether subsets of reading items functioned differently for students with disabilities in grades 4, 5, and 7 who received a read-aloud accommodation compared to students without a disability receiving no accommodation. Among 4th graders (n = 207), 20% of the 35 items showed DIF using MH, and 11% of the items showed DIF using SIBTEST. Among 5th graders (n = 121), 9% of the 32 items showed DIF using the MH procedure, and about 13% of the items showed DIF using the SIBTEST procedure. DIF was found for 9% of the 33 items on the test for 7th-grade students (n = 141) using both MH and SIBTEST. Although several items displayed measurement dissimilarity, there was not a pattern of DIF among the reading subtests that uniformly favored the read-aloud group.

DDF research. Marshall (1983) examined sex differences in distractor choice using the 6th-grade version of the California Survey of Basic Skills over a 3-year period. The author was interested in whether there were differences among gender, distractor choice, and the year the test was given, so she used loglinear model fitting, which allows main effects to be separated and interactions of focal variables to be categorized (Green et al., 1989). Of the 30 items on the test, 13% showed DDF between genders. Additionally, the pattern of errors was different depending on the year the test was taken, and 20% of the items showed DDF during the comparison between years. Marshall's study also used the incorrect answer choices to classify the types of errors made by males and females and found that the genders differed a great deal in the types of errors they made.

Green et al. (1989) also used loglinear model fitting to assess for DDF among test items in their study. They conducted a study on a form of the 1986 SAT verbal section, which contained 85 items. The study included 2,000 randomly selected White, Black, and Hispanic students. Among the different groups, 16% of the items exhibited DDF according to the set statistical significance level, and the options seemed to be performing the most differently among

Hispanic students. However, very few items showed differences by as much as 10%; most of the items showed differences between 2% and 3% between distractors.

Banks (2006) conducted a study examining DDF among 4,765 Black, Hispanic, and White examinees on the 5th-grade 20-item reading and language arts portions of the Terra Nova test, using aspects of Black culture in the distractors as a selection measure. After choosing items that exhibited certain aspects of the Black culture in the incorrect options and examining those groups of items for differential bundle functioning (DBF), Banks found that 11 items exhibited DBF. DBF, which is an extension of DIF, occurs when examinees with the same latent trait level have differing probabilities of answering the same groups (or bundles) of items correctly. These bundles are formed based on some predetermined criteria, such as item content or cluster analysis (Douglas, Roussos, & Stout, 1996). Banks further examined these 11 items for DDF using loglinear modeling, and of these items, eight exhibited DDF between Blacks and Whites. The author concluded that on the Terra Nova test, more distractors were likely to exhibit DDF than correct options were likely to exhibit DIF and thus suggested including DDF in the test analysis in order to provide a more thorough understanding of the items.

Implications for the Present Study

Two of the three reviewed DIF studies examined students with learning disabilities, and this is a fairly novel research focus. In those two studies, which included elementary, middle, and high school students, more items displayed DIF when the read-aloud accommodated group was compared to the students without learning disabilities who did not receive an accommodation than when any other group (e.g., students with disabilities who received extended time, setting accommodations, or no accommodations) was compared to students without learning disabilities who did not receive an accommodation. It is of interest to examine whether this holds true when students with disabilities who do not receive an accommodation are compared to students with disabilities who receive a read-aloud accommodation. The present study examines both comparisons.

The three DDF studies examined used loglinear model fitting to test for differences between groups in choice of distractors. Loglinear model fitting alone will not give information about individual distractors; it merely provides information on whether there is an association between group, distractor choice, and score level. The current study uses a DDF method that will provide information about individual distractor choices. This is of importance, since there have

been no known DDF studies of students with learning disabilities that have looked at individual item distractors.

Purpose of Study

This study examined whether different distractor choices functioned differently for students without learning disabilities than they functioned for students with learning disabilities who received no accommodation, students with learning disabilities who received a read-aloud accommodation, and students with learning disabilities who received some form of accommodation other than a read-aloud accommodation. Two examples of the other forms of accommodations that are intended to provide aid to those needing the accommodation, but not to change the construct of the test are receiving the test over multiple days and marking in the answer booklet. Such accommodations are usually specified in an individual's IEP/504 plan. The current study's purpose was twofold: (a) to examine the results of the DDF analysis to determine whether the distractors functioned differently for the various groups of students, and (b) to aid in determining whether the test may be modified for those students with learning disabilities by removing a distractor while maintaining adequate test validity and information.

Method

Sample

Data were sampled from approximately 460,000 4th-grade students who took a statewide criterion-referenced English language arts test. For the present study, English language learners from all groups of students (with and without disabilities) were excluded. Because of the very large size (approximately 300,000) of the students without disabilities group, this group was decreased in size through a random selection of 30,000 students. As a result, approximately 45,000 students were included in the current analysis. All of the students included in the analysis were those who tested on grade level when administered the standardized test during the 2004 school year. Students with a disability who were included in the sample performed significantly worse on the test than the remaining students in the sample. The percent of students within each proficiency level (*far below basic*, *below basic*, *basic*, *proficient*, and *advanced*) are found in Table 1. As can be seen from the table, the percentage of students with learning disabilities who scored *below basic* or *far below basic* in reading ability ranged from 62% to 73%, whereas among those students without disabilities, only 15% scored *below basic* or *far below basic*.

Table 1

Percentage of Students at Each Proficiency Level

Subgroup	Far below basic	Below basic	Basic	Proficient	Advanced	Total number of students
No disability	4%	11%	31%	30%	25%	30,000 [a]
Learning disability—no accommodation	32%	34%	25%	7%	2%	9,056
Learning disability—IEP/504	35%	38%	23%	4%	1%	4,727
Learning disability—IEP/504 & read-aloud	26%	36%	30%	6%	1%	1,371

[a] These students were randomly selected from a population of approximately 300,000 students.

Measure

The Grade 4 English language arts test, which consisted of 81 four-option multiple-choice items based on the state's curriculum standards, was used for the analysis. Six of the 81 items did not count towards the student's score since they were being pretested for the next year. As such, these items were excluded from the analysis, reducing the number of analyzed items to 75. There were 42 reading items and 33 writing items. Based on the test blueprints, 24% of the items were related to word analysis, fluency, and systematic vocabulary development; 20% were related to reading comprehension; 12% were related to literary response analysis; 24% were related to written and oral English language conventions; and 20% were related to writing strategies. The test also contained an essay question, which was not included in the analysis.

Procedure

Although there are several methods used to assess DDF, the two most commonly used methods are loglinear model fitting and standardized distractor analysis. Loglinear model fitting is used to achieve the most parsimonious model and is able to distinguish between items that are uniformly and nonuniformly biased (Mellenbergh, 1982). A drawback of loglinear modeling,

however, is that there is no way of distinguishing between distractors, which means the researcher will be unable to determine which distractor(s) in the item are causing the item to show DDF. Standardized distractor analysis (SDA), however, examines each distractor separately and thus is able to establish which distractors show DDF. This method is also able to account for both uniform and nonuniform DDF in an item. The current study therefore uses the SDA method of DDF to provide more detailed information. Although SDA is capable of identifying uniform and nonuniform DDF, information can also be lost at score extremes if there are not enough examinees in one or both groups in a particular interval. This issue will be discussed further in the Limitations section. It should also be noted that, because the two groups differ in proficiency and the internal matching variable includes both accommodated and nonaccommodated test scores, the DIF results should be interpreted with caution. However, several studies have been done comparing students with learning disabilities to students without learning disabilities using this methodology, and they have produced useful findings (Bolt, 2004; Bolt & Ysseldyke, 2006; Koretz & Hamilton, 1999).

SDA is an extension of the standardized *p*-difference used by Dorans, Schmitt, and Bleistein (1992), which is a formula used to test an item for DIF. SDA uses the same formula as the standardized *p*-difference, except it is generalized to include correct and incorrect options. The SDA formula is as follows:

$$\text{SDA}(i) = \frac{\sum \{W_s [P_{fs}(i) - P_{rs}(i)]\}}{\sum W_s},$$

where *i* is usually option A, option B, etc., W_s is the weighting factor at a particular score category, and P_{fs} and P_{rs} are the percent choosing option *i* in the focal and reference groups. W_s can be any one of the following:

1. the number of examinees at a particular score category in the total group
2. the number of examinees at a particular score category in the reference group
3. the number of examinees at a particular score category in the focal group
4. the relative frequency at a particular score category in the reference group.

However, most researchers suggest using (3), since it provides the greatest differences between the percent choosing an option among the two groups at the most common score levels of the focal group.

In a DIF analysis, using SDA allows for examinees to be matched using both thick and thin matching. Thick matching occurs when scores are grouped into larger intervals (e.g. 1-10, 11-20, but can be as large as the entire score range in the extreme case). Thin matching occurs when the scores are grouped into smaller intervals (e.g. 1-3, 4-6, but can be as small as one score per interval in the extreme case; Clauser & Mazor, 1998). The current study used thin matching with one score per interval.

According to Dorans et al. (1992), SDA items are classified as exhibiting negligible DDF $(|\text{SDA} < .05|)$, moderate DDF $(.05 \leq |\text{SDA}| < .10)$,or large DDF $(|\text{SDA}| \geq .10)$. In the present study, negative values are indicative of DDF favoring the reference group, while positive values are indicative of DDF favoring the focal group. Table 2 presents the different comparisons possible between groups.

Only those 10 test items that exhibited DIF were analyzed for DDF. However, it is possible for an item to show DDF without showing DIF. If this occurs, a distractor may be differentially attracting a particular group, but the distractor is not affecting the group's ability to choose the correct option. For this study, only the 10 items that previously showed DIF are of interest. The DDF analyses were used to aid in suggesting possible causes of the observed DIF for those items.

Results

This study examined only the 10 items found to exhibit DIF in a previous study. Of the 10 items examined, seven items were shown to exhibit DDF in this study. All seven of the items that displayed DDF occurred in a comparison that included the read-aloud group. For the comparison of students without a learning disability to students with a learning disability who received a read-aloud accommodation, nine items displayed DIF. Of those items, five showed moderate DDF favoring the students receiving a read-aloud accommodation in at least one of the distractors. An additional item showed large DDF in favor of students receiving the read-aloud accommodation in one option. There were also three distractors in two of the items that exhibited moderate DDF in favor of the students without a learning disability. Additionally, one of the

items that showed moderate DIF in favor of the read-aloud group contained an option that displayed DDF in favor of the group of students without disabilities. In this item, there was DIF in the correct option and DDF in all three distractors. This item may need further examination to discover a possible cause of this unexpected behavior.

Table 2

Reference-Focal Comparisons

Reference group	Focal group
Group 0 - No disability	Group 20 - Learning disability—no accommodation
Group 0 - No disability	Group 21 - Learning disability—IEP/504
Group 0 - No disability	Group 22 - Learning disability—IEP/504 & read-aloud
Group 20 - Learning disability—no accommodation	Group 21 - Learning disability—IEP/504[a]
Group 20 - Learning disability—no accommodation	Group 22 - Learning disability—IEP/504 & read-aloud

[a] Comparison did not show DIF so was not included in the DDF analyses. IEP = individualized education plan.

In the comparison between the students without learning disabilities and students with learning disabilities receiving no accommodation, the only item to exhibit DIF did not display DDF. Of the two items that showed DIF in the comparison between students without learning disabilities and students with learning disabilities not receiving the read-aloud accommodation, neither showed DDF. For the comparison between students with learning disabilities receiving no accommodation and those receiving a read-aloud accommodation, two items showed DIF, but only one of those two items displayed moderate DDF in favor of students receiving a read-aloud accommodation.

Table 3 summarizes the results of the DDF analyses. In the analyses, the proportion of examinees choosing a distractor in the reference group was always subtracted from the proportion of examinees choosing a distractor in the focal group, only as a matter of preference. Shaded boxes represent items that did not display DIF in the particular comparison. The R and F values represent DIF in favor of the reference and focal groups, respectively. Because the reference group was always subtracted from the focal group during DDF analysis, a +is

indicative of moderate DDF in favor of the focal group, and a "-" is indicative of moderate DDF in favor of the reference group. Large DDF in favor of the focal group is indicated by ++,and no items exhibited large DDF in favor of the reference group. A more detailed display of the results of the SDA can be found in Appendices A and B. Each set of figures in Appendix A plots a graph of the differences between the reference and focal groups for each item and its four options. The total score can be found on the horizontal axis and the difference in percent can be found on the vertical axis. For the correct option, if more of the curve is below the horizontal axis, the DIF occurs in favor of the reference group. If more of the curve is above the horizontal axis, the DIF occurs in favor of the focal group. With the distractors, the opposite is true. If more of the curve falls above the horizontal axis, the DDF occurs in favor of the reference group. If more of the curve falls below the horizontal axis, the DDF occurs in favor of the focal group. Appendix B provides a table of the actual values produced by the SDA.

Table 3

DDF Results Based on Group Comparisons and Distractor Choice

	Comparison groups																			
	0-20				0-21				0-22				20-21				20-22			
Item	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
3	R				R															
10										+		F						+		F
13									+	+	F	-								
25											++	F								
32											R	-								
33										F	+									
34										+	F									
45									R											
56									R											
64						R			-	R	-							R		

Note. Since the keyed option is correct, if an item shows DIF, positive values favor the reference group. However, with the distractors, if an item shows DDF, positive values favor the focal group, since the option is the incorrect option.+ represents moderate DDF in favor of the focal group; ++ represents large DDF in favor of the focal group; - represents moderate DDF in favor of the reference group. R represents DIF in favor of the reference group; F represents DIF in favor of the focal group. Shaded boxes represent items that did not exhibit DIF.

Discussion

Several of the distractors allowed for the creation of hypotheses as to why those distractors performed as they did. For example, in the comparison of the group without disabilities to the group receiving a read-aloud accommodation, the one item that displayed large DDF (in favor of the focal group) was the most difficult item that displayed DIF (based on the item difficulty index). An examination of the distractor in which DDF occurred showed that the distractor could be synonymous with the correct option, depending on which definition was chosen. Thus, students who knew the correct definition of the stem may have been drawn to this incorrect option because of its synonymous relationship to the correct option. Overall, six of the seven items that displayed DDF were moderately difficult or very difficult, based on the item difficulty index.

Additionally, in the comparisons between the group without disabilities and the group receiving a read-aloud accommodation, one of the items showed DDF in all three of its distractors (two favoring the focal group and one favoring the reference group). One of the distractors favoring the focal group seems to be a viable option given the story's content, leading the reference group to choose this incorrect option.

Another of the items that differed in the comparison of students without disabilities to those receiving a read-aloud accommodation relates to spelling. The two options showing DDF in favor of the reference group are words that are homophones, whose homophonic counterpart is more commonly used. Because the students receiving a read-aloud accommodation may be more familiar with the alternate spelling of the word, they may be led to believe the word is spelled incorrectly. Additionally, on a test of spelling where the student has to identify the word that is spelled incorrectly, a read-aloud accommodation could provide an extra source of difficulty, because the student is trying to listen as the words are being spelled.

Other distractors did not perform in a manner that allowed a specific hypothesis to be formed based on how the different groups performed. As such, those items exhibit DDF, but no explanation can be provided as to possible causes of the difference in behavior between the groups.

Because no single type of distractor exhibited DDF consistently among the different items, it appears that no single type of distractor can be eliminated without affecting the test used in this study. In fact, there were more distractors displaying DDF in favor of the students with learning disabilities than there were displaying DDF in favor of the students without learning

disabilities. In other words, students without learning disabilities were differentially drawn to the incorrect option more often than the students with learning disabilities. Even if a distractor could be eliminated from an item, further studies would be needed to ensure that there is not a resulting loss of information or decrease in the validity of test scores.

Conclusions

Seven of the items (17%) that assessed performance on the state's reading standards showed DDF, whereas only three of the items (9%) that assessed performance on the state's writing standards showed DDF. This may indicate a need for more emphasis to be placed on the development of reading items to aid in increasing construct validity among students with learning disabilities. Because the items showing DDF were distributed across different content and cognitive categories according to the test's blueprint, there were no content areas in which the various groups of students were consistently performing differently. Of the ten items that displayed DIF among the different groups, the comparison between the students without disabilities and students receiving the read-aloud accommodation contained the largest number of items showing DIF (nine), and of those items, seven (78%) displayed DDF. It should be noted that out of all the distractors that exhibited DDF, seven (64%) favored the students receiving the read-aloud accommodation. Additionally, the comparisons that included the read-aloud group contained 100% of the distractors that exhibited DDF, suggesting that the items may be behaving differently and thus causing measurement dissimilarity for the students who receive a read-aloud accommodation. Another possibility is that the way these items were read aloud (not the characteristics of the items themselves) may have encouraged students to select one distractor or the correct answer over the other choices.

Limitations

The way SDA is computed allows for cancellation to occur between different score points, making the procedure less sensitive to real differences in scores. For example, students at the lower end of the distribution in the focal group may be differentially attracted to a particular distractor, but students at the higher end of the distribution in the reference group may be differentially attracted to the same distractor. If the two score points have values that are similar in magnitude, but differ in sign, the distractor may produce a SDA value that shows only negligible DDF, when in fact moderate or even large DDF may exist at certain ability levels. An

individual examination of the different score points would illustrate whether the DDF that exists is uniform or nonuniform.

Also, in the comparisons between students with learning disabilities and those without learning disabilities, there were not enough students scoring at either end of the distribution to produce reliable results, resulting in a loss of information for extremely low performing and extremely high performing students. More students with learning disabilities fell in the lower score intervals than students without learning disabilities, and more students without learning disabilities fell in the higher score intervals than students with learning disabilities. Because of the large differences in ability between those students without disabilities and those with disabilities, it is a bit questionable whether the two groups can be adequately compared, and more research is needed in this area. DIF results are affected if there are empty cells in one or both of the two groups at a particular interval. The empty intervals are excluded from the analysis, causing information to be lost.

The previously mentioned issue of empty cells poses another important problem for the DIF and DDF analyses in the present study. In 3 of the 5 comparisons, students without learning disabilities were compared to students with learning disabilities. By definition of their disability, the students with learning disabilities perform at a lower level than their nonlearning disabled counterparts. However, even when the students with learning disabilities received an accommodation, information was still lost at the lower end (and sometimes upper end) of the score scale, because at the extremes there were not enough nonlearning disabled students to be compared to the students with learning disabilities.

This brings the issue of score comparability to the forefront. Do the scores of students with learning disabilities (with and without an accommodation) mean the same thing as the scores of students without learning disabilities? An investigation of this question is currently underway, but thus far no definitive findings have been obtained. The issue depends on whether the accommodation is viewed as, in fact, an accommodation of the test or as a modification of the test. If the latter view is taken, then the scores are not comparable. However, if the change is viewed as an accommodation— a method or device that provides students with disabilities the same opportunity to perform as their nondisabled peers while preserving the original meaning of the test scores (Cortiella, 2006)—the scores should be comparable. In this case, a DIF analysis

can be conducted without much concern. More research is also needed to determine whether the read-aloud accommodation alters the test in a manner that forfeits the test's validity.

Because there are differences between the two groups being compared as well as differing administration conditions, interpretation of the DIF results may present an issue, since DIF should only be performed when using a well-defined matching variable. Because the current study was done using operational data, the DIF and DDF analyses were performed using the information available, which included the possibility of somewhat differing matching variables. Nonetheless, the results still appear to be informative. Had experimental data been available, where the same groups of students with and without learning disabilities took the test under both accommodated and nonaccommodated conditions (using two different forms), an alternative method could be used. For example, one could match students with and without disabilities on the accommodated score and then examine DIF and DDF, comparing accommodated students with disabilities to nonaccommodated students without disabilities. It should be noted that a rational explanation can be provided to justify why certain distractors favored one group over the other, lending credence to the idea that the matching variables may not be different enough to preclude interpreting the results. Additionally, the majority of items that did exhibit DIF did so between the group that received the read-aloud accommodation and some other group, a finding which is consistent with the previous literature.

References

Banks, K. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *Applied Measurement in Education, 19*(2), 115-132.

Barton, K. (2006, April). *Approaches to developing modified alternate assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Bolt, S. E. (2004, April). *Using DIF analyses to examine several commonly-held beliefs about testing accommodations for students with disabilities*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Bolt, S. E., & Diao, Q. (2005, August). *Reading-aloud a reading test: Examining reading subskill performance*. Poster presented at the annual meeting of the American Psychological Association, Washington, DC.

Bolt, S. E., & Ysseldyke, J. E. (2006). Comparing DIF across math and reading/language arts tests for students receiving a read-aloud accommodation. *Applied Measurement in Education, 19*(4), 329-355.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues & Practice, 17*(1), 31-44.

Cortiella, C. (2006). *NCLB and IDEA*: *What parents of students with disabilities need to know and do*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved December 5, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Parents.pdf

Dorans, N. J., Schmitt, A. P., & Bleistein, C. A. (1992). The standardization approach to assessing comprehensive differential item functioning. *Journal of Educational Measurement, 29*(4), 309-319.

Douglas, J. A., Roussos, L. A., & Stout, W. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*(4), 465-484.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A method for studying differential distractor functioning. *Journal of Educational Measurement, 26*(2), 147-160.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum Associates, Inc.

Individuals With Disabilities Education Act of 1991, 20 U.S.C. § 1400 *et seq.* (1991).

Individuals With Disabilities Education Act of 1997, 20 U.S.C. § 1412(a)(17)(A). (1997).

Koretz, D., & Hamilton, L. (1999). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject* (CRESST Rep. No. 498). Los Angeles, CA: Center for Research on Standards and Student Testing. (ERIC Document Reproduction Services No. ED 440 148).

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Marshall, S. P. (1983). Sex differences in mathematical errors: An analysis of distractor choices. *Journal for Research in Mathematics Education, 14*(4), 325-336.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*(2), 105-118.

No Child Left Behind Act of 2001, 20 U.S.C. § 6301 *et seq.* (2002).

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,* 53-104.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Stoneberg, B. D., Jr. (2004). *A study of gender-based and ethnic-based differential item functioning (DIF) in the spring 2003 Idaho Standards Achievement Tests applying the simultaneous bias test (SIBTEST) and the Mantel-Haenszel chi square test*. Unpublished manuscript, The University of Maryland-College Park and the National Center for Education Statistics.

Thompson, S. J., Thurlow, M. L., Johnstone, C. J., & Altman, J. R. (2005). *2005 State special education outcomes: Steps forward in a decade of change*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved May 23, 2006, from http://education.umn.edu/NCEO/OnlinePubs/2005StateReport.htm/

Appendix A

Figures Indicating the Degree of DIF and DDF for Each Studied Item

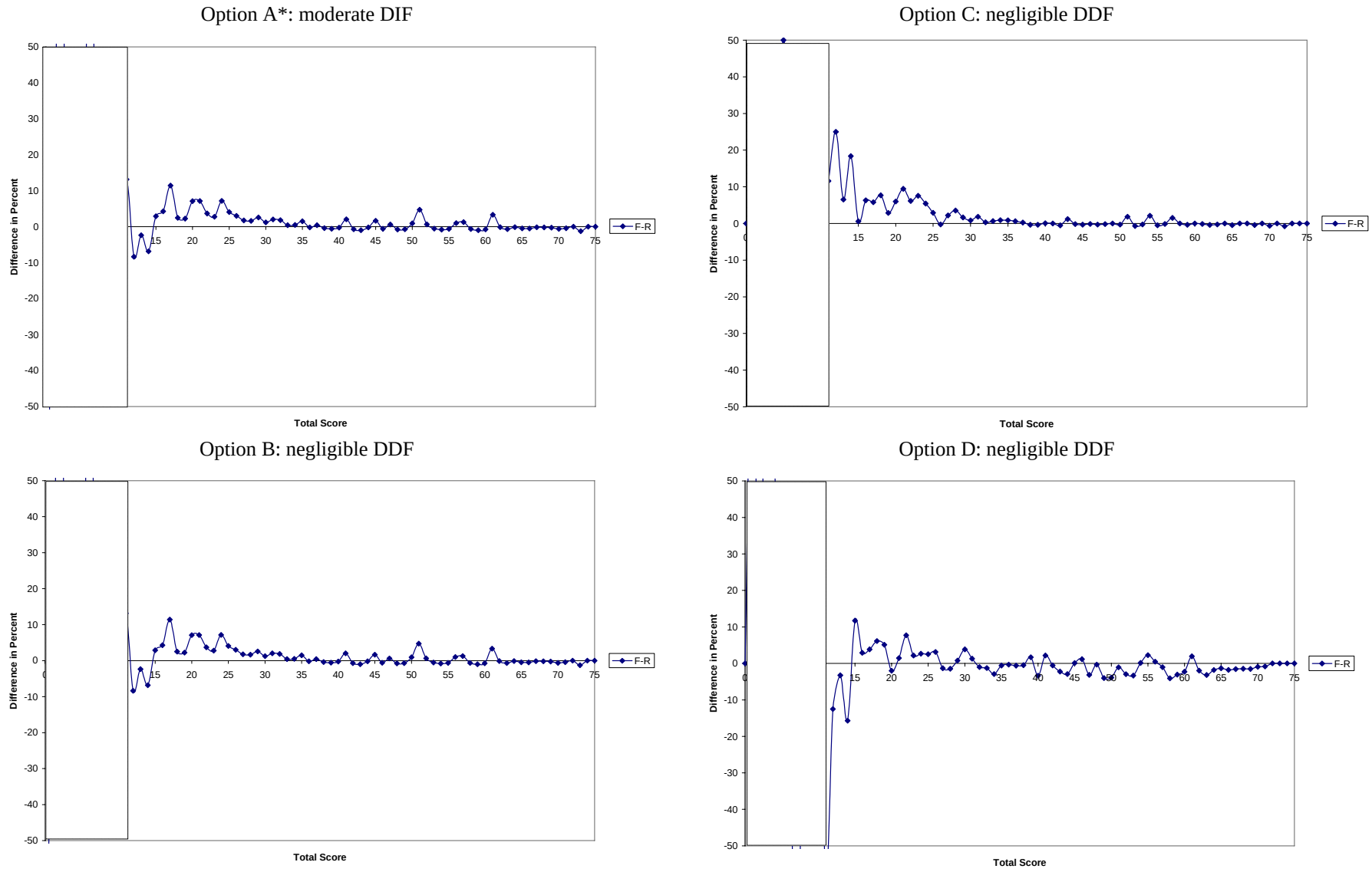


***Figure A1*. Item 3 differences (Group 0-20).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

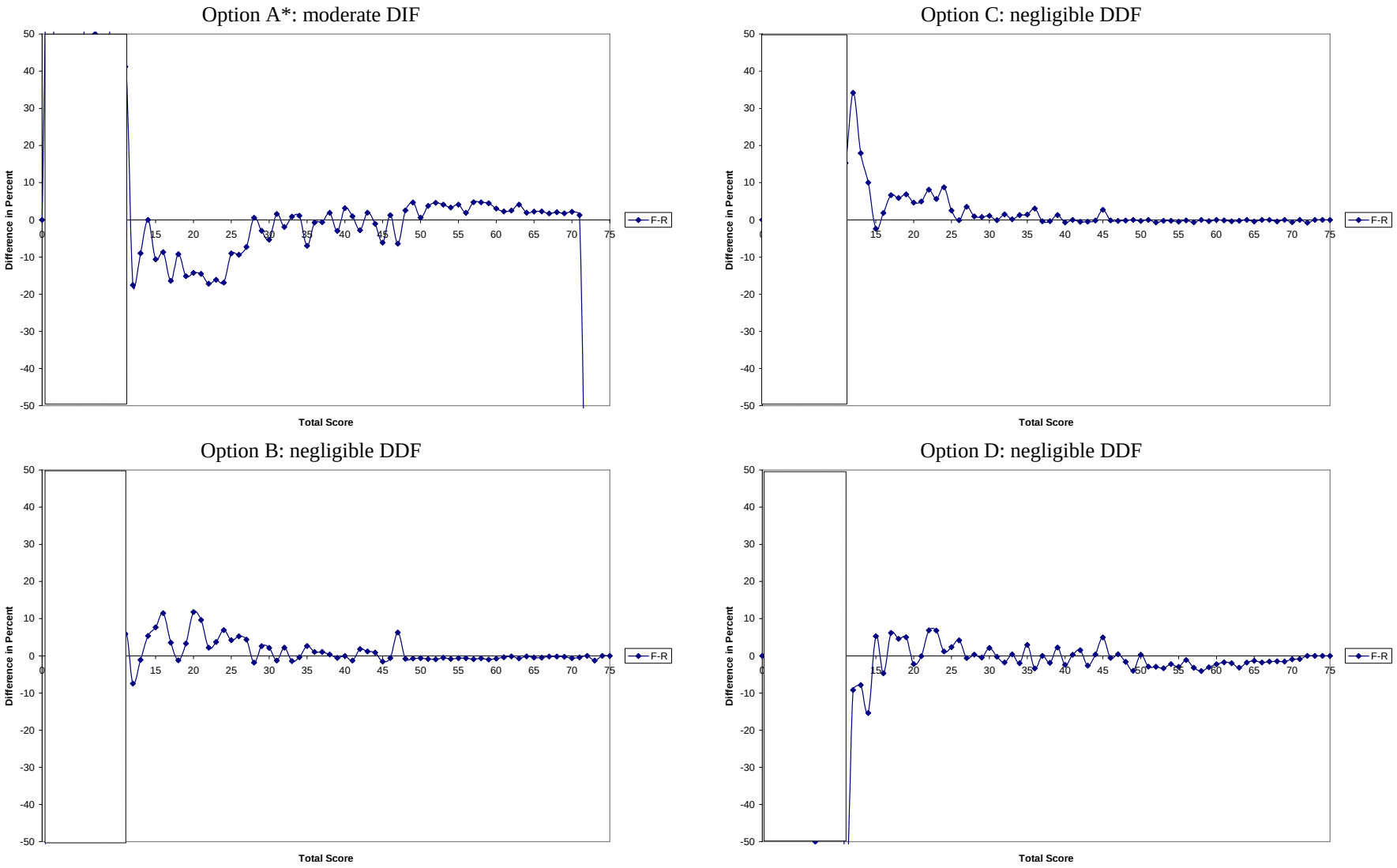


Figure A2. **Item 3 differences (Group 0-21).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

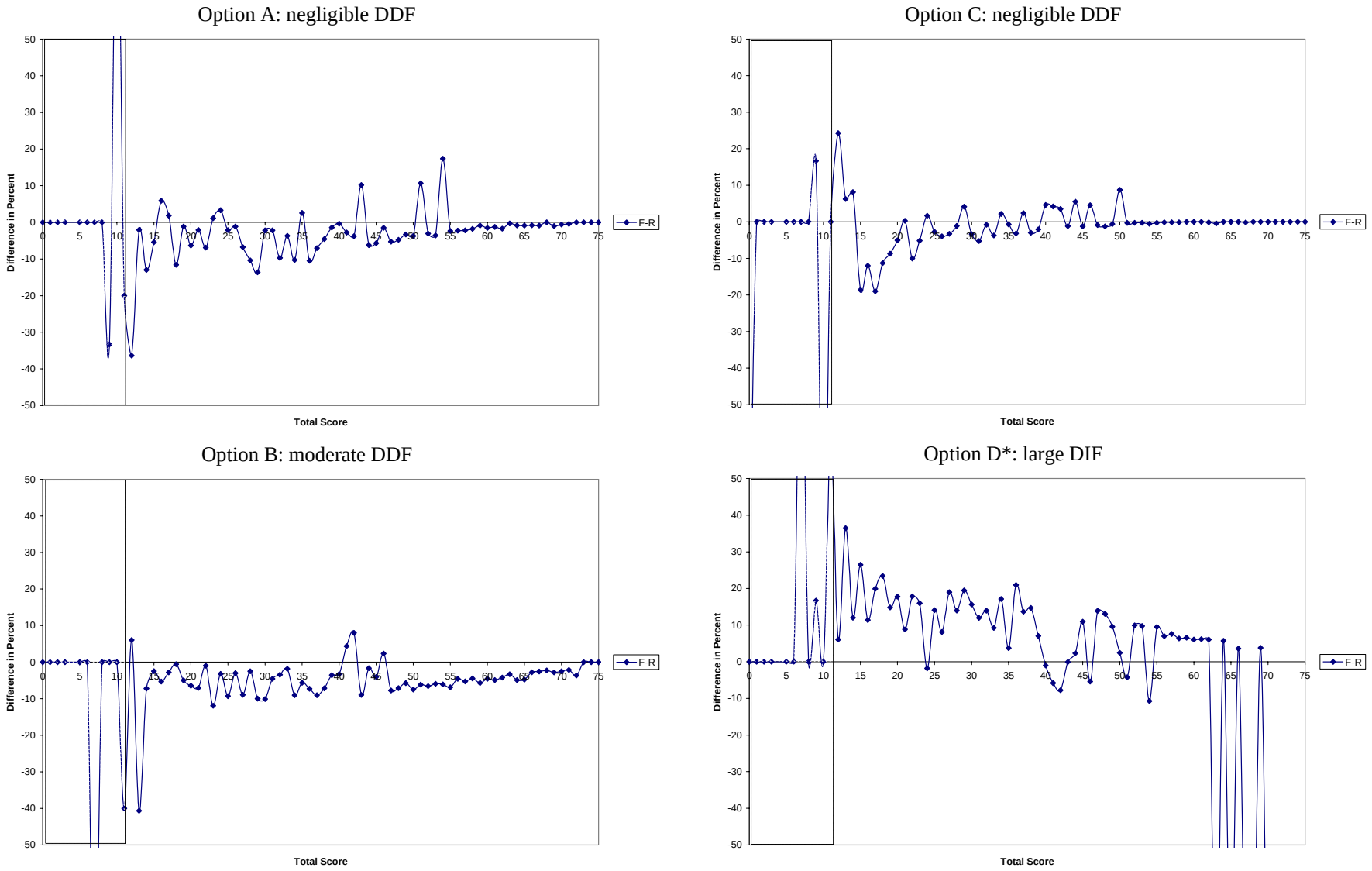


Figure A3. **Item 10 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

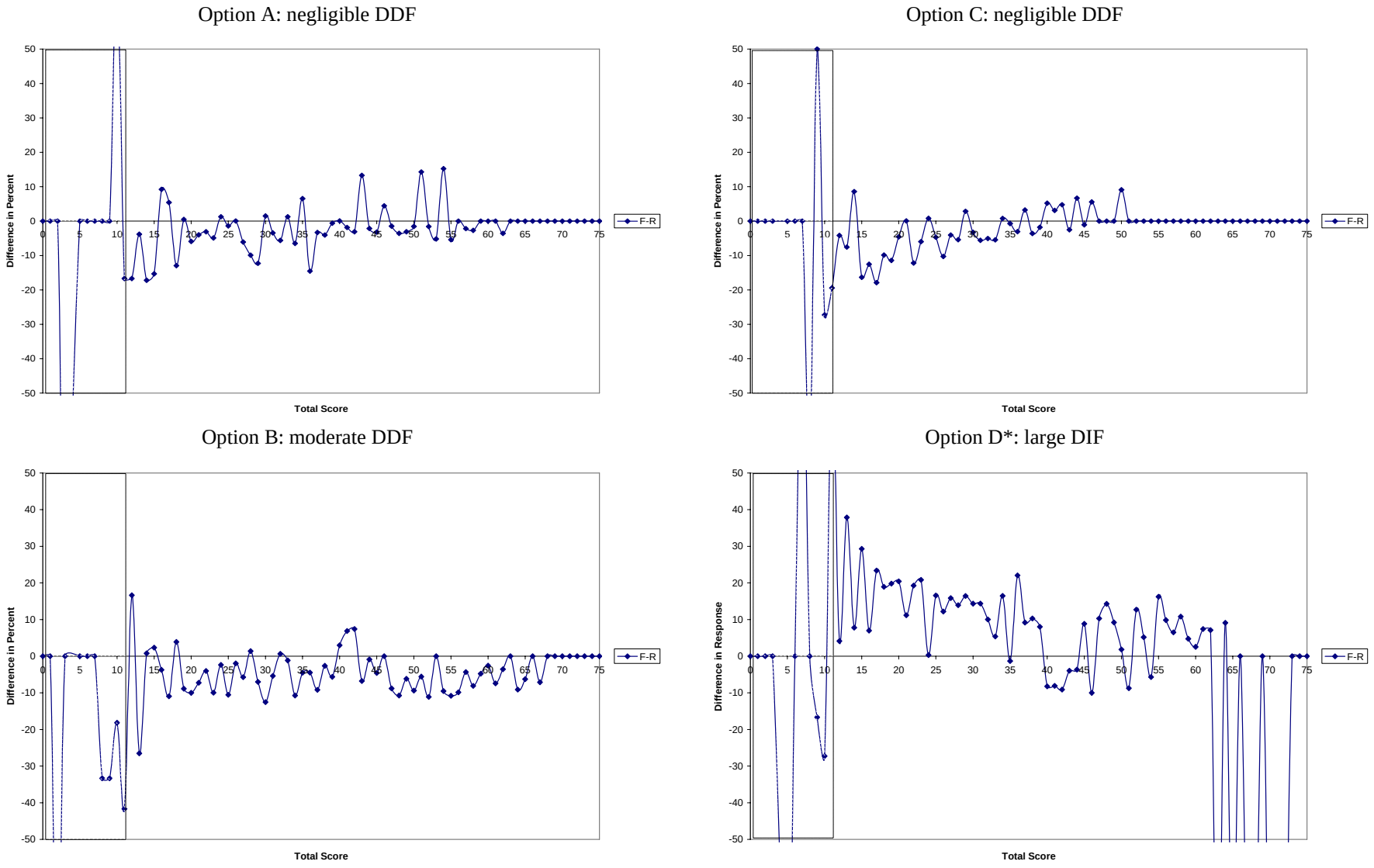


Figure A4. **Item 10 differences (Group 20-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

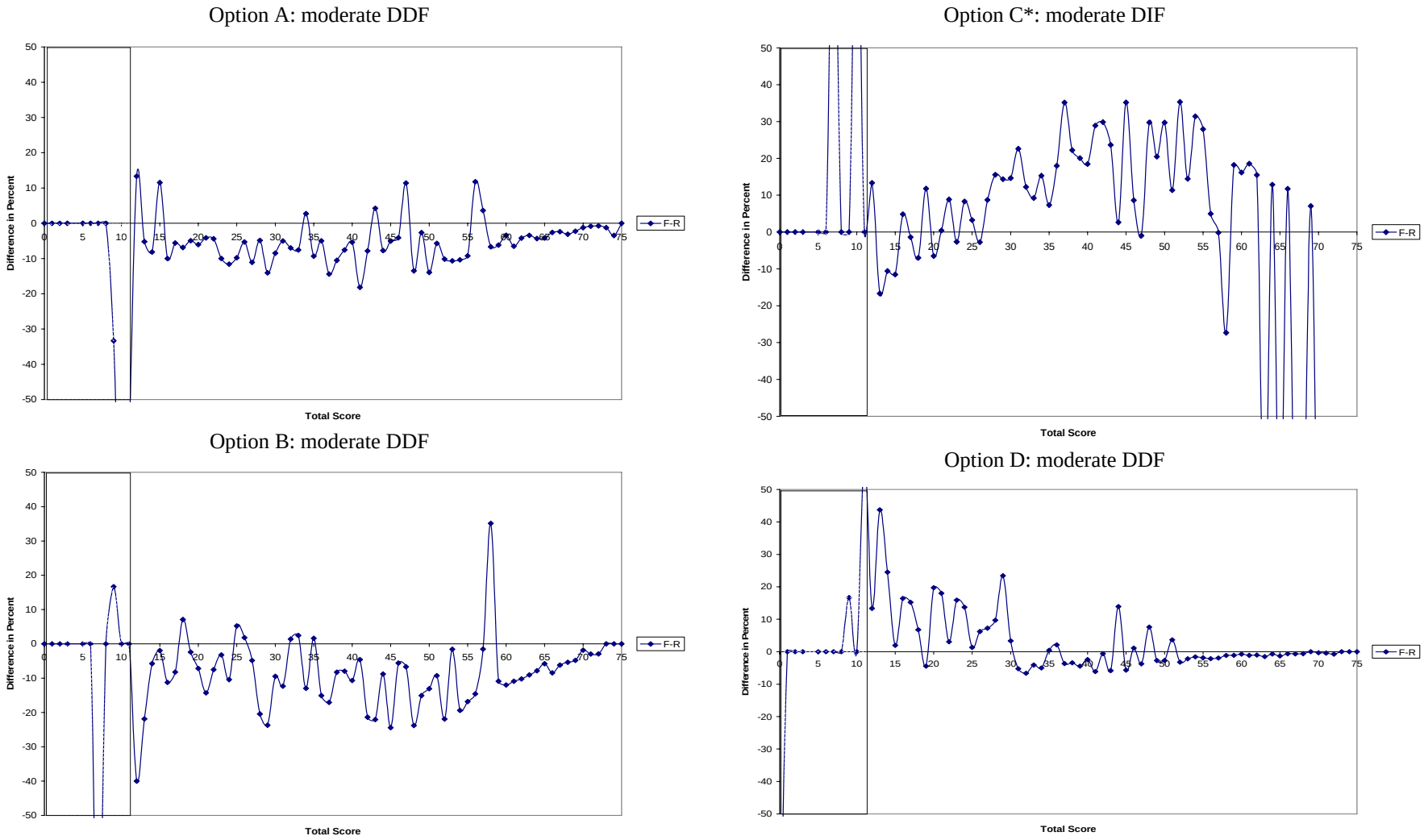


Figure A5. **Item 13 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

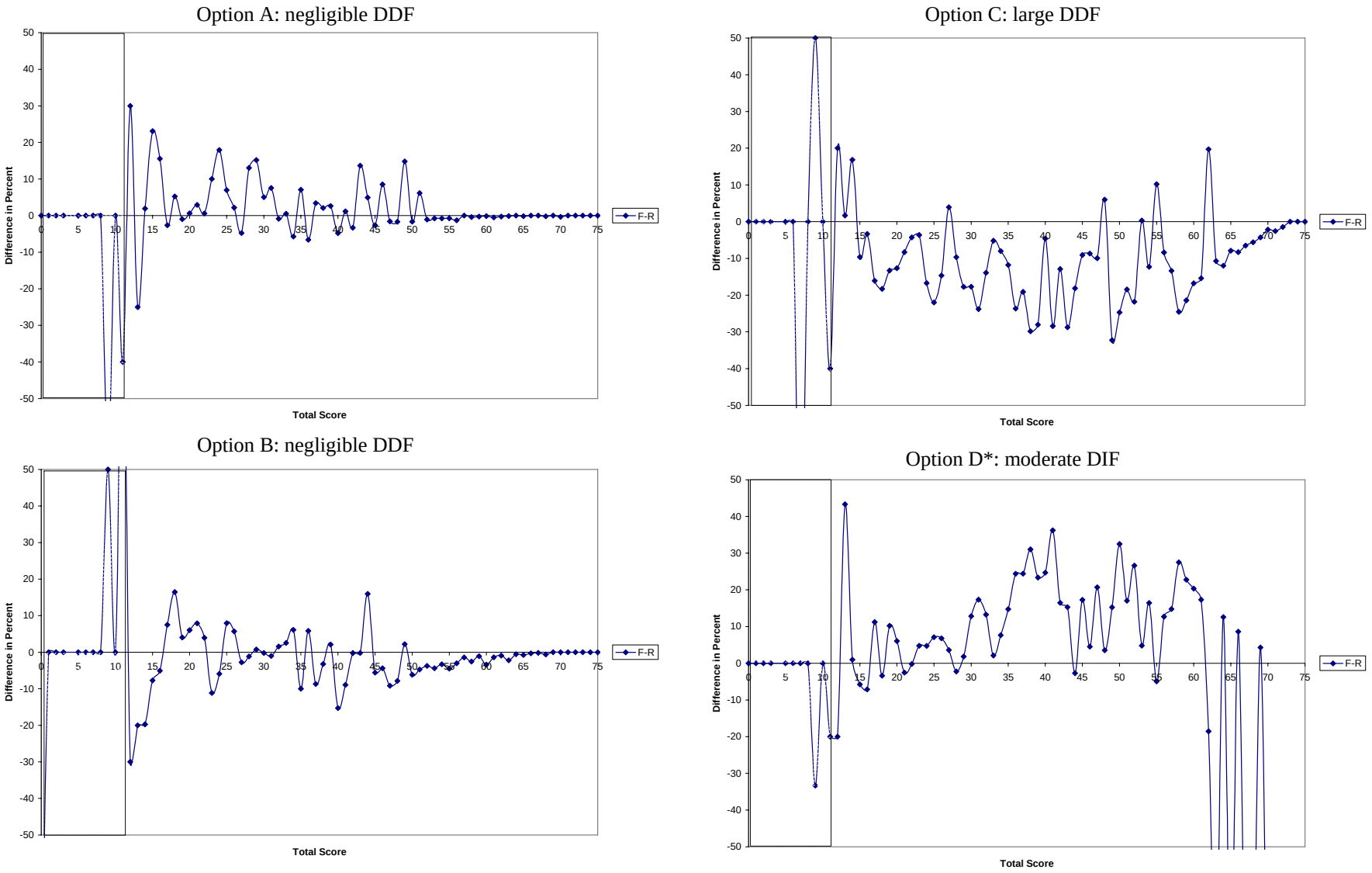


***Figure A6*. Item 25 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

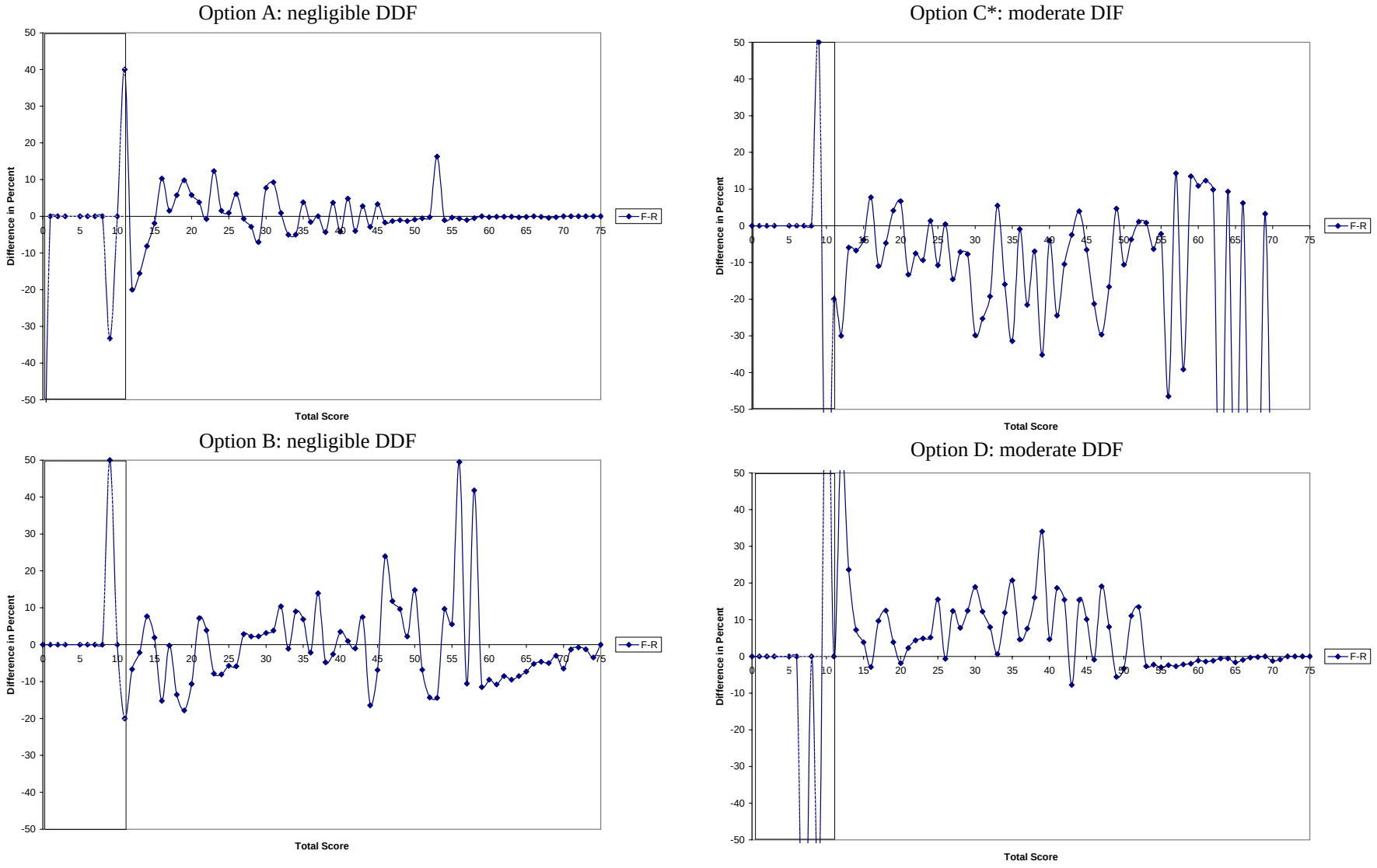


Figure A7. **Item 32 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

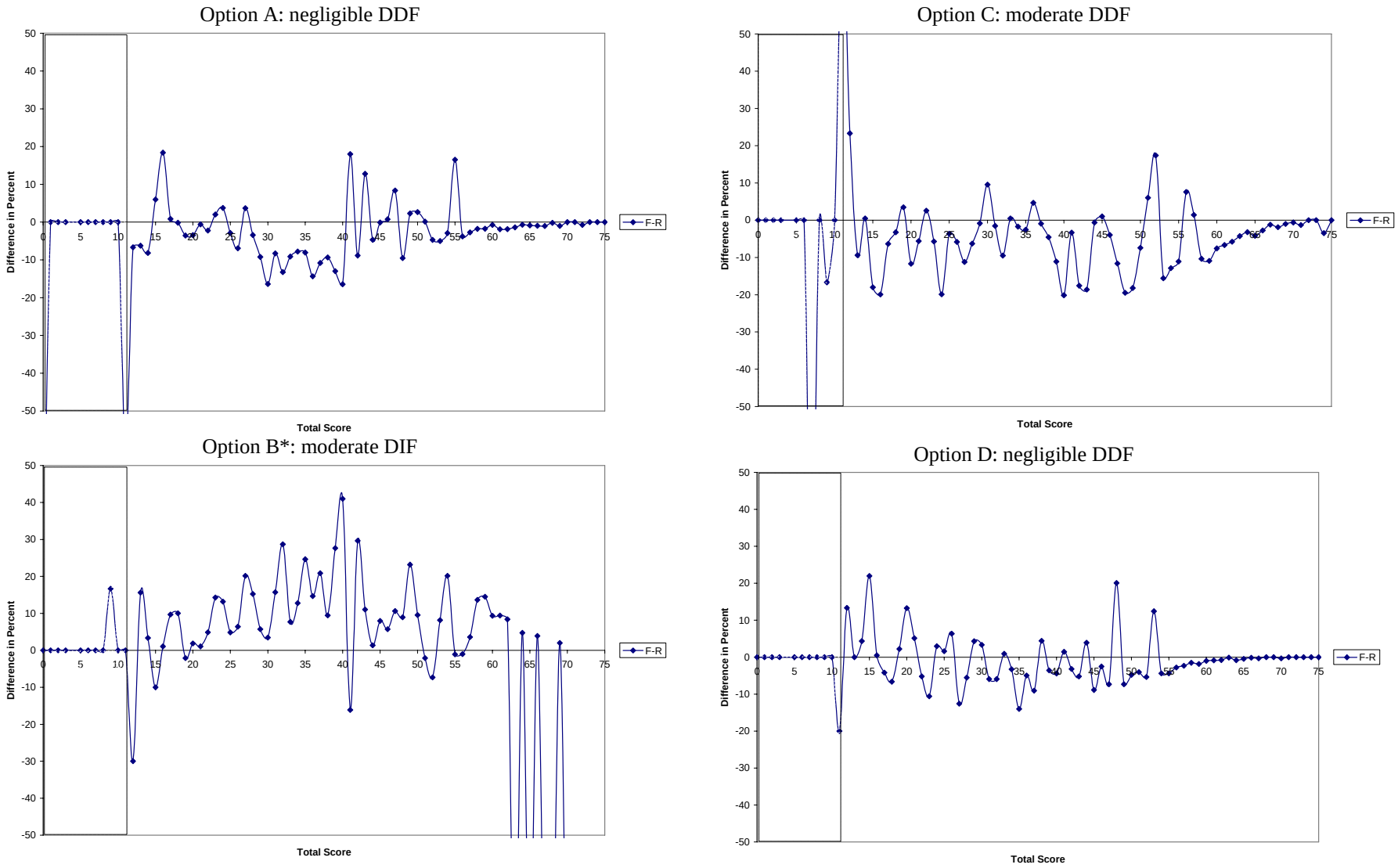


Figure A8. **Item 33 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

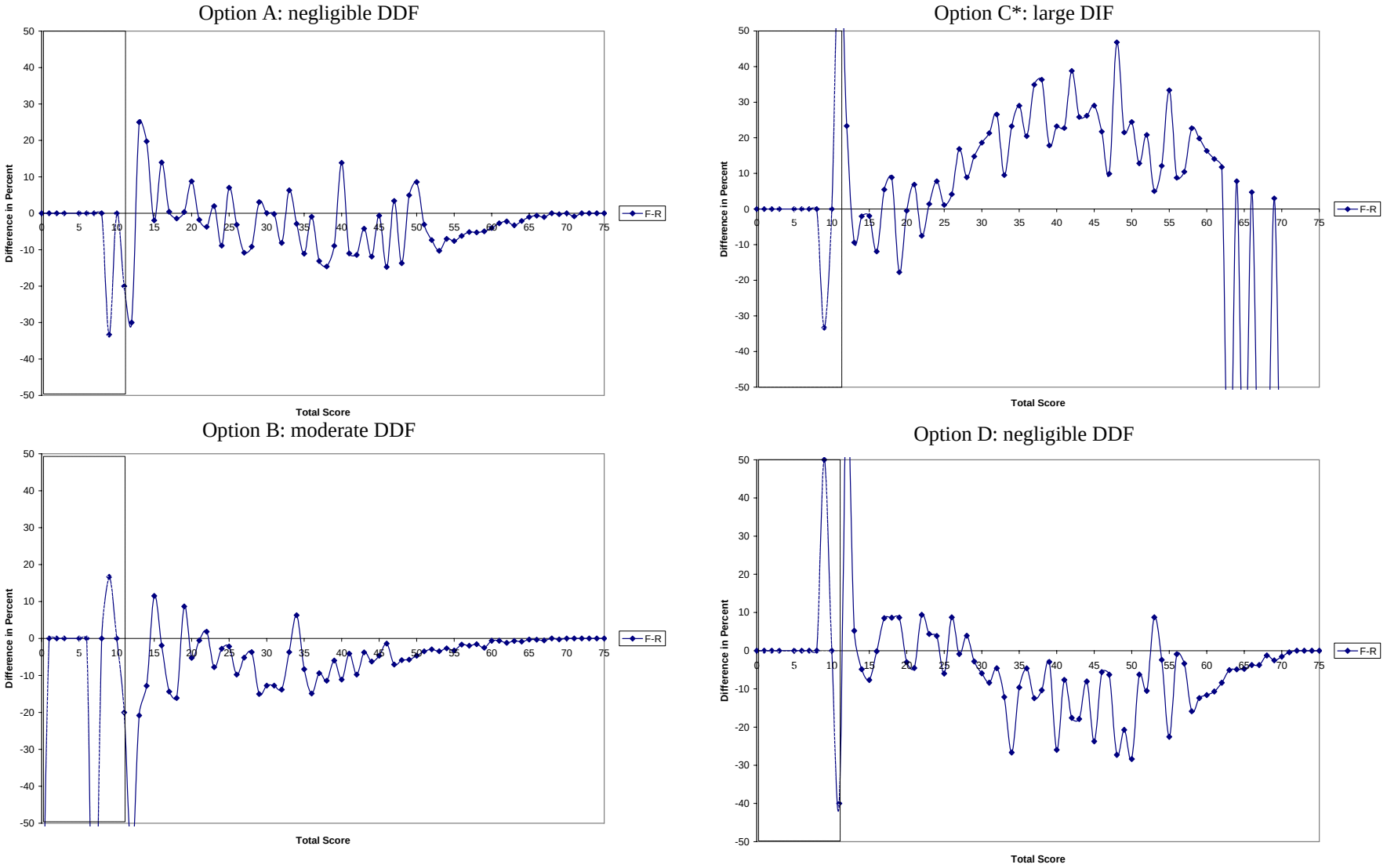


Figure A9. **Item 34 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

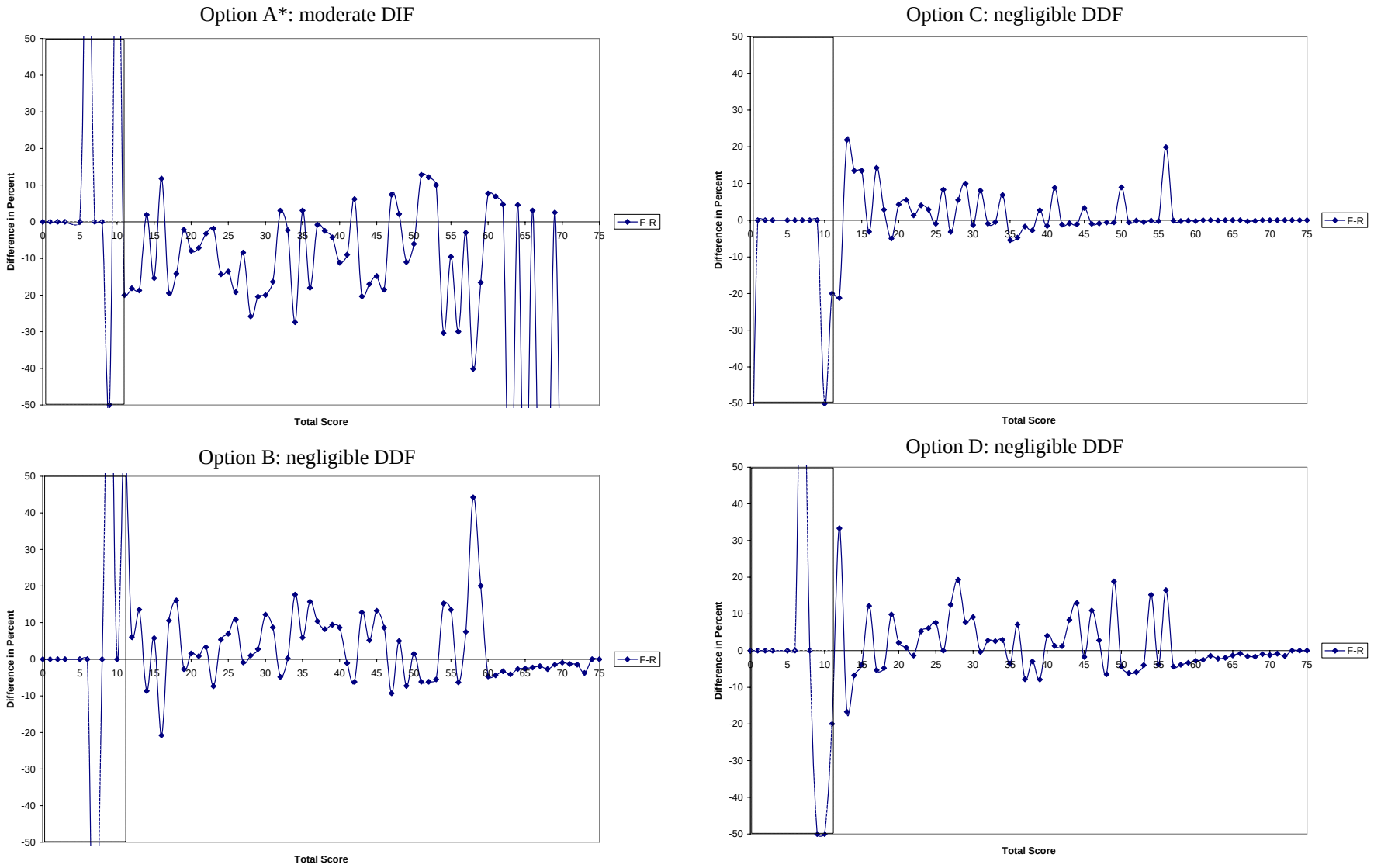


Figure A10. **Item 45 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

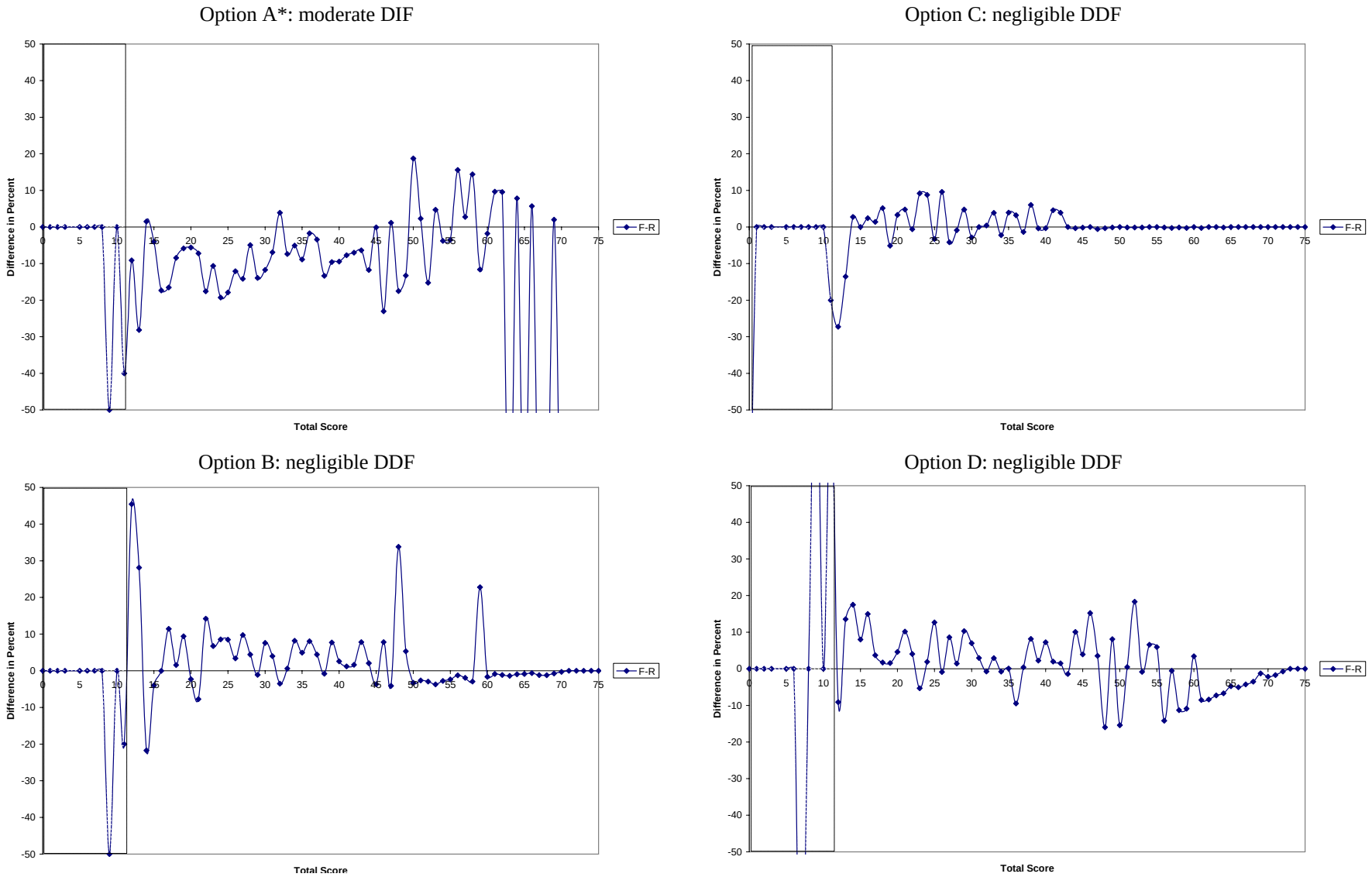


Figure A11. **Item 56 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

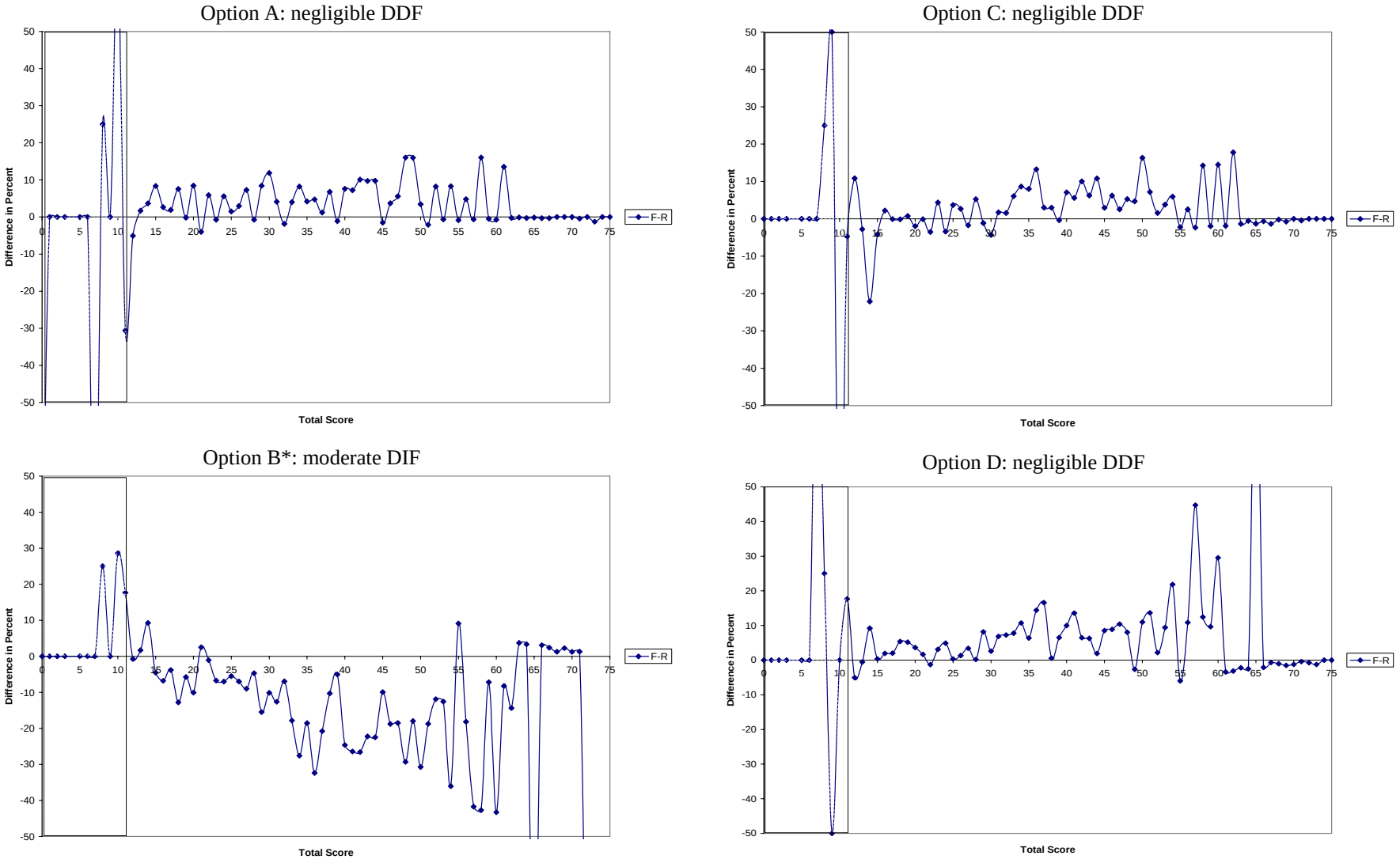
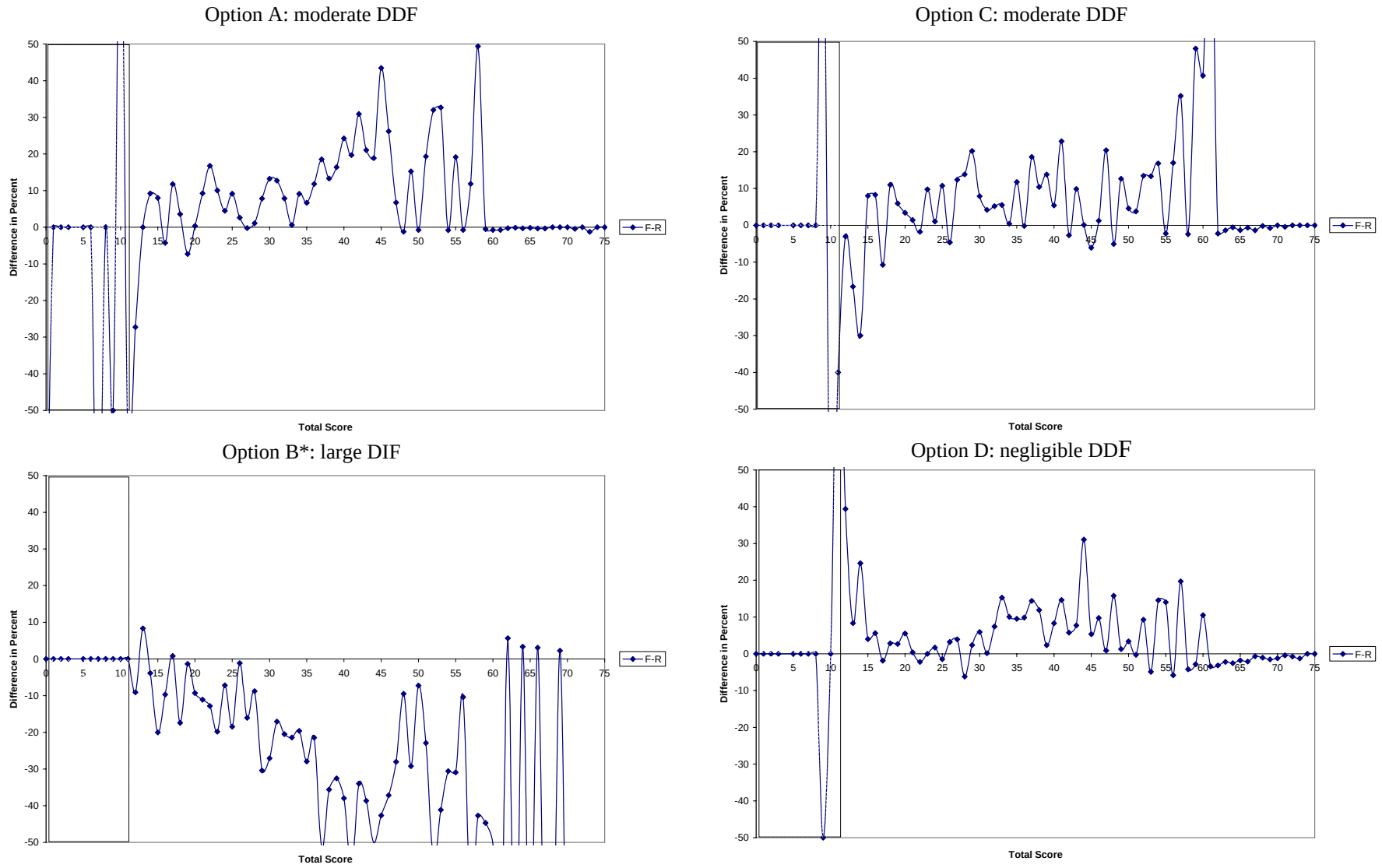


Figure A12. **Item 64 differences (Group 0-21).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.



Figurer A13. **Item 64 differences (Group 0-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

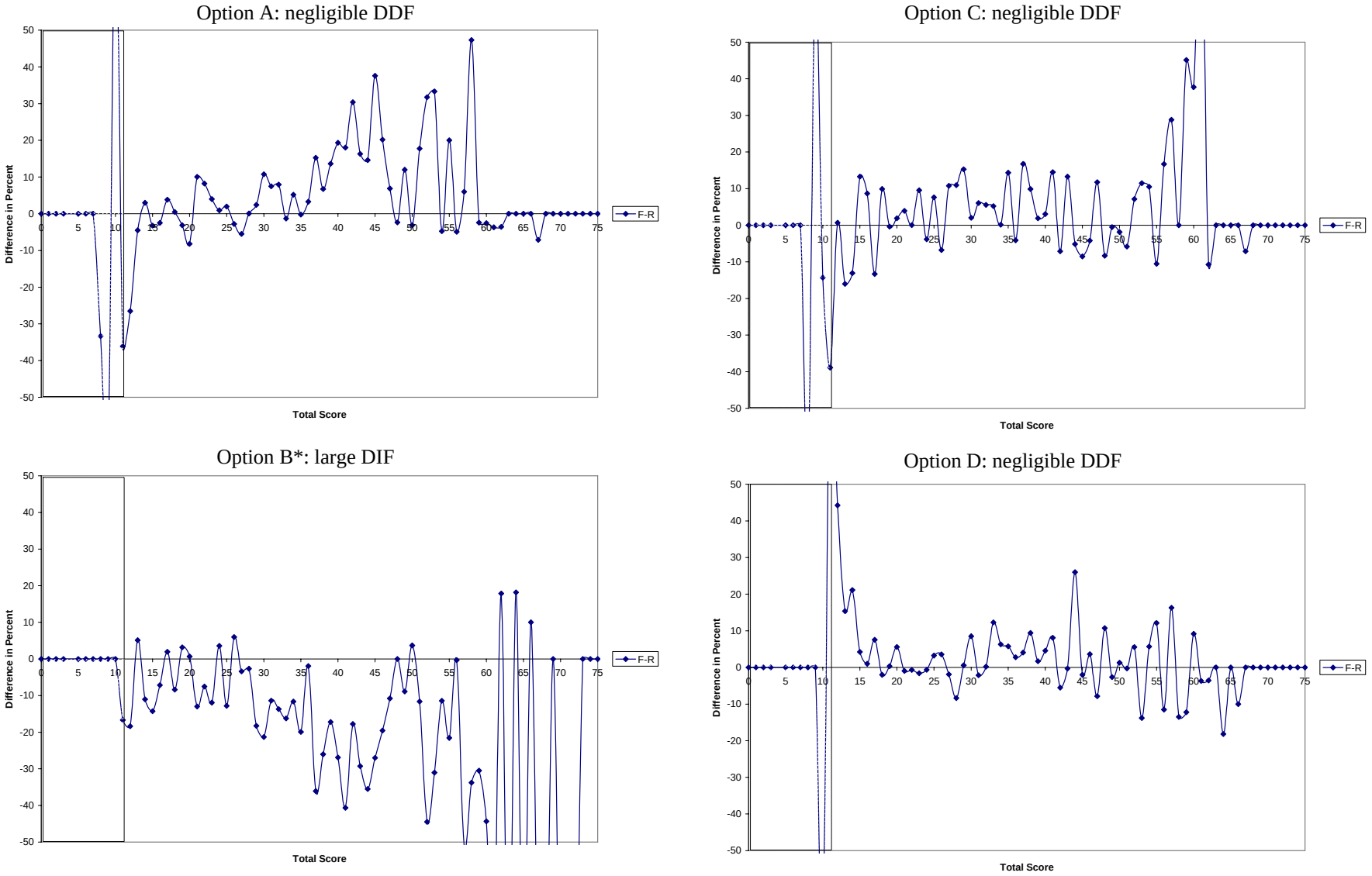


Figure A14. **Item 64 differences (Group 20-22).**

Note. Negative values for keyed response favor reference, positive values favor focal. Negative values for distractor responses favor focal, positive values favor reference. Correct option indicated by an asterisk.

Appendix B

DIF and DDF Values by Item

	Comparison groups																			
	0-20				0-21				0-22				20-21				20-22			
Item	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D	A	B	C	D
3	**0.07**	-	-	-	**0.08**	-	-	-												
10									0.04	*0.06*	0.03	-					0.03	*0.05*	0.04	-
13									*0.07*	*0.08*	-	-								
25									-	0	*0.13*	-								
32									-	0.01	**0.09**	-								
33									0.03	-	*0.06*	0.01								
34									0.02	*0.06*	-	0.03								
45									**0.09**	-	-	-								
56									**0.09**	-	-	-								
64					-	**0.09**	-	-	-	**0.19**	-	-					-	**0.11**	-	-

Note. Numbers in bold indicate the correct option that exhibits DIF. (Positive values indicate DIF in favor of the reference group; negative values indicate DIF in favor of the focal group.) Numbers in italics indicate the incorrect option that exhibits DDF. (Negative values indicate DDF in favor of the reference group; positive values indicate DDF in favor of the focal group.) Shaded portion represents items that did not show DIF for the groups compared.