

Differential Bundle Functioning on Mathematics and Science Achievement Tests: A Small Step Toward Understanding Differential Performance

Keith A. Boughton¹
Mark J. Gierl
Shameem Nyla Khaliq

Centre for Research in Applied Measurement and Evaluation
University of Alberta

Paper Presented at the Annual Meeting of the
Canadian Society for Studies in Education (CSSE)
Edmonton, Alberta, Canada
May 24 – 27, 2000

¹This paper can also be downloaded from the Centre for Research in Applied Measurement and Evaluation (CRAME) website: <http://www.education.ualberta.ca/educ/psych/crame/>. This research was supported with funds to the second author by the Social Sciences and Humanities Research Council of Canada.

Differential Bundle Functioning on Mathematics and Science Achievement Tests: A Small Step Toward Understanding Differential Performance

Almost all of the provinces and territories in Canada have an educational assessment program. These assessment programs are used to evaluate student performance on well-defined learning objectives put forth by the educational departments or ministries in each province. Alberta Learning, for example, measures student achievement in Grade 3, 6, and 9 in the content areas of Language Arts, Mathematics, Science, and Social Studies. Teachers are encouraged to include these scores for grading students and school results are frequently reported in local newspapers. In some cases, these tests account for a substantial percentage of a student's final course Grade. In British Columbia, the Grade 12 provincial examinations include over 19 subject areas (the highest found in Canada) with the exams being worth 40% of the student's final course grade. In Alberta, diploma examinations are administered in 11 subject areas at the Grade 12 level and account for 50% of the student's final course grade. The results of these tests are used to ensure that the province-wide standards are being met (Lafleur & Ireland, 1999). Universities also use these provincial exam scores for selection purposes. It is here where a single test score can have an important effect on a student since these scores contribute to the selection decisions at many universities. Consequently, test developers must ensure that their examinations are fair for all students.

Why would one group of examinees, such as Native, Female, or French examinees, receive a lower score on a particular item or group of items when these groups are matched on ability with Non-native, Male, or English examinees? Is it because one group lacks the sufficient knowledge to answer the item(s) or is there some unintended construct being measured by the item(s) (i.e., multidimensionality)? For example, if a test is said to measure science achievement and males and females, matched on science ability, significantly differ in their correct response to a particular item or a collection of items, then, it may be possible that the item(s) are biased towards one of the groups (e.g., either males or females). This outcome, if it occurs, calls into question the validity of the test and any inferences about group differences made from the test scores. To address this concern, psychometricians developed statistical procedures to screen items for potential bias. Note, however, that these statistical procedures can only flag items that function differentially between two groups and cannot indicate whether or not the item is biased or if the item shows impact (i.e., the item has

identified an outcome associated with a real group differences in ability between the two groups). Thus, one cannot tell using just statistical outcomes whether an individual item displays bias or impact.

The purpose of this paper is to summarize the research that the authors conducted last year and then to state the new direction the research has taken. Examples and comparisons will be used to demonstrate this new line of research as we seek to identify causes of differential group performance using a new psychometric technique called differential bundle functioning (DBF). The benefits of DBF analyses will be demonstrated using real data for both the Mathematics and Science achievement tests using the table of specifications (also called test specifications or blueprints) to structure the data. The purpose of this study is not to evaluate the tests or to suggest removal of any items that may be functioning differently, but to illuminate possible areas within the content blueprints that may result in differential performance for males and females and to illustrate a new statistical technique for identifying these differences.

Summary of Previous Research

The paper entitled, "Gender Differential Item Functioning in Mathematics and Science: Prevalence and Policy Implications" (Gierl, Khaliq, & Boughton, 1999) was presented last year at the annual meeting of the Canadian Society for the Study of Education. The study focused on differential item functioning or DIF (i.e., item-level analyses) in both mathematics and science. The purpose of our previous research was threefold: (a) to describe the testing program in the Student Evaluation Branch at Alberta Learning in regards to their fairness standards and protocols during test development, (b) to compare the consistency of three DIF statistical procedures used to detect gender DIF for both the Mathematics and Science provincial examinations in Grades 3, 6, and 9, and (c) to present policy implications based on our findings.

It was reported that the three statistical procedures (i.e., Simultaneous Item Bias Test, Mantel-Haenszel, and Logistic Regression; see also Gierl et al., 1999) used for identifying DIF items were relatively consistent, although not identical. SIBTEST flagged more items and is thus considered, in our view, to be the most liberal method for identifying DIF. Also, there were more DIF items found in Science than in Mathematics. However, when it came to substantively classifying a single item as

displaying bias or impact, the results were not conclusive even when the verbal protocols of students solving DIF items were analyzed. This finding was consistent with past research: Namely, that identifying and explaining the causes of DIF using judgmental methods have been ineffective (e.g., Angoff, 1993; Bond, 1993; Camilli & Shepard, 1994; Englehard, Hansche, & Rutledge, 1990; Gierl & McEwen, 1998; O'Neil & McPeck, 1993; Plake, 1980; Rengel, 1986; Sandoval & Miille, 1980). In fact, it was at this point that our research was stalled. To overcome this problem, the research team moved from the item level to the subtest or bundle level (i.e., we moved from studying DIF to DBF). Our hope was that we might be able to find reasons for group differences on a set of items found within the test specifications that might prove more useful in explaining why a bundle of items functions differentially between two groups matched on ability (Douglas, Roussos, & Stout, 1996). In effect, one should be able to more easily define the commonalities between a set of items chosen on some organizing principle, rather than trying to explain single-item outcomes.

Differential Item Functioning vs. Bundle Functioning

Researchers in the past have focused on item-level analyses. DIF is one of many examples. But with this focus, a considerable amount of confusion has also arisen when trying to interpret or explain why an item displays statistical DIF. DIF is a statistical concept without substantive meaning. Hence, items can be detected statistically using SIBTEST, for example, but these items still need a substantive interpretation in order to be classified as bias or impact. If the item is biased, then it must be revised or removed from the test. If the item displays impact, then further research is needed to explore why one group has a higher ability in the context of the item or group of items under investigation. Herein lies the problem of single-item DIF—when reviewing a single item, there are usually numerous hypotheses that can be generated as to why an item functions differentially because the logic applied to this analysis requires a statistically defined item to be interpreted substantively. Often, the results are inconclusive as reasons cannot be found to account for DIF on these statistically flagged items. To overcome this problem, Stout and Roussos (1999) stated:

In order to achieve a maximally statistically effective and substantively informative DIF/DBF analysis, the essential step in augmenting, indeed sometimes even replacing, a standard one at a time DIF analysis is to select bundles judged to be substantively homogeneously

and/or statistically dimensionally homogenous and then to analyze each selected bundle for DBF. When a homogeneous bundle is found to display DBF, it is often possible to reliably provide a substantive explanation for why the DBF has occurred (p. 3).

Here the logic changes: Instead of interpreting statistical outcomes, substantive hypotheses are first generated and then tested statistically. This logic is based on the DIF analysis paradigm, as described in the next section.

DIF Analysis Paradigm

The DIF analysis paradigm is based on Shealy and Stout's (1993) multidimensional model of DIF (MMD). MMD is a framework for understanding how DIF occurs. It is based on the assumption that multidimensionality produces DIF. Dimension refers to a substantive characteristic of an item that can affect the probability of a correct response on the item. The main construct that the test is intended to measure is the primary dimension. DIF items measure at least one dimension in addition to the primary dimension (Ackerman, 1992; Roussos & Stout, 1996a; Shealy & Stout, 1993). The addition of dimensions that produce DIF are referred to as the secondary dimensions. When primary and secondary dimensions characterize item responses, the data are considered multidimensional. The secondary dimensions are interpreted further. The secondary dimensions are considered auxiliary if they are intentionally assessed as part of the construct on the test. Alternatively, the secondary dimensions are considered nuisance if they are unintentionally assessed as part of the construct on the test. DIF that is caused by auxiliary dimensions is benign whereas DIF that is caused by nuisance dimensions is adverse.

Substantive DIF Analysis

The Roussos-Stout DIF analysis paradigm is built on the foundation provided by MMD. The first stage is a substantive analysis where DIF hypotheses are generated. The DIF hypothesis specifies whether an item or bundle designed to measure the primary dimension also measures a secondary dimension, thereby producing DIF, for examinees in either the reference or the focal group. The reference group is the majority group or the group to whom the focal group is compared. The focal group is the minority group or the particular group of interest in the DIF analysis. Roussos and Stout (1996a) contend that MMD can be used to design DIF-free items if test developers can generate

accurate DIF hypotheses based on their understanding of the underlying dimensional structure of the test data. This understanding may be enhanced by studying outcomes from previous DIF analyses, analyzing existing test data, formulating DIF hypotheses with content reviews, and testing bundles of items.

Statistical DIF Analysis

The second stage in the Roussos-Stout multidimensionality-based DIF analysis paradigm is statistical testing of the DIF hypotheses. The Simultaneous Item Bias Test (SIBTEST) can be used to test DIF hypotheses and quantify the size of DIF. With this statistical approach, the complete latent space is viewed as multidimensional, (Θ, η) , where Θ is the primary dimension and η is the secondary dimension. SIBTEST is designed to identify items or bundles in the secondary dimension. The statistical hypothesis tested by SIBTEST is:

$$H_0: \mathbf{b}_{UNI} = 0 \text{ vs. } H_1: \mathbf{b}_{UNI} \neq 0,$$

where \mathbf{b}_{UNI} is the parameter specifying the amount of DIF for an item or bundle. \mathbf{b}_{UNI} is defined as

$$\mathbf{b}_{UNI} = \int [P(\Theta, R) - P(\Theta, F)] f_F(\Theta) d\Theta,$$

where $P(\Theta, R) - P(\Theta, F)$, the difference in the probabilities of correct response for examinees from the reference and focal groups, respectively, conditional on Θ , $f_F(\Theta)$ is the density function for Θ in the focal group, and d is a scaling constant. \mathbf{b}_{UNI} is integrated over Θ to produce a weighted expected score difference between reference and focal group examinees of the same ability on an item or bundle. To operationalize this test, items on the exam are divided into the studied subtest and the matching (or sometimes called "valid") subtest. The studied subtest contains the item or bundle believed to measure the primary and secondary dimensions whereas the matching subtest contains the items believed to measure only the primary dimension. The matching subtest places the reference and focal group examinees into subgroups at each score level so their performances on items from the studied subtest can be compared.

To estimate \mathbf{b}_{UNI} , the weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the K subgroups is calculated by

$$\widehat{\mathbf{b}}_{UNI} = \sum_{k=0}^K p_k (\overline{Y_{Rk}^*} - \overline{Y_{Fk}^*}).$$

In this equation, p_k is the proportion of focal group examinees in subgroup k and $\overline{Y_{Rk}^*} - \overline{Y_{Fk}^*}$ is the difference in the adjusted means on the studied subtest item or bundle for examinees in the reference and focal groups, respectively, in each subgroup k . The means on the studied subtest item or bundle are adjusted to correct for any differences in the ability distributions of the reference and focal groups using a regression correction described in Shealy and Stout (1993). SIBTEST yields an overall statistical test for $\widehat{\mathbf{b}}_{UNI}$ that has a normal distribution with mean 0 and variance 1 under the null hypothesis of no DIF. If the alternative hypothesis is true, the null hypothesis is rejected if the test statistic exceeds the 100 $(1 - \alpha / 2)$ percentile point from the normal distribution using a non-directional hypothesis test. A statistically significant value of $\widehat{\mathbf{b}}_{UNI}$ that is positive indicates DIF against the focal group whereas a negative value indicates DIF against the reference group.

To aid in the interpretation of $\widehat{\mathbf{b}}_{UNI}$, Roussos and Stout (1996b, p. 220) proposed the following values for classifying DIF on a single item: (a) Negligible or A-level DIF: Null hypothesis is rejected and the absolute value of $\widehat{\mathbf{b}}_{UNI} < 0.059$, (b) Moderate or B-level DIF: Null hypothesis is rejected and $0.059 \leq |\widehat{\mathbf{b}}_{UNI}| < 0.088$, and (c) Large or C-level DIF: Null hypothesis is rejected and $|\widehat{\mathbf{b}}_{UNI}| \geq 0.088$. No comparable guidelines exist for classifying $\widehat{\mathbf{b}}_{UNI}$ on bundles.

Differential Bundle Functioning

Douglas et al. (1996) demonstrated how DBF “amplification” is a fundamental premise for studying test fairness at the bundle level. They suggested that DIF at the item level but in smaller quantities (e.g., several small or “A-level” items) that may go statistically undetected in the single item approach but can be detected using the DBF approach. It is only with the bundle approach that items may be gathered together and statistically tested as a group. An example is presented next to help clarify the issue surrounding DBF amplification.

Imagine a reading comprehension test containing a paragraph-based bundle of items concerning American professional football. If females are the focal group and males are the reference group, one might suspect that most of the items would display DIF in favor of males. However, the amount of DIF present in any single item could be quite small and thus be difficult to detect statistically. Nonetheless, over several items, small amounts of DIF can add up to an unacceptable level of DBF at the bundle level—in other words, an unacceptable level of DBF. This phenomenon is referred to as DIF amplification (Douglas et al., 1996, p. 468).

Nandakumar (1993) has studied SIBTEST's role in the detection of simultaneous "amplification" and "cancellation" and she found greater statistical power with the bundle approach compared to the single-item DIF approach. Amplification is caused by items acting in concert and each contributing to an unacceptable level of DBF. Cancellation, on the other hand, is caused by a bundle of items exhibiting DIF against one group while another bundle of items exhibits DIF against the alternate group and therefore, each is canceled out. A crucial and yet unresolved issue is in the creation of a valid subtest against which the suspect subtest items will be compared. There is an assumption that must be made when using these statistical procedures: Carefully constructed tests should have a majority of items that will be content and construct valid. To the extent that this holds true, a valid subtest can be found within the test and used to match examinees at different ability levels.

Method

The achievement tests in this study included the Grade 6 Mathematics test administered in 1996 and 1997 and the Grade 9 Mathematics tests administered in 1995 and 1996. Grade 6 and 9 Science achievement tests administered in 1997 and 1998 were also analyzed. The Mathematics tests for Grades 6 and 9 consisted of 50 multiple-choice items. The science test for Grade 6 (1997, 1998) consisted of 50 multiple-choice items and Science 9 (1997, 1998) contained 65 multiple-choice items. Overall, there were eight different tests analyzed in this study. Content areas from the test specifications for each test were used to form bundles that could be statistically tested for differential functioning. An example of the test specification from the Grade 6 Science achievement

test is presented in Table 1. Data from eight different randomly-selected students samples of 6000 males and 6000 females were analyzed.

For all DBF analyses, the content areas outlined in the test specifications were used to form bundles. For Grade 6 Mathematics, the test items were classified into five content areas: Numeration, operations and properties, measurement, geometry, and data management. The items from the Grade 9 Mathematics achievement test were also classified into five content areas: Number systems and operations, ratio and proportions, measurement and geometry, data management, and algebra. For Grade 6 Science, the test items were classified into five content areas: Evidence and investigation, air and aerodynamics, sky science, observation and inference, and trees and environment. The items from the Grade 9 Science test were classified into six content areas: Diversity of living things, fluids and pressure, heat energy: transfer and conversation, electromagnetic systems, chemical properties and changes, and environmental quality. These tests were specifically chosen because each subject-specific test (i.e., 1996 and 1997 Grade 9 Science) had the same content categories. This is fundamental because our screening process for bundle selection involves across year comparisons.

SIBTEST (Stout & Roussos, 1999) was used to identify and determine which bundles of items displayed DBF. However, before any of the bundles were statistically tested, a strict screening process was implemented. First, all of the DIF items were flagged using SIBTEST with a single-item analysis. Specifically, this means that each item was tested against all of the remaining items in the test. Then, plots of the beta-uni statistical indices were prepared by content area and reviewed for groups of items that cluster on the male or female side. If a bundle of items within a specific content domain favored one of the groups, then a reliability check was done by reviewing bundle analyses using data from the following year. The bundles had to meet this two-stage approach in order to be tested statistically. In order to be deemed practically significant, a content area needed to be flagged across two years and should produce a statistically significant effect.

Results

The descriptive statistics for the mathematics and science tests are presented in Tables 2 through 5. The mean total test score demonstrates that, for Grade 6 and 9 mathematics, males do

slightly better than females. For the science tests, the means were also slightly higher for the males compared to females across both administrations in Grade 6 and 9.

The first stage of the bundle analyses revealed for Grade 6 Mathematics (see Figures 1 and 2), that the content area called **measurement** slightly favored males across both years (1996 and 1997). As stated earlier, this was the first condition that must be satisfied in order to proceed with the statistical bundle analyses. Thus, the six items in the **measurement** bundle for Grade 6 1996 and the eight items in the 1997 bundle were tested separately using a directional hypothesis test in favor of males. This test resulted in a significant beta-uni of 0.185 for 1996 and 0.300 for 1997, $p < .01$ (see Table 6).

The first stage of the bundle analyses revealed for Grade 9 Mathematics (see Figures 3 and 4), that the content area called **algebra** favored females across both years (1995 and 1996). Thus, the 12 items in the **algebra** bundle for Grade 9 1995 and the 11 items in the 1996 bundle were tested separately using a directional hypothesis test in favor of females. This test yielded a significant beta-uni of -0.279 for 1995 and -0.412 for 1996, $p < .01$ (see Table 7). Overall, two bundles were classified statistically as favoring females on the mathematics tests.

The bundle analyses revealed for Grade 6 Science (see Figures 5 and 6) that the content area called **air and aerodynamics** favored males across both years (1997 and 1998). Thus, the 13 items in the **air and aerodynamics** bundle for Grade 6 1997 and the 15 items in the 1998 bundle were tested separately using a directional hypothesis test in favor of males. This test yielded a significant beta-uni of 0.423 for 1997 and 0.500 for 1998, $p < .01$ (see Table 8). Within these two tests there was also a bundle that favored females related to the content area called **observations and inferences**. There were nine items in the **observations and inferences** bundle for Grade 6 1997 and nine items in the 1998 bundle that were tested separately using a directional hypothesis test in favor of females. This test yielded a significant beta-uni of -0.440 for 1997 and -0.324 for 1998, $p < .01$ (see Table 8).

The bundle analyses revealed for Grade 9 Science (see Figures 7 and 8) that the content area called **fluids and pressure** favored males across both years (1997 and 1998). Thus, the 10 items in the **fluids and pressure** bundle for Grade 9 1997 and the 9 items in the 1998 bundle were tested separately using a directional hypothesis test in favor of males. It yielded a significant beta-uni of

0.181 for 1997 and 0.222 for 1998, $p < .01$ (see Table 9). For this test there was also two bundles that favored females related to the content areas called **diversity of living things** and **environmental quality**. There were eight items in the **diversity of living things** bundle for Grade 9 1997 and the 10 items in the 1998 bundle that were tested separately using a directional hypothesis test in favor of females which yielded a significant beta-uni of -0.213 for 1997 and -0.340 for 1998, $p < .01$ (see Table 9). There were 12 items in the **environmental quality** bundle for Grade 9 1997 and the nine items in the 1998 bundle that were tested separately using a directional hypothesis test in favor of females which yielded a significant beta-uni of -0.264 for 1997 and -0.222 for 1998, $p < .01$ (see Table 9). Overall, three bundles were classified statistically as favoring males with four bundles favoring females for the science tests.

Discussion and Conclusions

The purpose of this paper was to summarize the research that the authors of this article conducted last year and to outline the new direction our research has taken. Examples and comparisons were used to demonstrate our new line of research. Instead of conducting item-level analyses, groups of items called bundles were used to conduct differential bundle functioning analyses. The usefulness of DBF analyses were demonstrated using real data from the Alberta Learning Mathematics and Science achievement tests. Alberta Learning's test development process uses a table of specifications consisting of content areas and cognitive levels. SIBTEST was used to help us understand which content areas may favor males or females in Mathematics and Science by bundling and testing the items using content area clusters. The rationale for this approach is based on the Roussos-Stout DIF analysis paradigm. Our intent was not to evaluate the tests or to suggest removal of any items that may be functioning differently, but only to illuminate possible areas within the content blueprints that may be functioning differently and thus exposing multidimensionality within the tests. Systematic group differences were found in Mathematics and Science. Males consistently outperformed females in the mathematics content areas of measurement and algebra. Males also outperformed females in the science content areas of air and aerodynamics and fluids and pressures. Females, on the other hand, outperformed males in the science content areas of

observation and inference, diversity of living things, and environmental quality. These differences were stable and consistent over a two-year period.

The next phase of our research will be conducted by content specialists in an attempt to illuminate why some bundles of items favored one group over another. These bundles of items will then have had to pass a three-stage process (a) across year consistency, (b) followed by a statistical analysis, and (c) content review by specialists . If a group of items passes all three stages, then we should be able to predict items that will display DIF in subsequent administrations. This demonstration will show that we understand (and therefore can predict) the factors that produce differential performance. Our prediction study is currently underway.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. Journal of Educational Measurement, *29*, 67-91.
- Angoff, W. (1993). Perspective on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum.
- Bond, L. (1993). Comments on the O'Neill and McPeck paper. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 277-279). Hillsdale, NJ: Lawrence Erlbaum.
- Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. Newbury Park, CA: Sage.
- 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*, 465-484.
- Englehard, G., Hansche, L., & Rutledge, K. E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. Applied Measurement in Education, *3*, 347-360.
- Gierl, M. J., & McEwen, N. (1998, May). Differential item functioning on the Alberta Education Social Studies 30 Diploma Exams. Paper presented at the annual meeting of the Canadian Society for Studies in Education, Ottawa, Ontario, Canada.
- Gierl, M. J., Khaliq, S. N., & Boughton, K. A. (1999). Differential item functioning on Alberta Education mathematics achievement tests: Prevalence and policy implications. In X. Ma (chair), Improving Large-Scale Assessment in Education Research. Symposium conducted at the annual meeting of the Canadian Society for Studies in Education, Sherbrooke, Quebec, Canada.
- Lafleur, C., & Ireland, D. (1999). Canadian and provincial approaches to learning assessments and educational performance indicators. Paper submitted to the Canadian International Development Agency.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. Journal of Educational Measurement, *30*, 293-311.
- O'Neill, K. A., & McPeck, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), Differential item functioning (pp. 255-276). Hillsdale, NJ: Lawrence Erlbaum.
- Plake, B. S. (1980). A comparison of a statistical and subjective procedure to ascertain item validity: One step in the validation process. Educational and Psychological Measurement, *40*, 397-404.
- Rengel, E. (1986, August). Agreement between statistical and judgmental item bias methods. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. Applied Psychological Measurement, 20, 355-371.

Roussos, L. A., & Stout, W. F. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. Journal of Educational Measurement, 33, 215-230.

Sandoval, J., & Mille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. Journal of Consulting and Clinical Psychology, 48, 249-253.

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

Stout, W., & Roussos, L. (1999). Dimensionality-based DIF/DBF Package [computer program]. William Stout Institute for Measurement: University of Illinois.

Table 1

Blueprints for Science Grade 6 1997 and 1998

Content Area	Coverage
<p style="text-align: center;"><u>Evidence and Investigation</u></p> <p>Work cooperatively with others to design and carry out an investigation in which variables are identified and controlled; and recognize the importance of accuracy in observation and measurement, and apply suitable methods to record, compile, interpret, and evaluate observations and measurements gathered by self and group; and work cooperatively with others in designing and carrying out an investigation.</p>	11 Items (22%)
<p style="text-align: center;"><u>Air and Aerodynamics</u></p> <p>Describe properties of air, and the interactions of air with objects in flight and construct devices that move through air, and identify adaptations for controlling flight.</p>	13 Items (26%)
<p style="text-align: center;"><u>Sky Science</u></p> <p>Observe, describe, and interpret the movement of objects in the sky, and identify pattern and order in these movements.</p>	6 Items (12%)
<p style="text-align: center;"><u>Observation and Inference</u></p> <p>Apply observation and inference skills to recognize and interpret patterns, and to distinguish a specific pattern from among a group of similar patterns, and apply a knowledge of the properties and interactions of materials to the investigation and identification of a material sample.</p>	10 Items (20%)
<p style="text-align: center;"><u>Trees and Environment</u></p> <p>Describe characteristics of trees and the interaction of trees with other living things in the local environment.</p>	10 Items (20%)
Total	50 Items (100%)

Table 2

Descriptive Statistics for Grade 6 Mathematics Achievement Tests

Characteristic	1996		1997	
	Males	Females	Males	Females
No. of Examinees	6000	6000	6000	6000
No. of Items	50	50	50	50
Mean	37.56	36.59	35.93	34.26
Standard Deviation	8.31	8.20	8.37	8.46
Skewness	-0.81	-0.71	-0.52	-0.36
Kurtosis	0.04	-0.03	-0.41	-0.61

Table 3

Descriptive Statistics for Grade 9 Mathematics Achievement Tests

Characteristic	1995		1996	
	Males	Females	Males	Females
No. of Examinees	6000	6000	6000	6000
No. of Items	40	40	45	45
Mean	30.51	28.75	34.70	33.14
Standard Deviation	9.90	9.88	10.90	11.02
Skewness	-0.24	-0.31	-0.34	-0.18
Kurtosis	-0.72	-0.88	-0.73	-0.89

Table 4

Descriptive Statistics for Grade 6 Science Achievement Tests

Characteristic	1997		1998	
	Males	Females	Males	Females
No. of Examinees	6000	6000	6000	6000
No. of Items	50	50	50	50
Mean	32.23	30.54	34.94	33.66
Standard Deviation	7.64	7.40	8.15	8.21
Skewness	-0.42	-0.27	-0.66	-0.47
Kurtosis	-0.22	-0.37	0.03	-0.33

Table 5

Descriptive Statistics for Grade 9 Science Achievement Tests

Characteristic	1997		1998	
	Males	Females	Males	Females
No. of Examinees	6000	6000	6000	6000
No. of Items	55	55	55	55
Mean	38.24	36.11	37.89	35.49
Standard Deviation	9.06	9.13	8.70	8.89
Skewness	-0.64	-0.40	-0.63	-0.40
Kurtosis	-0.02	-0.43	0.08	-0.35

Table 6

Differential Bundle Functioning Results for the Grade 6 1996 and 1997 Mathematics Achievement

Bundle	No. of Items	Beta-Uni	Favors
<u>1996</u>			
Measurement	6	0.185*	Males
<u>1997</u>			
Measurement	8	0.300*	Males

* $p < .01$.Note.

The matching subtest used in each year was created by combining item from the content areas 1, 2, 4, and 5 with the exception of items displaying C-level DIF.

Table 7

Differential Bundle Functioning Results for the Grade 9 1995 and 1996 Mathematics Achievement

Bundle	No. of Items	Beta-Uni	Favors
<u>1995</u>			
Algebra	12	-0.279*	Females
<u>1996</u>			
Algebra	11	-0.412*	Females

* $p < .01$.Note.

The matching subtest used in each year was created by combining item from the content areas 1, 2, 3, and 4 with the exception of items displaying C-level DIF.

Table 8

Differential Bundle Functioning Results for the Grade 6 1997 and 1998 Science Achievement

Bundle	No. of Items	Beta-Uni	Favors
<u>1997</u>			
Air and Aerodynamics	13	0.423*	Males
Observations and Inferences	9	-0.440*	Females
<u>1998</u>			
Air and Aerodynamics	15	0.500*	Males
Observations and Inferences	9	-0.324	Females

*p<.01.

Note.

The matching subtest used in each year was created by combining item from the content areas 1, 3, and 5 with the exception of items displaying C-level DIF.

Table 9

Differential Bundle Functioning Results for the Grade 9 1997 and 1998 Science Achievement

Bundle	No. of Items	Beta-Uni	Favors
<u>1997</u>			
Diversity of Living Things	8	-0.213*	Females
Fluids and Pressure	10	0.181*	Males
Environmental Quality	12	-0.264*	Females
<u>1998</u>			
Diversity of Living Things	10	-0.340*	Females
Fluids and Pressure	9	0.222*	Males
Environmental Quality	9	-0.222*	Females

*p<.01.

Note.

The matching subtest used in each year was created by combining item from the content areas 3, 4, and 5 with the exception of items displaying C-level DIF.

Figures

Figure 1. A plot of all of the mathematics items in Grade 6 for 1996 by content area

Figure 2. A plot of all of the mathematics items in Grade 6 for 1997 by content area

Figure 3. A plot of all of the mathematics items in Grade 9 for 1995 by content area

Figure 4. A plot of all of the mathematics items in Grade 9 for 1996 by content area

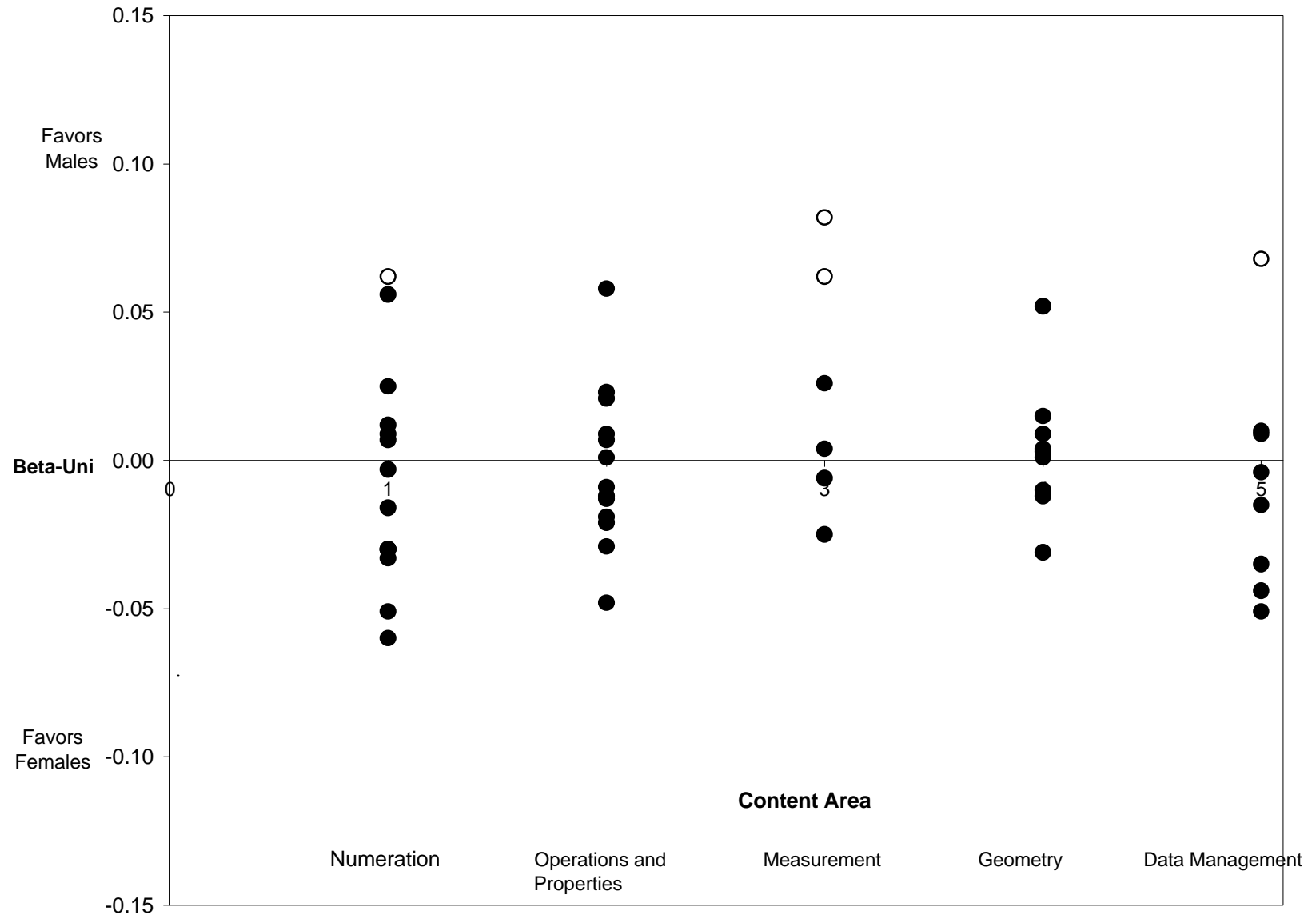
Figure 5. A plot of all of the science items in Grade 6 for 1997 by content area

Figure 6. A plot of all of the science items in Grade 6 for 1998 by content area

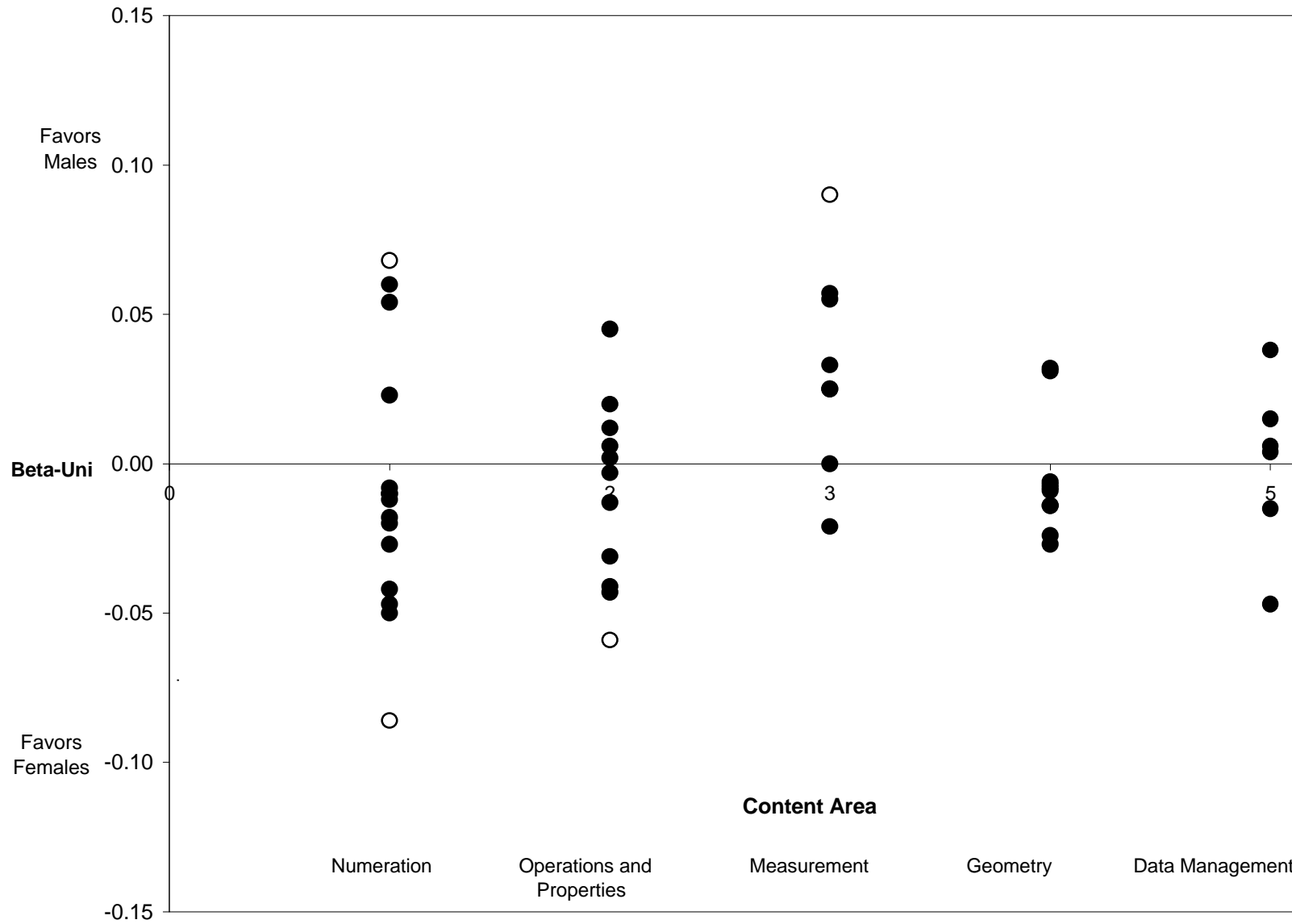
Figure 7. A plot of all of the science items in Grade 9 for 1997 by content area

Figure 8. A plot of all of the science items in Grade 9 for 1998 by content area

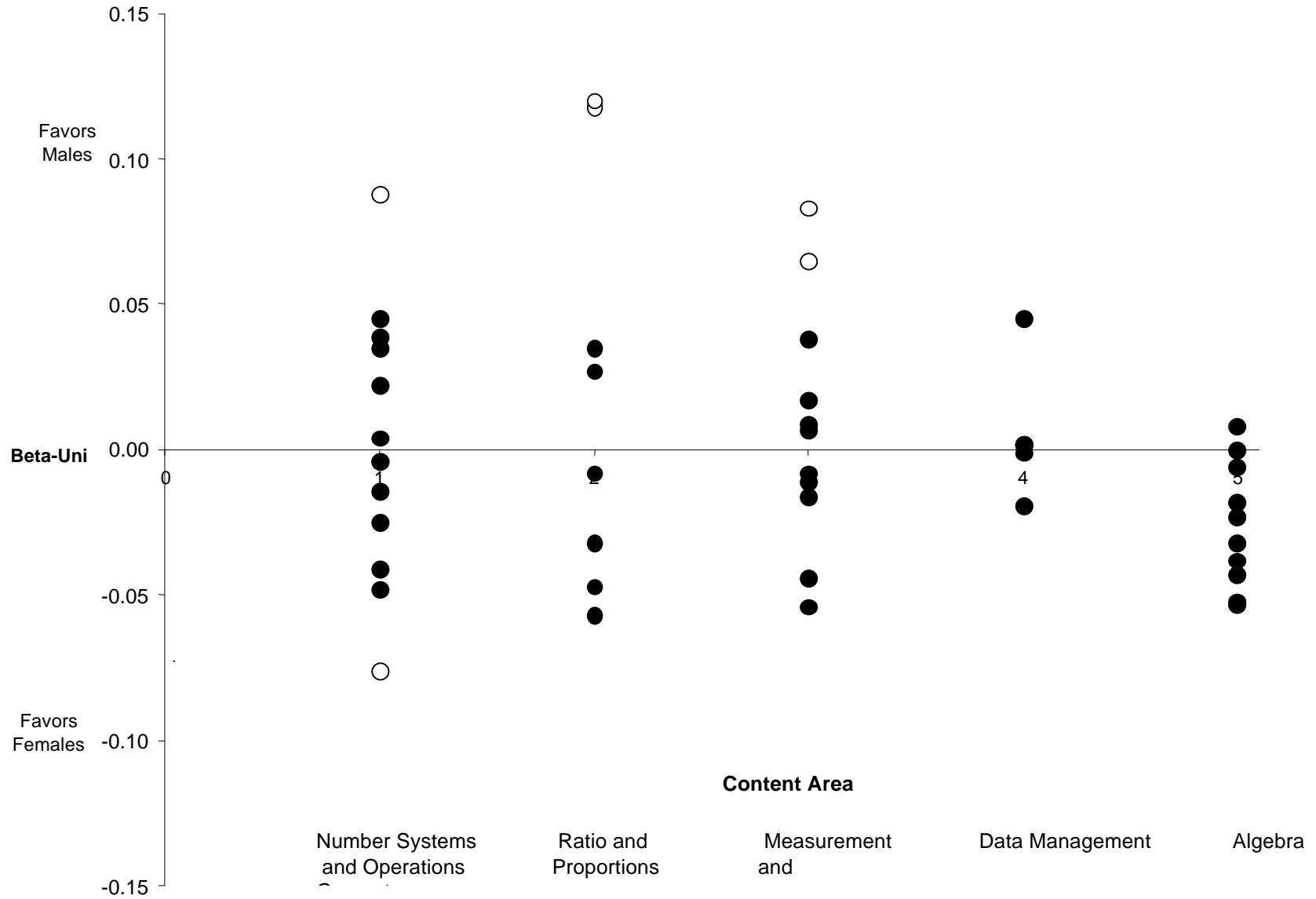
Grade 6 1996 Mathematics



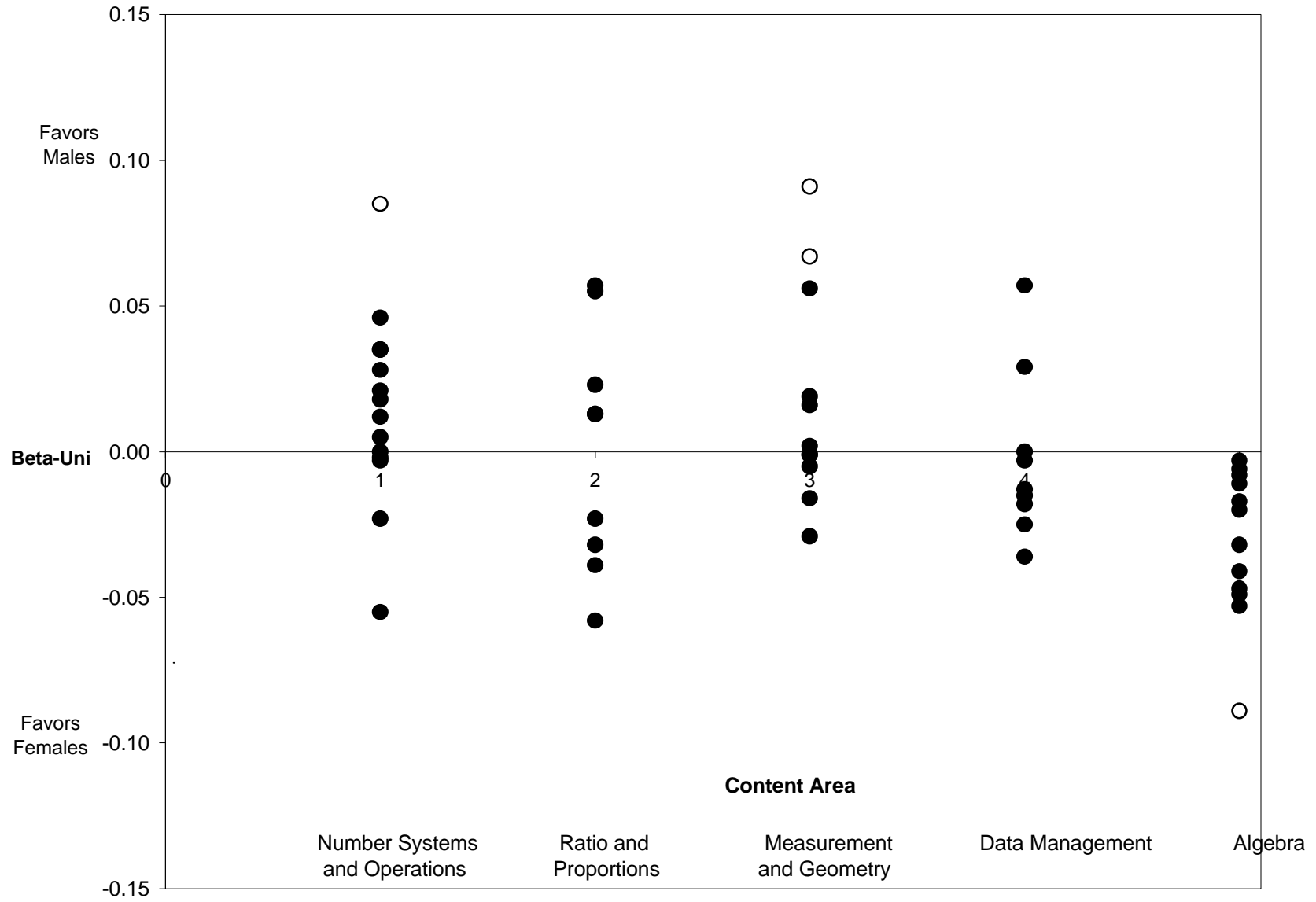
Grade 6 1997 Mathematics



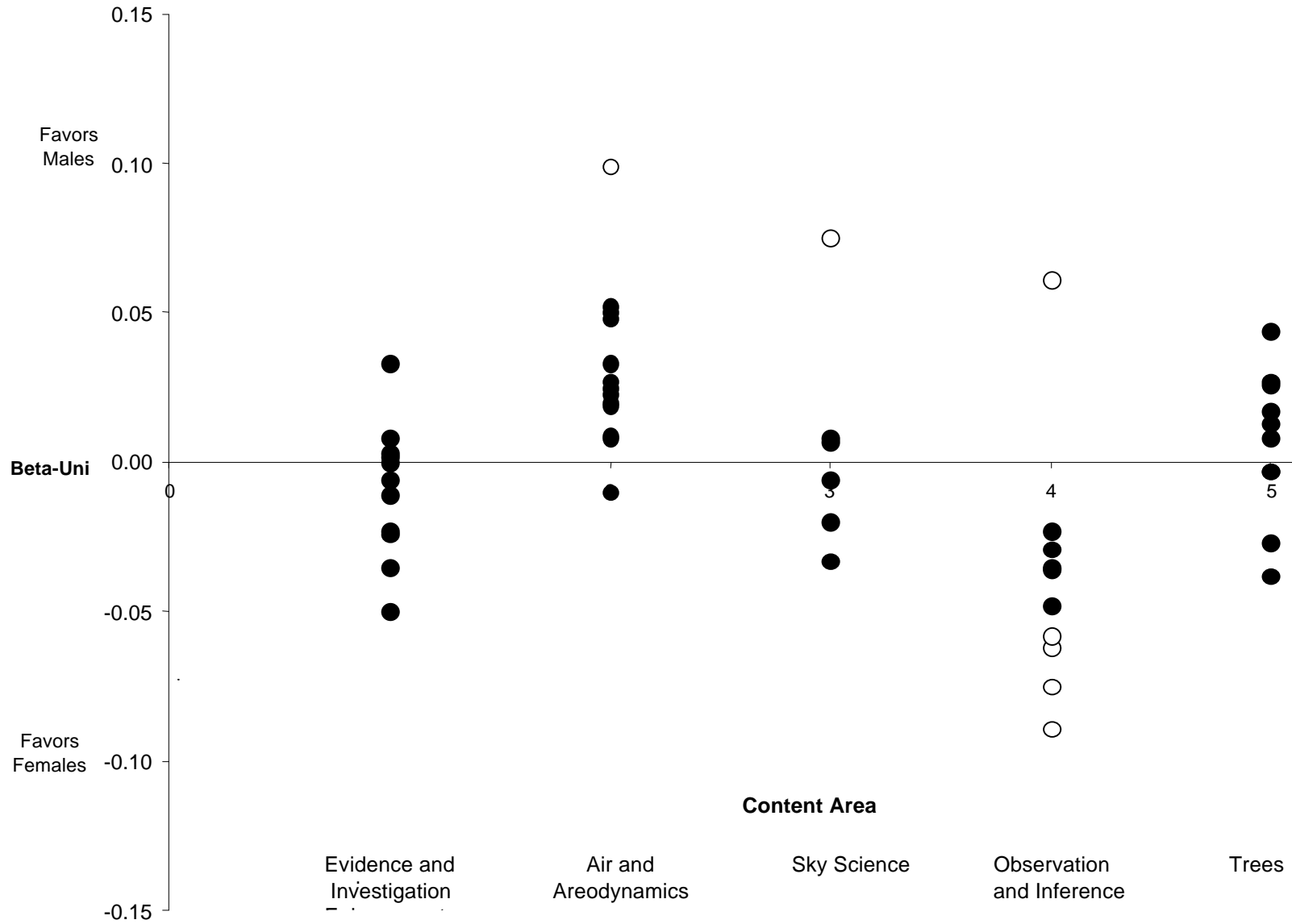
Grade 9 1995 Mathematics



Grade 9 1996 Mathematics



Grade 6 1997 Science



Grade 6 1998 Science

