Robustness of Ability Estimation to Multidimensionality in CAST with Implications to Test Assembly

Yanwei Zhang
AICPA
Ratna Nandakumar
University of Delaware

Paper Presented at the Annual Conference of the National Council of Measurement in Education, San Francisco, CA
April 2006

Abstract

Computer Adaptive Sequential Testing (CAST) is a test delivery model that combines features of the traditional conventional paper-and-pencil testing and item-based computerized adaptive testing (CAT). The basic structure of CAST is a panel composed of multiple testlets adaptively administered to examinees at different stages. Current applications of CAST reply on the item response theory (IRT) and assume a unidimensional IRT model for scoring. This study evaluated the robustness of CAST when tests were constructed, administered, and scored by a unidimensional IRT model but item responses were multidimensional. Various conditions of multidimensionality were simulated in item pools, as well as different levels of content misclassification through manipulation of the correspondence between content area and dimension of items. An automated test assembly (ATA) process constructed CAST panels from the item pools, each representing a unique combination of multidimensionality and content misclassification. Administration of the panels was simulated and multidimensional response data were scored by the unidimensional IRT model. The ability scores, routing decisions, and pass-fail decisions were evaluated against "true" ability scores and decisions to assess the impacts of multidimensionality and content misclassification. Results showed that, when multidimensionality was mild as measured by the angle distance between item clusters, unidimensional ability estimates and routing decisions were not sensitive to the level of content misclassification in item pools.

Introduction

Computerized Adaptive Sequential Testing (CAST) is a test delivery model that combines features of traditional paper-and-pencil examination and item-based computerized adaptive testing (CAT). Its basic structure consists of a panel of multiple stages sequentially administered to examinees. At each stage, there are different groups of items called modules or testlets, and examinees are required to take one of the modules. Which module an examinee takes depends on their ability estimates from the previous stage or stages. That is, examinees are routed during the testing process and they will take different test forms, or pathways, of the examination, each made of a different combination of modules. Figure 1 shows a CAST panel made of three stages, with one panel at the first stage, and two panels at both second and third stages. Each possible combination of modules of a CAST panel can be assembled to meet statistical and non-statistical criteria prior to exam administration, which enables CAST to claim two significant benefits – the assurance of quality and parallelism of test forms and the control of item exposure. The CAST model has produced promising results in a number of research studies and in the field test forms for the United States Medical Licensing Examination Step 1 in 1997 (Luecht & Nungester, 1998, 2000; Luecht, Brumsfield, & Breithaupt, 2002; Luecht & Burgin, 2003). CAST was also adopted by the computerized Uniform Certified Public Accountants (CPA) Exams launched in April 2004 (Melican, Breithaupt, & Mills, 2005) for the multiple choice questions (MCQ) of the tests.

This study is the first attempt, to our knowledge, to evaluate the robustness of computerized adaptive sequential testing (CAST) to the violation of the unidimensionality assumption. The basic research design follows the logic of previous studies on CAT ability estimation with multidimensional data (e.g. Ackerman, 1991; Folk & Green, 1989). First, multidimensional item response data are generated and then calibrated by a unidimensional IRT model. Unidimensional

item parameter estimates are used to assemble CAST panels. Second, administration of panels is simulated. During the administration, item responses are generated to be multidimensional, while ability estimation, assuming unidimensional data, is based on item parameter estimates used to assemble the panels. Finally, the estimated abilities are evaluated against "true" abilities to assess the impact of multidimensionality.

However, unlike previous studies, this study focuses on the potential capacity of CAST to "control" the dimensional structure of a test by prescribing the content specification of items in the panel assembly process. The rationale for such a focus is explained as follows.

Traditionally, the content and the dimension of items are closely related. Items from a specific content area tend to lie along one dimension in the latent trait space and items from a different content area tend to lie along a different dimension. A typical example is a mathematics test that contains algebra items and trigonometry items, corresponding to an algebra dimension and a trigonometry dimension (Zhang & Stout, 1999). In real tests, however, a perfect correspondence between content specification and statistical dimensionality rarely exists. More often we will see items from the same content area lie along different dimensions, and items on one dimension may come from different content areas. Because dimensionality is a statistical feature that secures an item response model to be monotone and locally independent (Nandakumar & Ackerman, 2004), the dimensionality structure may or may not correspond to the content specifications of a test assigned by test developers or content experts (for example, the ACT math test studied in Ackerman, 1991). The reason for this to happen is probably because test development and dimensionality assessment are usually two separate processes, and items may be assigned to content areas not necessarily aligned with their statistical dimensions. In this study, the statistical dimension represents the true identity of an item. Hence, the content area assigned to the item by test developers can be either correct or incorrect. In other words, there is correct classification of item content and there is misclassification of item content, depending on whether or not the content matches the dimension for an item.

Given the above reasoning, this study tries to create three scenarios to distinguish between different levels of content misclassification of items while CAST panels are assembled. In the first scenario, a perfect classification of items is achieved so that items from one content area are associated with one and only one dimension, and vice versa. This is an ideal situation but not very likely to manifest in real tests. In the second, most realistic scenario, minor misclassification of items happens so that a moderate number of items from one dimension are assigned to a content area "by mistake". Put differently, items from on dimension can be associated with two content areas. The third scenario, more realistic than the first scenario, stands for severe misclassification of item content by which most items from one dimension are assigned to a "wrong" content area.

A major advantage of CAST is to improve test form quality by offering adequate control over test assembly. If test data are in fact multidimensional, and if the multidimensional structure reflects itself in the content classification of items, the test assembly process that controls content specification can be expected to maintain a given dimensional structure in the test. By the same token, if test data are in fact multidimensional, and if the multidimensional structure fails to reflect itself in the content classification of items, the test assembly process will not maintain the dimensional structure in the test when correct content classification is mistakenly assumed. Whether the dimensional structure is maintained or not should have some impact on ability estimation and important decisions based on ability estimates. By utilizing the above three scenarios of content misclassification to evaluate unidimensional ability estimation from multidimensional data, this study intended to answer the question of the robustness of unidimensional ability estimation from a new perspective. In short, this study evaluated unidimensional ability estimation in CAST as a function of the interaction of the level of multidimensionality and the level of content misclassification.

Method

Simulated data were used to investigate the objectives of this study. Various multidimensional data were simulated corresponding to different conditions of multidimensionality. For data associated with each multidimensional condition, unidimensional item parameters were estimated, and a content classification was assigned to each item. In this manner, an item pool was created for each condition. The item pool was used to construct a CAST panel. The CAST panel was then administered to simulees and item response data was generated under the true multidimensional model. Simulees' abilities were again estimated using the unidimensional model and evaluated appropriately. The data preparation and analyses are depicted in Figure 2, where each box represents a major step of this study. These steps are briefly described below as an overview, and more detailed descriptions of each step will be provided in the following sections.

Step 1: Item response data were generated to mimic different two-dimensional data conditions based on a MIRT model with two compensatory abilities. In order to generate item responses, item parameters for two clusters of items (Cluster I and Cluster II) were generated corresponding to two particular composites of the two dimensions (θ_1 and θ_2). For each two-dimensional data condition, the following steps 2 to 5 were carried out.

Step 2: For each two-dimensional data condition, unidimensional item parameters were estimated from item responses generated in Step 1. After calibration, items were assigned a content classification (Content A or Content B) three times, each resulting in an item pool, where the correspondence between the content and the cluster for some of the items varied.

Step 3: An automated test assembly (ATA) process was used to select items from the item pool into the CAST panel that met both content and statistical targets.

Step 4: Administration of the panel to examinees was simulated based on the two-dimensional MIRT model used for data generation in Step 1. Original item parameters and abilities were used in this process. Examinee abilities were estimated under the assumption of unidimensionality with known item parameters estimated in Step 2. Examinees were placed on different routes during the panel administration and were given a pass/fail decision based on a passing score.

Step 5: Finally, examinees' ability estimates, routing decision during the panel administration, and the pass-fail decision were evaluated against the "true" abilities and the "true" decisions.

In summary, the CAST panel was created using "incorrect" item parameters. Namely, unidimensional item parameter estimates of multidimensional response data were used. However, in simulating the administration of the panel, data were generated using the two-dimensional model, because these were two-dimensional items. Again, examinee ability levels were estimated "incorrectly" under the unidimensional model.

In an ideal and "correct" situation, two-dimensional parameter estimates should be used in forming the item pool; two-dimensional item parameter estimates should be used to construct CAST panels; and finally, two abilities should be estimated for each examinee after administration of panel items.

Step 1: Generation of Multidimensional Data

This section elaborates Step 1 in Figure 2. Before explaining the details of Step 1, the MIRT model will be described as it forms the foundation for further explanation.

Specification of MIRT Model

This study used the three-parameter logistic model with two compensatory abilities (Reckase & McKinley, 1983, 1991; Reckase, 1997) to simulate different conditions of two-dimensional data. The model is presented as

$$P_i(\theta_{1j},\theta_{2j}) = c_i + \frac{1-c_i}{1+\exp[-1.7[a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)]}. \quad (1)$$

Here $P_i(\theta_{1j},\theta_{2j})$ stands for the probability of correct response to a dichotomous item *i* by an examinee *j* with two abilities $(\theta_{1j},\theta_{2j})$. The parameters a_{1i} and a_{2i} represent, respectively, the discrimination parameters of the item *i* on θ_1 and on θ_2, whereas d_i is the difficulty parameter of item *i*, and c_i is the guessing parameter of item *i*.

In this model, each item in the test is driven by two abilities simultaneously and it measures a composite of θ_1 and θ_2. In the two-dimensional space, the direction of an item composite can be determined by its angle in the θ_1 and θ_2 plane. Figure 3 shows such a space in which the line l_i represents the vector for item *i,* which is a composite of θ_1 and θ_2. The angle α is the distance of the vector from the θ_1 axis. When α is less than 45°, the composite relies more heavily on θ_1 then on θ_2. When α is larger than 45°, the composite relies more heavily on θ_2 then on θ_1. In other words, as the angle α increases, θ_2 contributes more to the composite.

By this kind of presentation, items within distinct narrow fans or clusters are said to measure distinct composites of abilities. The degree of multidimensionality in the test can be determined by three factors. 1. The angle distance between two item clusters, $\gamma = \beta - \alpha$, where α is the angle of items in Cluster 1 from the θ_1 axis, and β is the angle of items in Cluster 2 from the θ_1 axis. 2. The number of items in each cluster. 3. The correlation between the two abilities ($\rho_{\theta_1\theta_2}$). This is displayed in Figure 4. It can be seen that items in Cluster 1 are at angle α with the θ_1 axis, and items in Cluster 2 are at angle β with the θ_1 axis. Hence, the angle distance is the difference between the angles of the two clusters in the θ_1 and θ_2 space, namely, $\gamma = \beta - \alpha$. Smaller γ indicates lower multidimensionality in the test. The number of items in

each cluster also determines the degree of multidimensionality. The larger the number of items within one cluster relative to the other, the lower is the degree of multidimensionality. A test with equal number of items in the two clusters has higher degree of multidimensionality than a test with unbalanced clusters. At the same time, when the other two factors are constant, smaller correlation between the two abilities results in higher level of multidimensionality.

Specification of Data Conditions

As described in the previous section, three factors determine the dimensional structure of item response data. In this study, three levels of γ were considered: γ=30°, with α=30° and β=60°; γ=60°, with α=15° and β=75°; and γ=90°, with α=0° and β=90°. In this study each item pool consisted of 1000 items. Three combinations of number of items in each cluster (in the item pool) were used: N_I = 500 and N_{II} = 500; N_I = 700 and N_{II} = 300; N_I = 900 and N_{II} = 100, where N_I denotes the number of items in Cluster 1 and N_{II} denotes the number in Cluster 2. Two levels of correlation between the abilities ($\rho_{\theta_1\theta_2}$) were used, 0.3 and 0.7. These three factors were completely crossed producing eighteen two-dimensional conditions. In addition, there was also one unidimensional condition where all items fall in one cluster (Cluster 1). In total there were nineteen conditions.

Table 1 lists all the data conditions. In the condition code, D1 refers to unidimensional data, and D2 refers to two-dimensional data. The two-dimensional conditions are organized first by $\rho_{\theta_1\theta_2}$, then by γ, and finally by N_I and N_{II}. For example, D2E2 refers to a condition with two item clusters, where Cluster I has 700 items and Cluster II has 300 items; the angle between the clusters is 60° and the correlation between θ_1 and θ_2 is 0.7.

For all two-dimensional conditions, items in the pool were randomly split into two clusters with appropriate numbers in each cluster. Items in Cluster 1 were

simulated to be at angle α from the θ_1 axis, and items in Cluster 2 were simulated to be at angle β from the θ_1 axis.

Simulation of Data

For each data condition in Table 1, dichotomous responses of 5000 examinees to the 1000 items were generated to embody the specified dimensional structure. The simulation procedure is detailed below.

Simulation of Unidimensional Data

Unidimensional data were generated using the unidimensional IRT model given in

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + \exp[-a_i(\theta - b_i)]} \ . \qquad (2)$$

Examinee abilities were sampled from the standard normal distribution. Instead of generating item parameters, unidimensional item parameter estimates (a_i, b_i, and c_i) from one subject of the Paper-and-Pencil Uniform CPA Exams administrated from 1999 to 2003 were used in this study. There were originally 1286 items. To reduce the impact of extreme item parameters on the analyses, 286 items were removed, and the parameters (a, b, c) of the remaining 1000 items were summarized in Table 2.

For each examinee, for each item, the probability of getting the item correct for the examinee (P_i) is computed using Equation 2 and compared to a random number from the uniform distribution with interval (0, 1). If the P_i value was greater than the random number, the item was considered to be answered correctly and a score of 1 was assigned. Otherwise, a score of 0 was assigned.

Simulation of Two-dimensional Data

Two-dimensional data were generated using the MIRT model given in Equation 1. The two ability values (θ_1 and θ_2) were sampled from a bivariate normal distribution with a specified value of correlation (0.3 or 07). Two dimensional item

parameters were generated from the unidimensional item parameters used in the unidimensional data simulation as follows.

Two-dimensional discrimination parameters of an item *i* were generated as (Kim, 1994; Yu & Nandakumar, 2001; Nandakumar, Yu, & Zhang, 2003):

$$a_{1i} = a_i \cos(\alpha_i), a_{2i} = a_i \sin(\alpha_i)$$

where a_i is the unidimensional discrimination parameter and α_i is the angle of the item vector in the θ_1 and θ_2 space. For each two-dimensional data condition shown in Table 2, the angle was defined by the value of α for Cluster 1 and the value of β for Cluster 2. The difficulty and guessing parameters were defined as:

$$d_i = b_i, \; c_i = c_i$$

where b_i is the unidimensional difficulty parameter and c_i is the guessing parameter. The generated two-dimensional item parameters are summarized in Table 3.

To obtain dichotomous responses, for each examinee, for each item, the probability of getting the item correct for the examinee (P_i) is computed using Equation 1 and compared to a random number from the uniform distribution with interval (0, 1). If the P_i value was greater than the random number, the item was considered to be answered correctly and a score of 1 was assigned. Otherwise, a score of 0 was assigned.

In summary, this data generation process produced eighteen multidimensional data sets and one unidimensional data set, corresponding to the nineteen data conditions in Table 1. Each data set had 5000 examinees and 1000 items.

Step2: Creation of Unidimensional Item Pools

The following sections describe the creation of unidimensional item pools for different conditions. Each item in the pool is associated with two pieces of information: unidimensional parameter estimates and a content specification.

Calibration of Unidimensional Item Parameters

For each data set created in Step 1, unidimensional item parameters were estimated (calibrated) using the BILOG software. BILOG (Mislevy & Bock, 1990) was developed for binary responses following unidimensional IRT models. It utilizes a marginal maximum likelihood estimation method (MML) via iterative EM algorithm and Newton-Gauss (Fisher Scoring) methods. The EM algorithm is an iterative procedure for finding the maximum likelihood estimates of parameters of IRT models in the presence of unobserved random variables.

The IRT model that was used to generate the unidimensional data was assumed for the calibration of unidimensional and multidimensional data. For each data set, the calibration process produced a set of item parameter estimates ($\hat{a}$, $\hat{b}$, and $\hat{c}$).

Assignment of Content Codes to Calibrated Items

Content specification is a major input to the assembly of CAST panels to distribute items according to the content "blueprint" of a test in order to ensure the validity and parallelism of test forms. In this study, content specification of items was manipulated to introduce different scenarios in the correspondence between content and dimensionality in the item pools. The manipulation was intended to mimic typical misclassifications of content for items in real tests.

In an ideal situation, the content and the dimensionality of an item should be perfectly matched. In other words, an item is either categorized as Content A if it belongs to Cluster I (which is the first dimension), or as Content B if it belongs to Cluster II (which is the second dimension). To create three scenarios (X1, X2, and

X3), each item was reassigned a content code three times. In the scenario X1, the first 500 items in the pool were always classified as Content A and the second 500 items were classified as Content B (N_A = 500, N_B = 500). In the scenario X2, the first 700 items in the pool were always classified as Content A and the second 300 items were classified as Content B (N_A = 700, N_B = 300). In the scenario X3, the first 900 items in the pool were always classified as Content A and the last 100 items were classified as Content B (N_A = 900, N_B = 100).

Among these three scenarios, only one scenario has the true specification of content for all its items. For example, in the condition D2A1 first 500 items are of content A, and the next 500 items are of content B (N_I = 500, N_{II} = 500). After reassigning the content specification of these items, the scenario X1 matches perfectly with the true content specification of items in the pool, whereas scenarios X2 and X3 are misclassification of item content. In X2, 200 items were misclassified as content B. In X3, 400 items were misclassified as content B. Another example is the condition D2A3, where there are 900 items of content A and 100 items of content B in the item pool (N_I = 900, N_{II} = 100). After the reassignment of content, the scenario X3 corresponds to the correct classification of content of items, while X1 and X2 correspond to misclassification of content for 400 items and 200 items respectively.

Table 4 shows all the conditions of Table 2 with three scenarios of content assignment. Since there are 19 different conditions, each with three scenarios, there are 57 item pools in total. Each item pool consists of 1000 items with parameter estimates ($\hat{a}$, $\hat{b}$, and $\hat{c}$) and content codes (Content A or Content B).

Step 3: Assembly of CAST Panels

For a given scenario of each two-dimensional condition, a CAST panel was established by utilizing an automated test assembly program. The following

sections describe the process that includes choosing a panel design, deriving statistical targets, and assembling modules and panels.

Description of the 1-2-2 Panel Design

Figure 1 demonstrates a 1-2-2 panel. This panel has five modules or testlets, one of medium difficulty at Stage One, two modules (medium and hard) at Stage Two, and two modules (medium and hard) at Stage Three. Hence there are four possible routes or pathways in this panel. The two primary routes are: Medium-Medium-Medium (MMM) and Medium-Hard-Hard (MHH). The two secondary routes are: Medium-Medium-Hard (MMH) and Medium-Hard-Medium (MHM).

In this study each module had 20 items, thus there were 60 items on each route and a total of 100 items in the panel. When the panel was administered, all examinees would take the medium module at Stage One. Their performance on those 20 items would determine whether they were routed to a medium or a hard module at Stage Two. When they finished the second module, their performance on the 40 items they took so far would determine whether they were routed to a medium or a hard module at Stage 3. When they finished the module at Stage 3, their final performance would be estimated based on all the 60 items they had answered.

The 1-2-2 design was chosen to resemble the model adopted by the Computerized Uniform CPA Exams that certify top CPA candidates in the United States. For certification exams such as the CPA Exams, high measurement precision is required for two groups of candidates: those who are around the passing standard, and those who fail as they need diagnostic information on their performance. Those two groups roughly correspond to the two primary pathways in a 1-2-2 panel.

Derivation of the Target Test Information Function (TTIF)

The assembly of CAST panels requires target test information function (TTIF) as the statistical target (van der Linden, 1998; Luecht, 1992, 1998). The TTIF indicates the amount of test information desired across the latent proficiency scale.

The TTIF also indirectly helps control the distribution of item difficulty for each module in the panel.

Because a panel is made of modules of varied difficulty, a different TTIF is required for each module. In practice, however, the key to generating TTIF is to focus on the primary routes (Luecht, 2000) and then distribute the target information among the modules on the routes. There are several strategies to create TTIF for the primary routes, all involving a technique that generates TTIF at specific locations on the ability scale. The technique is called the Average Maximum Information or AMI (Luecht, 2000). AMI is conceptually similar to simulating multiple adaptive tests without replacement. It locates certain points on the ability scale where the maximum information is required, and then generates feasible information functions from available item parameters. Because AMI is an empirical way to produce realistic targets based on the item pool of a testing program, it allows the testing program to construct many parallel test forms over time. This study implemented the following AMI algorithm (Luecht 2000) to derive the TTIF for the item pools created above.

First, a particular point on the ability scale (θ scale) is located that corresponds to the location for the desired maximum information of a given TIF assuming a normal (0, 1) distribution of θ. For this study, the location of the maximum information was set at 1.0 for the hard route ($\theta_H = 1.0$), and at 0.0 for the medium route ($\theta_M = 0.0$).

Then, for each item in the pool, item information is computed at the two selected locations, that is, IIF_M for θ_M and IIF_{H} for θ_H. Next, the item pool is sorted in descending order by IIF_M and IIF_H. After sorting, for a particular number of items (denoted as n, equal to the test length 60 in this case), another number (denoted as m) is chosen for the so-called "maximally informative replications" without replacement.

The "maximally informative replications" works in the following way: for the item pool sorted by IIF_M and IIF_H, a total of $n \times m$ most informative items are selected at the two locations for maximum information, θ_M and θ_H. This process mimics using an adaptive item selection algorithm to build m non-overlapping test

forms of length n that are maximally informative at the two θ points. In general a larger m makes the derived information targets more robust (Luecht, 1992). In this study, m was fixed to 5, so $60 \times 5 = 300$ most informative items were selected from a pool of 1000 items.

Then, the total information of the selected items is computed at a series of θ points and then divided by m to obtain the mean TTIF for each primary route:

$$TTIF_{jk} = T_j(\theta_k) = \frac{\sum_{i=1}^{n \times m} I_i(\theta_k)}{m}.$$

where $j = H, M$, corresponding to the two primary pathways, and $k = 1, \ldots, 61$, representing 61 equally spaced points from -.3 to +.3 on the θ scale.

For a panel of equally sized modules at each stage, such as the panels in this study, $TTIF_{Hk}$ and $TTIF_{Mk}$ are divided by the number of stages (3 in this case) at each θ point (θ_k, $k = 1, \ldots, 61$), which results in two different TTIF for Stage One, $TIF_{H(1)k}$ and $TIF_{M(1)k}$. Then the two information values are averaged at each θ point to produce a single target $TTIF_{(1)k}$ for Stage One, i.e. $TTIF_{(1)k} = \sum TTIF_{j(1)k} / 2$.

At the final step, the total-test-length TTIF ($TTIF_{Hk}$ and $TTIF_{Mk}$) is multiplied by the percentage of items at Stage Two and Stage Three. Let the total test length be n, and the length of the first stage be n_1, the percentage is $p = (n - n_1) / n$, and the TTIF for Stage Two and Stage Three is then $TTIF' = p\ (TTIF_{jk})$. After adding the TTIF of Stage One, $TTIF_{(1)k}$, to the TTIF of Stage Two and Stage Three ($TTIF'$), the full-length target for each primary route is obtained. That is, $TTIF_{Mk} = TTIF_{Mk}' + TTIF_{(1)k}$ for the medium route, and $TTIF_{Hk} = TTIF_{Hk}' + TTIF_{(1)k}$ for the hard route.

After the primary-route TTIF is obtained, it can be divided by the proportional size of the stages to get TTIF for each module along the route. Because IRT test information is additive, TTIF of a route can be broken apart as easily as they can be put together. When modules have the same number of items, the divider is simply the number of stages. In this study, it was $TTIF_{Mk} / 3$ for each medium module, and $TTIF_{Hk} / 3$ for each hard module.

For each item pool in this study, the above procedure derived the TTIF for the five modules in the 1-2-2 panel as part of the input to the following test assembly process.

Assembly of Modules and Panels

To build test forms for CAST, an automated test assembly (ATA) process selects items from an item pool into modules at the different stages of a panel. The selection simultaneously satisfies both statistical and non-statistical constraints. Technically, the assembly can be viewed as a multiple-objective function, constrained optimization problem. This study used a heuristic designed for large-scale test assembly called the normalized weighted absolute deviation heuristic (NWADH) (Luecht, 2000) implemented in a DOS-based shareware CASTISEL (Luecht, 1998b).

CASTISEL uses the NWADH heuristic to build multistage tests that satisfy multiple statistical objectives, such as TTIF for each module, and limited content specifications. The current version of the software allows for only one level of content classification such as in this study. The program requires four input files: an item bank file, a target test information file, a content constraint file, and a command file.

The item bank file stores the IRT parameters and content codes of each item. In other words it represents the item pool. The target test information file hosts the TTIF derived for each module of a panel. The content constraint file specifies how the items are distributed among content areas. Unlike the TTIF required for each module, the content distribution is controlled for the panel as a whole. CASTISEL uses a "partitioning algorithm" (Luecht & Nungester, 1998) to allocate the total-test content requirements to the different stages in the panel. The modules at the same stage are constructed to satisfy the same partition of content. The command file contains the syntax for the program, specifying data input and controlling how the panel should be built, such as the number of items in each module, the number of modules at each stage, and the number of stages in the panel.

CASTISEL delivers several output files. Among them the item sequence file provides the sequence numbers of the items selected for each module. In other words, it represents an assembled panel.

From each item pool described in Table 3, the CASTISEL program was used to select five modules, each of 20 items, and assemble them into a 1-2-2 panel. The distribution of content in the panel followed the same ratio of Content A and Content B items in the item pool. In other words, if an item pool had 700 items in Content A and 300 in Content B, the panel of 100 items assembled from the pool would have 70 items from Content A and 30 from Content B.

This study assembled one panel from each item pool to ensure the quality of the panel. The assembly of multiple panels was not attempted, because the current version of CASTISEL is unable to build simultaneously multiple panels with equivalent quality from the same item pool. As a panel corresponds uniquely to an item pool that also uniquely represents an item pool in Table 4, each panel was identified with the corresponding item pool in the rest of the paper. For example, the panel D2A2/X1 refers to the panel built from the item pool representing the X1 scenario under the D2A2 data condition.

Step 4: Simulation of Panel Administration

After a panel was assembled, the study simulated its administration to examinees by generating responses by 1000 simulees to the 100 items selected into the panel. For two-dimensional data, although panel construction was based on unidimensional item parameter estimates, two-dimensional data were simulated using the MIRT model to reflect the true nature of data.

In order to create a sparse data matrix, a full item response matrix was generated first without any missing data. Then the full matrix was transformed into a sparse one by substituting missing values for the responses to items not supposed to be answered by some examinees.

Item responses were generated to be either two-dimensional or unidimensional. If a panel was assembled from an item pool associated with a two-dimensional condition, the response data were made two-dimensional. For such a panel, the two-dimensional item parameters (a_1, a_2, d, and c) of the 100 items were identified, and 1000 pairs of ability values (θ_1 and θ_2) were sampled from a bivariate normal distribution where $\rho_{\theta_1\theta_2}$ was fixed to 0.3 or 0.7 depending on the specification of the condition.

For the unidimensional condition, the unidimensional ability θ was sampled from the univariate standard normal distribution. The unidimensional item parameters of the 100 items in the panel were identified from Step 1. Item responses were generated also by the unidimensional IRT model.

The transformation from a full response matrix to a sparse matrix was realized by estimating ability score in the first two stages and routing the examinees to different modules based on the estimates. At the end of Stage one, ability estimate ($\hat{\theta}$) was computed based on the 20 items in the first medium module for all 1000 examinees. If $\hat{\theta}$ was greater than or equal to 0.5 (the midpoint between the two targets where the TIF was maximized for each primary route), the examinee was placed on the hard route for Stage two. Otherwise the examinee was routed to take another medium module. Because examinees would not have answers to the items in the module they were not routed to, responses to the items not seen were replaced by missing values. Similarly, at the end of Stage two, ability score was computed based on examinees' performance on their first 40 items. Again, if $\hat{\theta}$ was greater than or equal to 0.5, examinees were directed to the hard route at Stage three, otherwise to the medium route. As these examinees would not have answers to the module they did not take, their responses to the items in that module were changed to missing values. By then, most examinees would have stayed in their original route (MMM or MHH), but a few of them could be shifted from the medium to the hard (MMH), or vise versa

(MHM). At the end of Stage Three, the ability of examinees was estimated for the last time based on all the 60 items they answered.

It should be noted that, at each of three Stages, unidimensional ability was estimated under the assumption that item parameters are known. The item parameters used in the ability estimation were those stored in the item pools to assemble the panels.

The above data generation, ability estimation, and routing was replicated 200 times for each panel. Each replication sampled a different set of 1000 θ_1/θ_2 pairs for two-dimensional data, or a different set of 1000 θ values for unidimensional data.

Step 5: Evaluation of Ability Estimation, Routing Decision, and Pass-fail Decision

Item responses generated for the panel administration simulated in Step 4 were either two-dimensional or unidimensional. However, ability estimation during and after the panel administration assumed unidimensional data and used the item parameters estimated by a unidimensional IRT model. This study evaluated the ability estimation from three perspectives - how well $\hat{\theta}$ retrieved its "true" value; how correctly examinees were routed during the panel administration; and how accurately examinees were classified as passing or failing the exam given a particular passing score.

Evaluation of Ability Estimation

The main objective is to compare the estimated ability ($\hat{\theta}$) with the "true" ability. For unidimensional data, the true ability θ_T was the single ability value used to simulate item responses of each simulee during the data simulation process in Step 4. For two-dimensional data, since two abilities generated the item responses, θ_T was defined as a weighted linear combination (WLC) of θ_1 and θ_2 used to simulate two-dimensional item responses for each simulee in Step 4, i.e. $\theta_T = w_1\theta_1 + w_2\theta_2$.

There is no widely accepted definition of "true" ability as a combination of θ_1 and θ_2, especially when the two-dimensional structure in item responses was determined simultaneously by multiple factors as in this study. In this study the weights (w_1 and w_2) were assigned to reflect the proportions of items in two content areas (Content A and Content B) specified by the assembly process for a given panel. Thus for a real test the weights would reflect the emphasis on different content areas specified by test developers for that test. The WLC then would symbolize the "expected" or "desired" content-based composite score from the perspective of test development.

Hence, if a panel was assembled with 50 items from Content A and 50 items from Content B, then $w_1 = 0.5$, $w_2 = 0.5$. If a panel was assemble with 70 items from Content A and 30 items from Content B, then $w_1 = 0.7$, $w_2 = 0.3$. If a panel was assembled with 90 items from Content A and 10 items from Content B, then $w_1 = 0.9$, $w_2 = 0.1$. As one can recall from Section 3.4.3, the ratio of Content A and Content B in the panel followed the same ratio of Content A and Content B items in the item pool. The above combinations of w_1 and w_2 therefore reflected the content specification of the item pools listed in Table 4. From Table 4 one can infer that the panels ended with /X1 always had $w_1 = 0.5$ and $w_2 = 0.5$, the panels ended with /X2 had $w_1 = 0.7$ and $w_2 = 0.3$, and the panels ended with /X3 had $w_1 = 0.9$ and $w_2 = 0.1$.

The evaluation of how well θ_T was recovered by $\hat{\theta}$ was accomplished using three methods. The first is the Pearson correlation ($r_{\theta_T\hat{\theta}}$) between θ_T and $\hat{\theta}$. A larger correlation indicates better recovery of θ_T by $\hat{\theta}$. The second is the root mean square difference (RMSD) between $\hat{\theta}$ from θ_T, which is defined as

$$RMSD_{\hat{\theta},\theta_T} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\hat{\theta}_j - \theta_{Tj})^2} .$$

where *J* stands for the total number of examinees. A smaller RMSD indicates better recovery of θ_T by $\hat{\theta}$. The third method is a graphic display that plots $\hat{\theta}$ against θ_T across the ability scale (-3 to +3). In this plot, the Y-axis provides the scale for both θ_T and $\hat{\theta}$, and the X-axis represents the simulees ordered by θ_T from the lowest value (the left end of the X-axis) to the highest value (the right end of the X-axis). For each simulee, the values of $\hat{\theta}$ and θ_T are plotted. There are two lines in the plot, one smooth (θ_T) and the other zigzagged ($\hat{\theta}$). The line of $\hat{\theta}$ either goes above or below the line of θ_T at each point on the X-axis. The "scattering" or deviance around the θ_T line indicates the size of error in ability estimation.

Evaluation of Routing Decision

This study applied two methods to evaluate the accuracy of routing decisions. First, a two-by-two table was computes for routing decisions at the end of Stage one and Stage two; and the proportion of simulees falling into each cell of the table was computed. Second, a chi-square (χ^2) statistics was used to examine whether there was agreement between the decisions based on $\hat{\theta}$ and θ_T.

As described in Step 4, simulees are routed twice during the 1-2-2 panel administration, first time (initial routing) at the end of the Stage One, and second time (final routing) at the end of Stage Two. Each time simulees can be placed on the hard (H) route or the medium (M) route based on their performance on the items they had taken. Routing decisions were made based both on θ_T (true routing) and on $\hat{\theta}$ (estimated routing). Thus for each replication of a given panel, at the end of Stage One and Stage Two, the following two-by-two frequency table can be formed between the two routing decisions (on θ_T and on $\hat{\theta}$):

Routing Decision Based on $\hat{\theta}$	Routing Decision Based on θ_T	
	To Hard Route	To Medium Route
To Hard Route	A	B
To Medium Route	C	D

In this table the cell A contains the number of simulees placed on the hard route by both θ_T and $\hat{\theta}$. The cell B contains the number of simulees placed on the medium route by θ_T but on the hard route by $\hat{\theta}$. Similarly cells C and D hold the frequency, respectively, of simulees placed on hard route by θ_T but on the medium route by $\hat{\theta}$, and of simulees placed on the medium route by both θ_T and $\hat{\theta}$. In other words, the cells A and D represent the agreement between true routing and estimated routing.

At the end of Stage One and Stage Two, a given simulee was coded as "Correct" if the estimated routing matched the true routing (corresponding to cells A and D in the above table). Otherwise the routing was coded "Wrong" (corresponding to cells B and C in the above table). Because routing was performed at the end of two stages for each simulee, routing codes fell into one of the following four categories: "Correct-Correct", "Wrong-Correct", "Wrong-Wrong", and "Correct-Wrong", where the first word referred to the initial routing and the second word to the final routing.

Chi-square test is the second method to evaluate routing decisions. The purpose is to assess the level of agreement between routing decisions based on θ_T and on $\hat{\theta}$. The null hypothesis tested is that the routing based on $\hat{\theta}$ was independent from the routing based on θ_T. A large χ^2 value would reject the null hypothesis and testify that there is significant correlation between the two decisions.

For each replication of a panel, a chi-square test was run twice against the two-by-two frequency table shown above, first for the initial routing, and then for the final routing.

Evaluation of Pass-fail Decision

Similar to routing decisions, the pass-fail status of simulees can be determined based on θ_T (true status) and on $\hat{\theta}$ (estimated status), both against a passing score of 1.0 on the θ scale. The passing score of 1.0 is chosen because it is the point on the θ scale where the test information function (TIF) for the hard route of the panel was maximized. The pass-fail decision for an examinee can be "correct" or "incorrect". A pass-fail decision is "correct" when the estimated status based on $\hat{\theta}$ matched the true status based on θ_T, and "incorrect" when these two statuses did not match.

For a panel of unidimensional data or two-dimensional data, the accuracy of pass-fail decisions is evaluated by the type I and type II errors. In this study, a type I error occurred when an examinee who should have failed passed the exam, and a type II error occurred when an examinee who should have passed failed.

Summary of Results

The results for panels with unidimensional item responses are included in Tables 5 through 8. One can see that for unidimensional data, the correlation between estimated and "true" abilities was high ($r_{\theta_T\hat{\theta}} > 0.96$) and their RMSD was low (around 0.27). Among the four routes of the panel, the two primary routes (MMM and MHH) had higher accuracy in ability estimation than the two secondary routes (MMH and MHM). For each panel, about 90% of routing decisions were correct and more than 92% of pass-fail decisions were correct. For pass-fail decisions, both type I and type II errors were below 4%.

Results for two-dimensional data are presented graphically in Figures 6 to 14. Results indicate that, in general, the angle difference between the two item clusters (γ) was a decisive factor. When the angle distance between the two item clusters were relatively small (γ = 30° or 60°), Figure 6 and 7 show that larger correlation ($r_{\theta_T\hat{\theta}}$) and smaller RMSD between $\hat{\theta}$ and θ_T were observed in the X1 panel, regardless how items in the item pool for that panel were distributed between the two clusters, or to what extent the content of items was misclassified. The characteristic of the X1 panel is that the two content areas had equal number of items. The only exception to the above observation was the D2B3 condition where γ = 60° and 90% of items in the item pool were in the first item cluster (i.e. dimension). For this condition, the X2 panel had larger $r_{\theta_T\hat{\theta}}$ and smaller RMSD than the X1 and X3 panels. When the two item clusters were most widely apart in the two-dimensional space (γ = 90°), results of $r_{\theta_T\hat{\theta}}$ and RMSD favored the panels where fewer items were misclassified in content. For this level of γ, how well $\hat{\theta}$ recovered θ_T was positively related with the level of correct content classification of items in the item pool. For each data condition with γ = 90°, the panel with zero misclassification of content always had the highest $r_{\theta_T\hat{\theta}}$ and the lowest RMSD. The "scattering" or deviance around the θ_T line, indicating the size of error in ability estimation, are displayed in Figures 8, 9 and 12 and confirm above findings.

Similar results were found with the routing decisions for panels of two-dimensional data (Figures 11 and 12). When γ = 30° or γ = 60°, the X1 panel for each two-dimensional condition always had the highest proportion of correct routing decisions (again except for D2B3 where the X2 panel had the highest proportion). When γ = 90°, the panel without any content misclassified items had the highest

proportion of correct routing decisions. Chi-square analyses on the agreement between the routing decisions based on $\hat{\theta}$ and those based on θ_T further revealed the above results.

Evaluation of pass-fail decisions for two-dimensional data led to the following findings (Figures 13 and 14). On average, about 92% of the decisions were correct, almost as good as that for unidimensional data. In particular, there were eleven panels where less than 90% of the pass-fail decisions were correct, eight with $\rho_{\theta_1\theta_2} = 0.3$, and three with $\rho_{\theta_1\theta_2} = 0.7$. In all these eleven panels, content was misclassified for either 200 items or 400 items in the item pool. Across all panels approximately 8% of the pass-fail decisions were incorrect. These incorrect decisions were either of type I error or of type II errors. The two types of errors had very different characteristics as discussed below.

The type II error was defined as the proportion of examinees failed who should have passed. For all two-dimensional data conditions, the smallest type II error was observed in the X1 panel, and the largest type II error was found in the X3 panel. Type II errors also increased as the angle distance between the two item clusters increased. Thus panels with $\gamma = 90°$ had larger type II errors than panels with $\gamma = 60°$ which, in turn, had larger errors than panels with $\gamma = 30°$. Larger type II errors were observed in panels with $\rho_{\theta_1\theta_2} = 0.7$ than in the corresponding panels with $\rho_{\theta_1\theta_2} = 0.3$.

The type I error was defined as the proportion of examinees passing who should have failed. Results showed that type I errors behaved very differently from type II errors. First, panels with larger γ had smaller type I errors. Smaller type I errors were observed in panels with $\gamma = 60°$ than in panels with $\gamma = 30°$, and even smaller errors were observed in panels with $\gamma = 90°$. Second, panels with $\rho_{\theta_1\theta_2} = 0.7$

had smaller type I errors than the corresponding panels with $\rho_{\theta_1\theta_2} = 0.3$. Finally, the size of type I errors was directly related with the level of correspondence between content and dimension of items. For each two-dimensional data condition, the smallest type I error was always observed in the panel without any content misclassification.

Results also showed that both types of errors were confined only to the two hard routes, MHH and MMH, particularly to the MHH route. This observation applied to both two-dimensional data and unidimensional data.

In summary, results of this study suggest that, given a particular level of $\rho_{\theta_1\theta_2}$, the angle distance between the two item clusters (γ) determined how $\hat{\theta}$ was related to θ_T. When γ was relatively small (30° or 60°), $\hat{\theta}$ seemed to be the simple average of θ_1 and θ_2, or $\theta_T = 0.5\theta_1 + 0.5\theta_2$. This was true regardless of how many items were in each of the two clusters, or how many items were misclassified between the content and the dimension in an item pool. When θ_T was defined differently, i.e. $\theta_T = 0.7\theta_1 + 0.3\theta_2$ or $\theta_T = 0.9\theta_1 + 0.1\theta_2$, $\hat{\theta}$ failed to represent θ_T well enough, even though the weights of θ_1 and θ_2 may actually reflect the proportions of items between the two clusters in the item pool, i.e. there was no misclassification of content in the item pool. When γ was large (90°), $\hat{\theta}$ and θ_T were related in a different manner. For this level of γ, $\hat{\theta}$ recovered θ_T very well if θ_T was defined in such a way that the weights of θ_1 and θ_2 correctly reflected the proportions of items between the two clusters in the item pool. That is, θ_T was best recovered in the panels without any misclassified content. In these panels the content specification of items truly corresponded to the dimensional structure of the test.

The above relationship between $\hat{\theta}$ and θ_T was revealed in $r_{\theta_T\hat{\theta}}$, RMSD, and graphic displays of $\hat{\theta}$ against θ_T. This relationship between $\hat{\theta}$ and θ_T also determined the results when routing decisions based on $\hat{\theta}$ were evaluated against routing decisions based on θ_T.

Discussion

One can conclude that, when multidimensionality in the item responses was not severe as measured by the angle distance between the two item clusters, unidimensional ability estimates and the routing decisions based on the estimates were not sensitive to the level of content misclassification in the item pool. In this case, panels without any content misclassification did not necessarily have ability estimates more highly correlated with their "true" values, and for that reason, these panels did not necessarily have more correct routing decisions than the panels with content misclassification.

One can also conclude that, only when multidimensionality in item responses was severe in terms of the angle distance between the two item clusters, the quality of unidimensional ability estimates and routing decisions became sensitive to the level of content misclassification in the item pool. In this case, panels without any content misclassification had more accurate ability estimates and routing decisions than panels with content misclassification.

However, in both cases discussed above, results of this study are assuring because they indicate that content misclassification does not necessarily affect the

unidimensional ability estimation and routing decisions of CAST. For all the panels without content misclassification, and for about half of the panels with some content misclassification, both the alignment between estimated and "true" abilities and the agreement between the routing decisions based on estimated and "true" abilities were comparable to the panels of truly unidimensional data.

On the other hand, results of pass-fail decisions give a different message. For pass-fail decisions, type I errors were the smallest for panels without any content misclassification, regardless of the level of multidimensionality in the data. For certification and licensure examinations, the type I error (as defined in this study) is much more important than the type II error. This is because passing an unqualified candidate poses serious damage to the public interest. Although content misclassification, to some degree, is not a serious concern for $\hat{\theta}$ to recover θ_T, it plays a significant role in the accuracy of pass-fail decisions. From this perspective, content misclassification should be sufficiently controlled by a testing program that uses CAST for certification or licensure purpose.

The significance of this study is twofold. First, it is first known assessment of the robustness of unidimensional ability estimation and ability-related decision makings in the Computerized Adaptive Sequential Testing (CAST) when test data are truly multidimensional. Second, and more importantly, it is the first study ever that evaluates the joint implications of both multidimensionality and content misclassification in the item pool. The content specification of items in a real test may not reflect accurately the underlying dimensional structure of the test. Whenever there is disagreement between the dimension and the content of items, one can say that misclassification of content has been introduced into the item pool. When

misclassified content specification of items is used as an explicit criterion for test construction, as in the assembly of CAST panels, the unidimensionally-estimated ability scores and related decisions on examinees may fail to represent the scores and decisions expected based on the content specification in the item pool. This study suggests that the above concern is legitimate, and its consequence may be severe when it involves the pass-fail decisions on examinees.

As for any study that uses simulated data, the findings of this study are restricted by the prescribed multidimensional data conditions and the levels of content misclassification. Items from a real test may as well belong to more than two dimensions or two content areas. The findings are also restricted by the definitions of "true" ability in two-dimensional item responses. Another limitation involves the restraints imposed by the automated test assembly program used in this study. As we know, CASTISEL is not capable of selecting multiple panels with parallel characteristics from a single item pool. As a result, only one panel from each item pool was analyzed. The observations made by this study may not be sustainable when multiple panels are created from an item pool. Another problem is the lack of clear quantification of multidimensionality when it is controlled by several factors simultaneously. A good measure of multidimensionality in data will be helpful to make the results more interpretable. For example, a dimensionality assessment (DIMTEST or DETECT, for example) can be performed on item response data, and the estimated degree of multidimensionality can be related to the results of this study.

References

Ackerman, T. A. (1991). The use of unidimensional parameter estimates of multidimensional items in adaptive testing. *Applied Psychological Measurement, 15*, 13-24.

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement, 13*, 373-389.

Luecht, R. M. (2000). Implementing the computer-adaptive sequential testing (CAST) framework to mass produce high quality computer-adaptive and mastery tests. Unpublished manuscript.

Luecht, R. M. (1998a). Computer-assisted test assembly using optimization heuristics. *Applied Psychological Measurement*, *22*, 224-236.

Luecht, R. M. (1998b). CASTISEL (computer program). Greensboro, NC: Author.

Luecht, R. M. (1992). Generating target information functions and item specifications in test design. Paper presented at the Annual Meeting of the

Luecht, R. M., & Nungester, R. J. (2000). Computer-adaptive sequential testing. In W. J. van der Linden, & C. A. W. Glas (Eds.). *Computerized Adaptive Testing: Theory and Practice*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Luecht, R. M., & Nungester, R. J. (1998). Some practical examples of computerized adaptive sequential testing. *Journal of Educational Measurement, 35,* 239-249.

Luecht, R. M., Brumsfield, T., & Breithaupt, K. (2002). *A testlet assembly design for the Uniform CPA Examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Luecht, R. M., & Burgin, W. (2003). *Test information targeting strategies for adaptive multistage testing designs*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Melican, G., Breithaupt, K. & Mills, C (2005). *Multi-stage testing and case studies in a fully-functioning licensing examination*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Mislevy, R. L., & Bock, R. D. (1990). BILOG (computer program). Chicago, IL: Scientific Software, Inc.

Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences*. Newbury, CA: Sage.

van der Linden, W. J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement*, 22, 195-211.

Yu, F., & Nandakumar, R. (2001). Poly-detect. *Journal of Educational Measurement.*

Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64, 213-249.*

Table 1: List of Data Conditions

Data Type	Correlation (θ_1, θ_2)	Angle Distance γ [a]	Number of Items Cluster I /Cluster II	Data Condition
Unidimensional	N/A	N/A	1000/0	D1
Multidimensional	.3	30°	500/500	D2A1
			700/300	D2A2
			900/100	D2A3
		60°	500/500	D2B1
			700/300	D2B2
			900/100	D2B3
		90°	500/500	D2C1
			700/300	D2C2
			900/100	D2C3
	.7	30°	500/500	D2D1
			700/300	D2D2
			900/100	D2D3
		60°	500/500	D2E1
			700/300	D2E2
			900/100	D2E3
		90°	500/500	D2F1
			700/300	D2F2
			900/100	D2F3

[a] γ is the difference between β (Cluster II) and α (Cluster I) , i.e. $\gamma = \beta - \alpha$.

Table 2: Summary Statistics of Original Unidimensional Item Parameters (N=1000)

Parameter	Mean	S.D.	Max	Mean
a	0.63	0.22	1.25	0.25
b	-0.09	1.04	2.5	-2.50
c	0.23	0.06	0.35	0.10

Table 3: Summary Statistics of Generated Two-dimensional Discrimination Parameters

Data Condition	Item Cluster	N	Parameter	Mean	S.D.	Max	Min
D2A1/D2D1[a]	I	500	a_1	0.55	0.19	1.08	0.22
			a_2	0.32	0.11	0.62	0.13
	II	500	a_1	0.32	0.11	0.62	0.13
			a_2	0.55	0.19	1.08	0.22
D2A2/D2D2[a]	I	700	a_1	0.56	0.19	1.08	0.22
			a_2	0.32	0.11	0.62	0.13
	II	300	a_1	0.31	0.11	0.60	0.13
			a_2	0.53	0.18	1.04	0.22
D2A3/D2D3[a]	I	900	a_1	0.55	0.19	1.08	0.22
			a_2	0.32	0.11	0.62	0.13
	II	100	a_1	0.30	0.11	0.59	0.13
			a_2	0.52	0.19	1.02	0.22
D2B1/D2E1[a]	I	500	a_1	0.61	0.21	1.20	0.24
			a_2	0.16	0.06	0.32	0.07
	II	500	a_1	0.17	0.06	0.32	0.06
			a_2	0.62	0.21	1.21	0.24
D2B2/D2E2[a]	I	700	a_1	0.62	0.21	1.21	0.24
			a_2	0.17	0.06	0.32	0.07
	II	300	a_1	0.16	0.06	0.31	0.06
			a_2	0.59	0.21	1.16	0.24
D2B3/D2E3[a]	I	500	a_1	0.62	0.21	1.21	0.24
			a_2	0.17	0.06	0.32	0.06
	II	500	a_1	0.16	0.06	0.30	0.07
			a_2	0.58	0.21	1.14	0.25
D2C1/D2F1[a]	I	500	a_1	0.63	0.22	1.24	0.25
			a_2	0.00	0.00	0.00	0.00
	II	500	a_1	0.00	0.00	0.00	0.00
			a_2	0.64	0.22	1.25	0.25
D2C2/D2F2[a]	I	700	a_1	0.64	0.22	1.25	0.25
			a_2	0.00	0.00	0.00	0.00
	II	300	a_1	0.00	0.00	0.00	0.00
			a_2	0.61	0.21	1.20	0.25
D2C3/D2F3[a]	I	900	a_1	0.64	0.22	1.25	0.25
			a_2	0.00	0.00	0.00	0.00
	II	100	a_1	0.00	0.00	0.00	0.00
			a_2	0.60	0.21	1.18	0.26

[a]: Statistics are the same for these two conditions.

Table 4: List of Data Conditions after Content Assignment

Data Type	Corr. (θ_1, θ_2)	Angle Distance γ [a]	Num of Items Cluster I / II	Data Condition	Num of Items Content A/B	Item Pool (Panel)
Unidimensional	N/A	N/A	1000/0	D1	1000/0	D1/X1
					500/500	D1/X2
					700/300	D1/X3
Two-dimensional	.3	30°	500/500	D2A1	500/500	D2A1/X1
					700/300	D2A1/X2
					900/100	D2A1/X3
			700/300	D2A2	500/500	D2A2/X1
					700/300	D2A2/X2
					900/100	D2A2/X3
			900/100	D2A3	500/500	D2A3/X1
					700/300	D2A3/X2
					900/100	D2A3/X3
		60°	500/500	D2B1	500/500	D2B1/X1
					700/300	D2B1/X2
					900/100	D2B1/X3
			700/300	D2B2	500/500	D2B2/X1
					700/300	D2B2/X2
					900/100	D2B2/X3
			900/100	D2B3	500/500	D2B3/X1
					700/300	D2B3/X2
					900/100	D2B3/X3
		90°	500/500	D2C1	500/500	D2C1/X1
					700/300	D2C1/X2
					900/100	D2C1/X3
			700/300	D2C2	500/500	D2C2/X1
					700/300	D2C2/X2
					900/100	D2C2/X3
			900/100	D2C3	500/500	D2C3/X1
					700/300	D2C3/X2
					900/100	D2C3/X3

(To be continued)

[a] γ is the difference between β (Cluster II) and α (Cluster I) , i.e. $\gamma = \beta - \alpha$.

Table 4 (continued): List of Data Conditions after Content Assignment

Data Type	Corr. (θ_1, θ_2)	Angle Distance γ [a]	Num of Items Cluster I / II	Data Condition	Num of Items Content A/B	Item Pool (Panel)
Two-dimensional	.7	30°	500/500	D2D1	500/500	D2D1/X1
					700/300	D2D1/X2
					900/100	D2D1/X3
			700/300	D2D2	500/500	D2D2/X1
					700/300	D2D2/X2
					900/100	D2D2/X3
			900/100	D2D3	500/500	D2D3/X1
					700/300	D2D3/X2
					900/100	D2D3/X3
		60°	500/500	D2E1	500/500	D2E1/X1
					700/300	D2E1/X2
					900/100	D2E1/X3
			700/300	D2E2	500/500	D2E2/X1
					700/300	D2E2/X2
					900/100	D2E2/X3
			900/100	D2E3	500/500	D2E3/X1
					700/300	D2E3/X2
					900/100	D2E3/X3
		90°	500/500	D2F1	500/500	D2F1/X1
					700/300	D2F1/X2
					900/100	D2F1/X3
			700/300	D2F2	500/500	D2F2/X1
					700/300	D2F2/X2
					900/100	D2F2/X3
			900/100	D2F3	500/500	D2F3/X1
					700/300	D2F3/X2
					900/100	D2F3/X3

[a] γ is the difference between β (Cluster II) and α (Cluster I) , i.e. $\gamma = \beta - \alpha$.

Table 5: Correlation (Standard Deviation) [a] between $\hat{\theta}$ and θ_T for Panels with Unidimensional Data

Panel	Correlation (Std.) Stage 1	Correlation (Std.) Stage 2	Correlation (Std.) Stage 3
D1/X1	0.85	0.94	0.96
	(0.025)	(0.016)	(0.022)
D1/X2	0.87	0.94	0.96
	(0.021)	(0.037)	(0.027)
D1/X3	0.87	0.94	0.97
	(0.031)	(0.029)	(0.018)

[a] Correlation and standard deviation are based on 200 replications.

Table 6: Correlation (Standard Deviation) [a,] between $\hat{\theta}$ and θ_T by Route for Panels with Unidimensional Data

Panel	Route	N[b]	Correlation (Std.)		
			Stage 1	Stage 2	Stage 3
D1/X1	MHH	481	0.60	0.80	0.88
			(0.031)	(0.028)	(0.017)
	MHM	70	0.11	0.68	0.80
			(0.053)	(0.063)	(0.063)
	MMH	42	0.17	0.55	0.75
			(0.065)	(0.054)	(0.055)
	MMM	404	0.73	0.90	0.94
			(0.026)	(0.034)	(0.027)
D1/X2	MHH	448	0.53	0.79	0.87
			(0.029)	(0.032)	(0.022)
	MHM	55	0.04	0.54	0.78
			(0.048)	(0.051)	(0.047)
	MMH	60	0.36	0.60	0.69
			(0.055)	(0.051)	(0.044)
	MMM	439	0.75	0.90	0.94
			(0.038)	(0.041)	(0.035)
D1/X3	MHH	466	0.55	0.83	0.89
			(0.026)	(0.025)	(0.020)
	MHM	50	0.08	0.74	0.84
			(0.049)	(0.044)	(0.045)
	MMH	63	0.16	0.46	0.81
			(0.066)	(0.051)	(0.055)
	MMM	429	0.74	0.88	0.94
			(0.039)	(0.028)	(0.027)

[a] Correlation and standard deviation are based on 200 replications.
[b] Numbers of examinees have been rounded to integers. They may not add up to 1000 due to averaging over replications.

Table 7: Proportions of Correct and Incorrect Final Routing Decisions for Panels with Unidimensional Data

Panel	Final Routing Decision [a]	Proportion % (Std.) [b,c]
D1/X1	Correct	90.3 (.054)
	Wrong	10.1 (.031)
D1/X2	Correct	90.9 (.049)
	Wrong	9.2 (.029)
D1/X3	Correct	89.1 (.050)
	Wrong	10.8 (.033)

[a] Final routing decisions (based on $\hat{\theta}$) are evaluated against "true" routing decisions (based on θ_T) at the end of Stage 2.
[b] Proportions are averaged over 200 replications.
[c] Proportions may not add up to 100% due to averaging over replications.

Table 8: Results of Pass-Fail Decisions for Panels with Unidimensional Data

Panel	True Status [a]	Estimated Status [b]	Error	Proportion (Std.) [c,d]				
				All	MMM	MHH	MHM	MMH
D1/X1	Pass	Pass		14.1 (.04)	0 (0)	29.0 (.06)	0 (0)	2.3(.02)
	Fail	Fail		78.1 (.06)	100 (0)	55.0 (.07)	100 (0)	95.5(.07)
	Pass	Fail	Type II	4 (.02)	0 (0)	8.1(.03)	0 (0)	2.3(.01)
	Fail	Pass	Type I	3.8 (.02)	0 (0)	7.9(.02)	0 (0)	0(0)
D1/X2	Pass	Pass		11.8(.03)	0 (0)	26.3(.04)	0 (0)	1.9(.01)
	Fail	Fail		81.6(.05)	100 (0)	59.2(.05)	100 (0)	97.4(.08)
	Pass	Fail	Type II	3.6(.02)	0 (0)	7.8(.02)	0 (0)	1.6(.02)
	Fail	Pass	Type I	3(.02)	0 (0)	6.7(.02)	0 (0)	0(0)
D1/X3	Pass	Pass		14.1(.04)	0 (0)	30.3(.04)	0 (0)	2.0(.02)
	Fail	Fail		79.9(.07)	100 (0)	57.1(.05)	100 (0)	97.1(.07)
	Pass	Fail	Type II	2.9(.02)	0 (0)	6.0(.03)	0 (0)	1.7(.02)
	Fail	Pass	Type I	3.1(.03)	0 (0)	6.7(.02)	0 (0)	0.04(.00)

[a] True Status is the pass-fail status based on θ_T at the end of Stage3.
[b] Estimated Status is the pass-fail status based on $\hat{\theta}$ at the end of Stage 3.
[c] Proportions are based on 200 replications.
[d] Proportions may not add up to 100 due to averaging over replications.

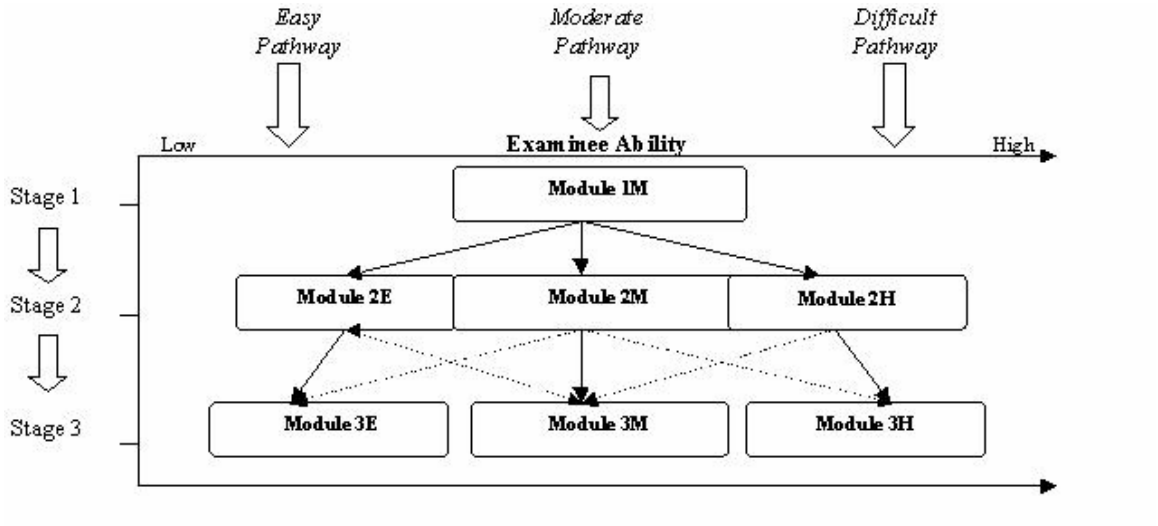


Figure 1
Diagram of a 1-3-3 CAST Panel
(Adapted from Luecht & Nungester, 2000)

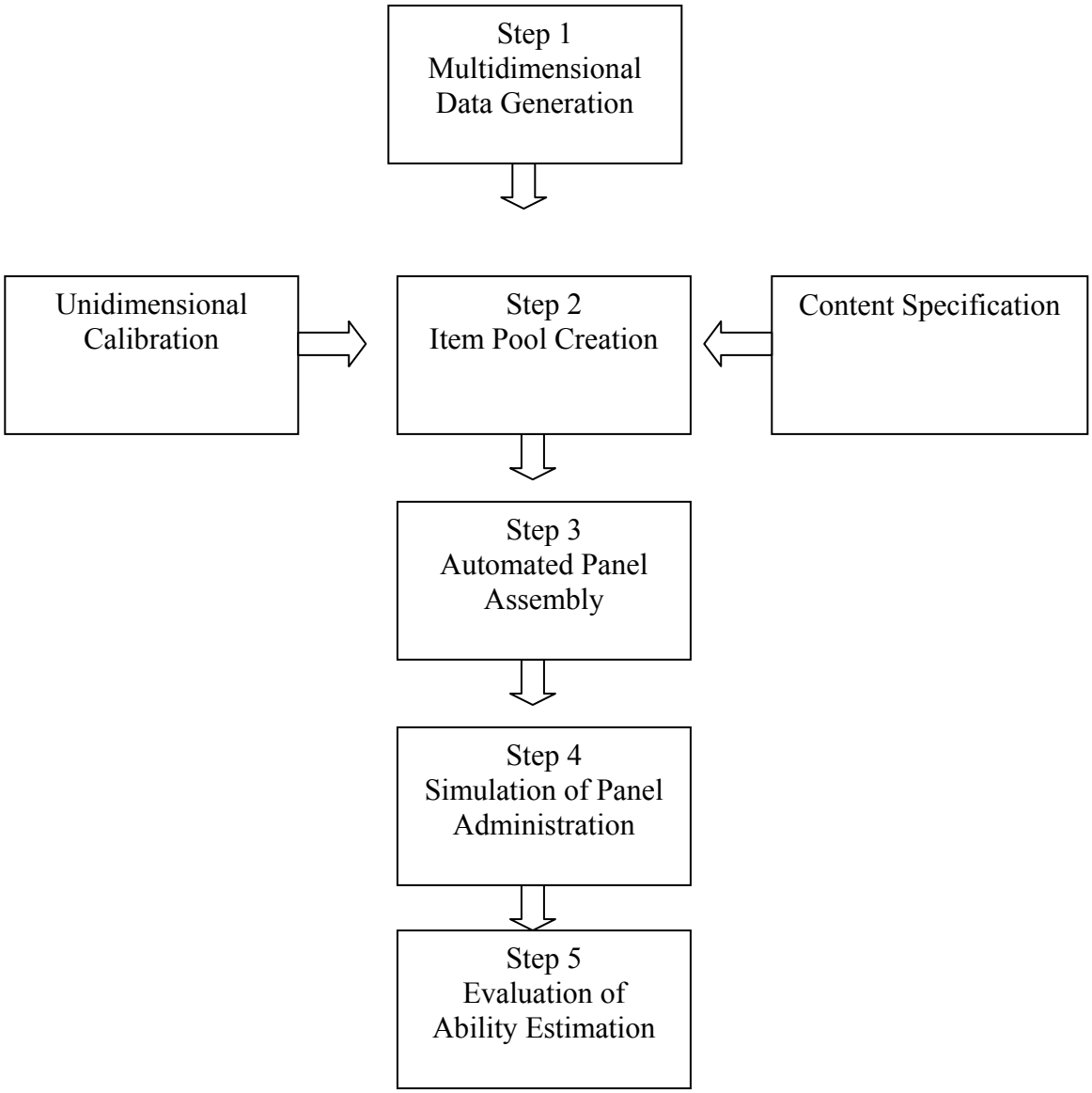


Figure 2
Major Steps of Data Preparation and Analysis

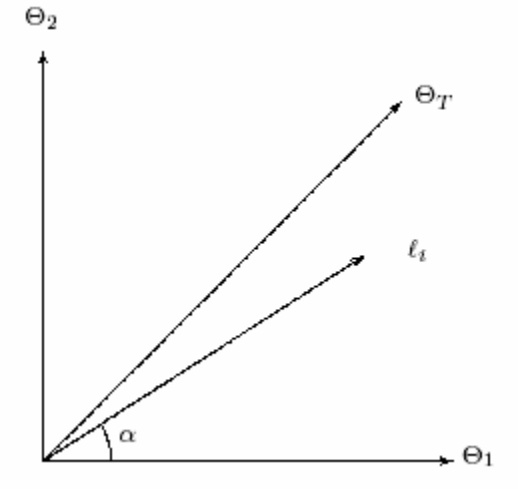


Figure 3
Vector Representation of Items in a Two-Dimensional Space

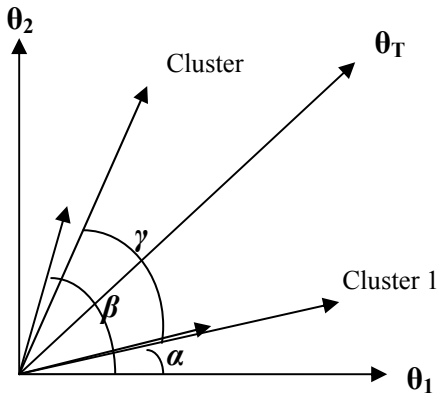


Figure 4
Vector Presentation of Item Clusters in a Two-Dimensional Space

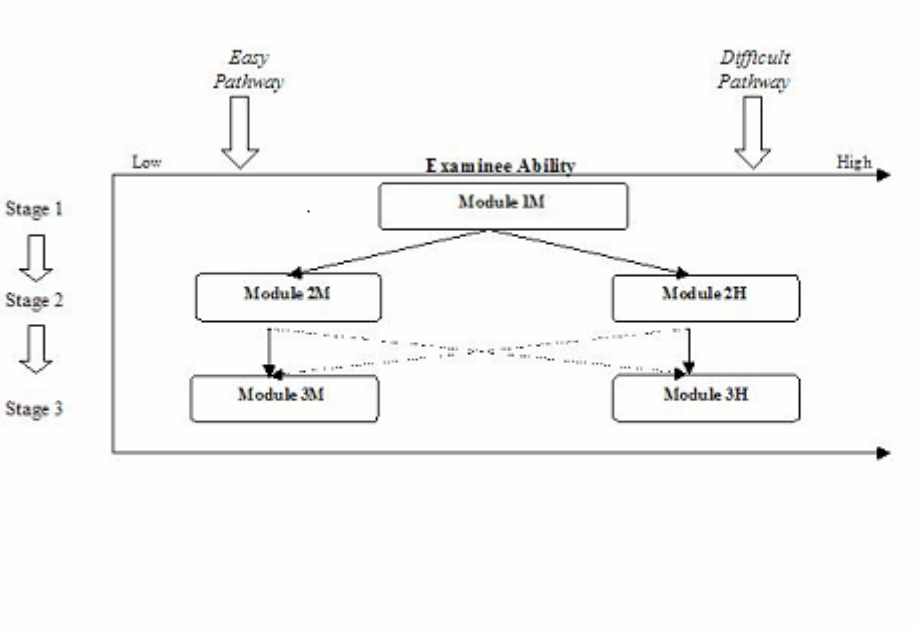
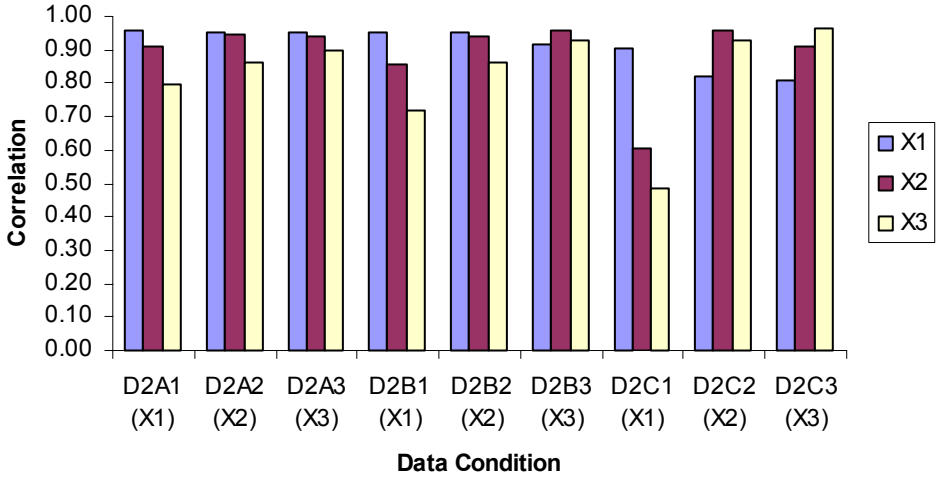
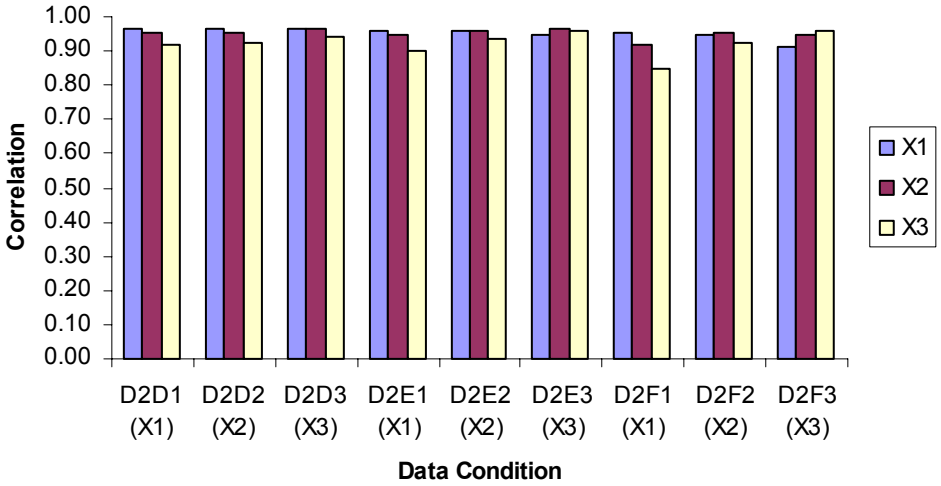


Figure 5
Diagram of a 1-2-2 CAST Panel



$\rho_{\theta_1\theta_2} = 0.3$

(a)

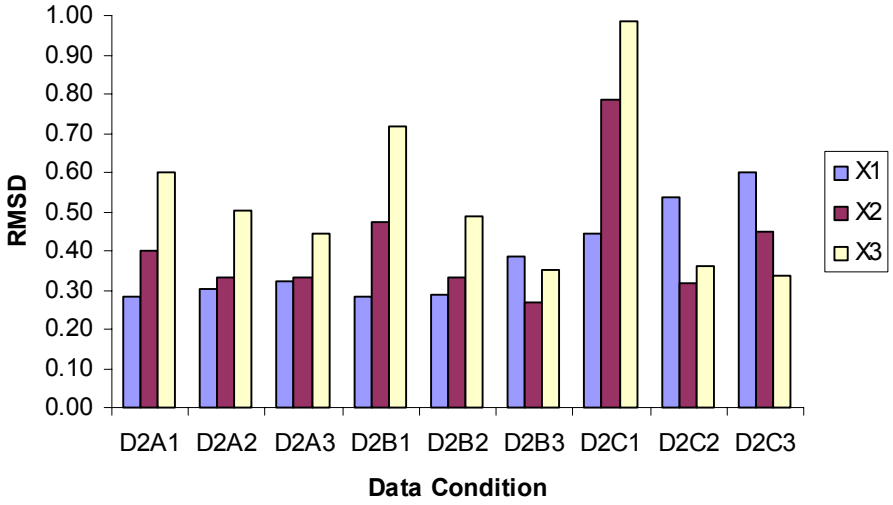


$\rho_{\theta_1\theta_2} = 0.7$

(b)

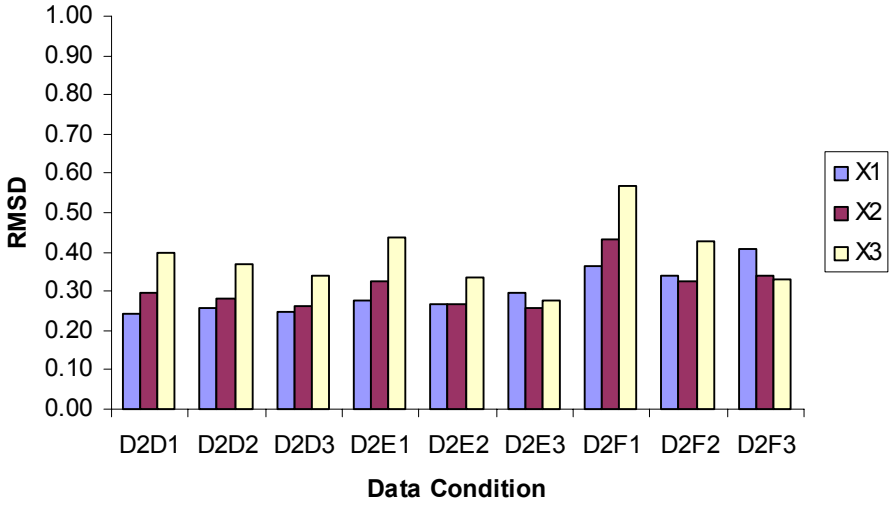
Figure 6

Correlation between $\hat{\theta}$ and θ_T for panels with two-dimensional data



$\rho_{\theta_1\theta_2} = 0.3$

(a)



$\rho_{\theta_1\theta_2} = 0.7$

(b)

Figure 7
RMSD between $\hat{\theta}$ and θ_T for panels with two-dimensional data

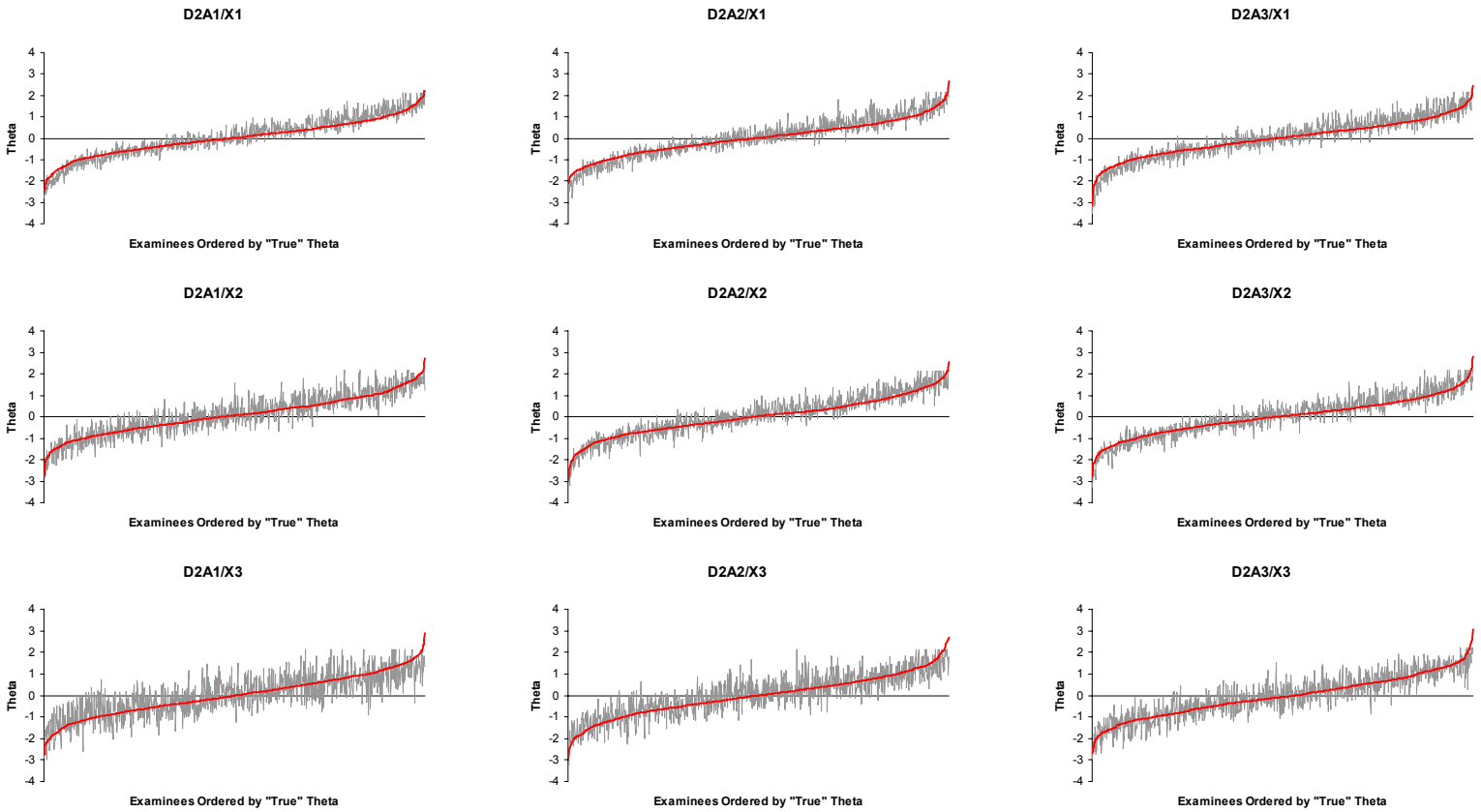


Figure 8

Deviation of $\hat{\theta}$ from θ_T for panels of two-dimensional data: $\rho_{\theta_1\theta_2} = 0.3$ and $\gamma = 30°$

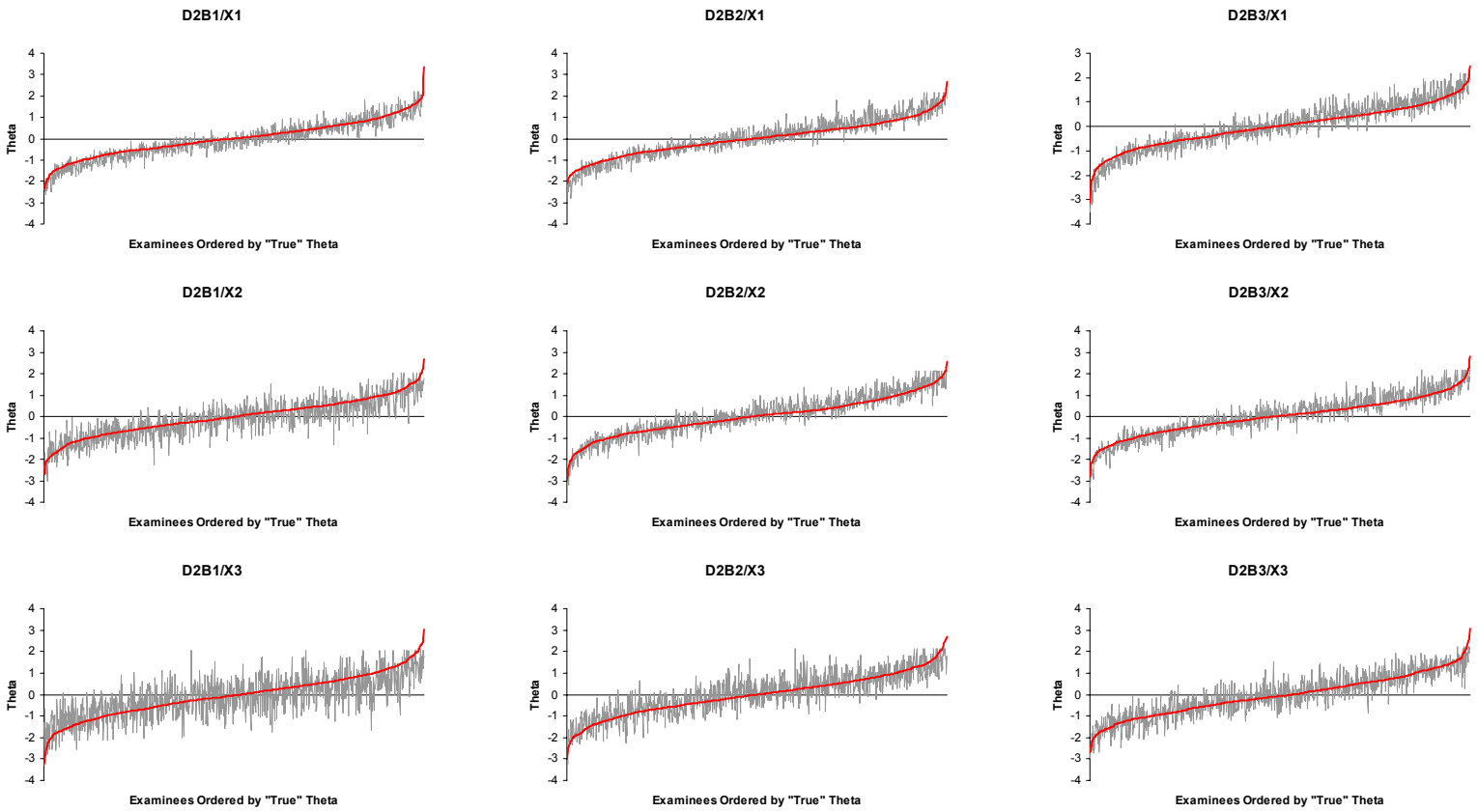


Figure 9

Deviation of $\hat{\theta}$ from θ_T for panels of two-dimensional data: $\rho_{\theta_1\theta_2} = 0.3$ and $\gamma = 60°$

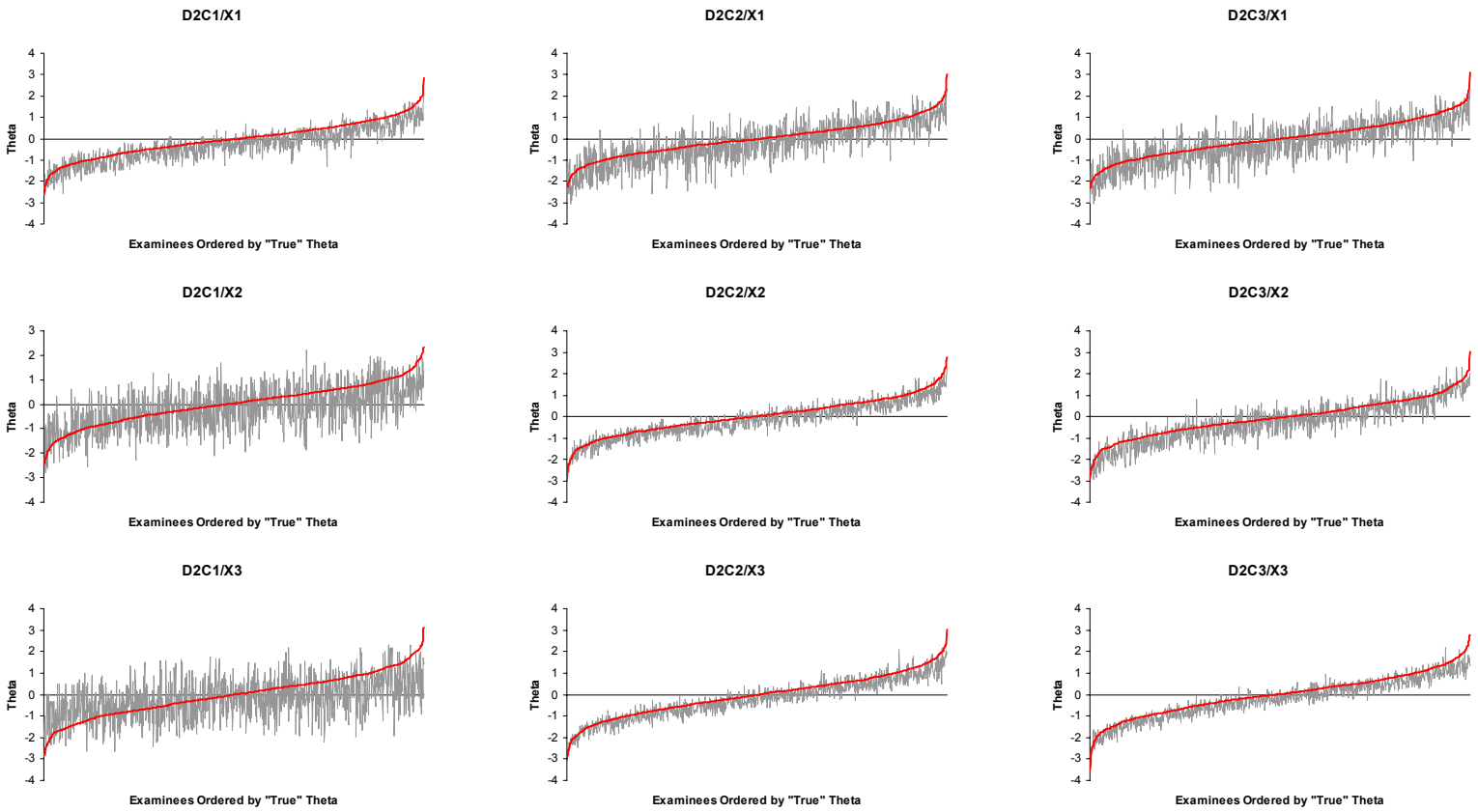
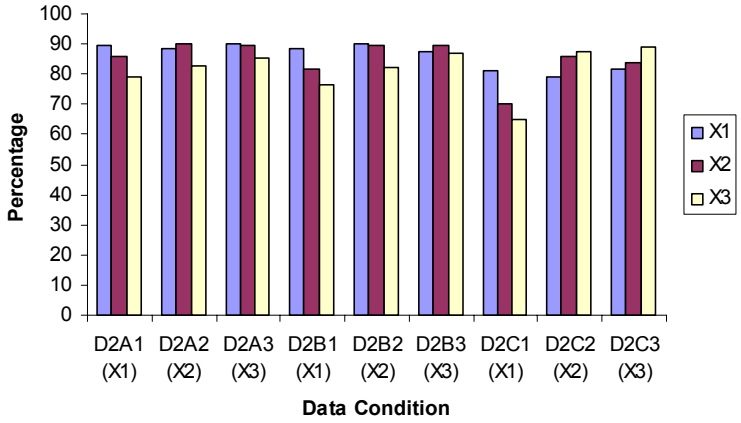


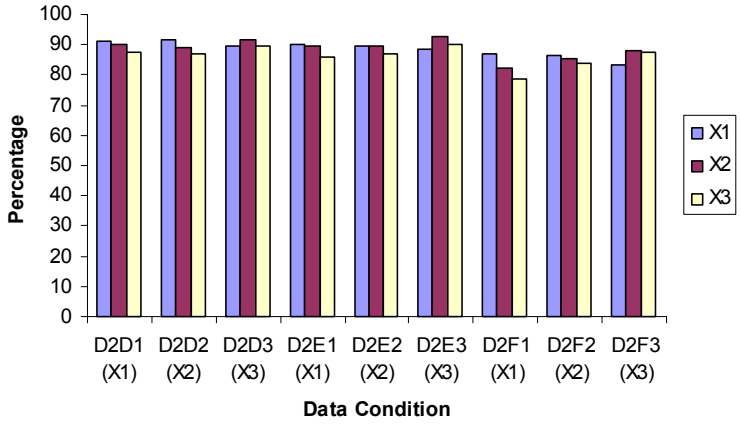
Figure 10

Deviation of $\hat{\theta}$ from θ_T for panels of two-dimensional data: $\rho_{\theta_1\theta_2} = 0.3$ and $\gamma = 90°$



$\rho_{\theta_1\theta_2} = 0.3$

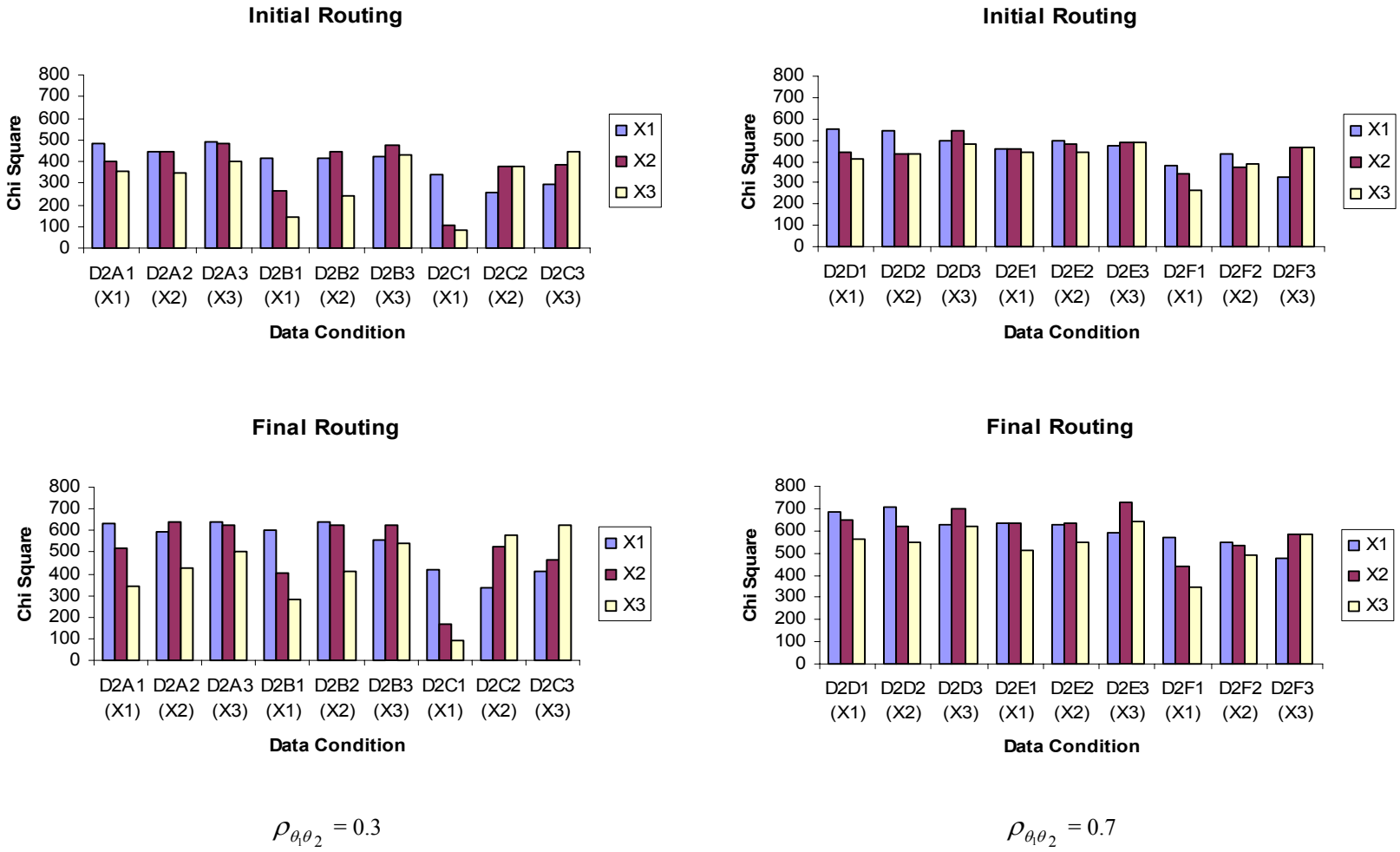
(a)



$\rho_{\theta_1\theta_2} = 0.7$

(b)

Figure 11
Proportions of correct final routing decisions
for panels with two-dimensional data



$\rho_{\theta_1\theta_2} = 0.3$ $\rho_{\theta_1\theta_2} = 0.7$

Figure 12
Chi-square statistics on the independence between estimated routing decisions and "true" routing decisions

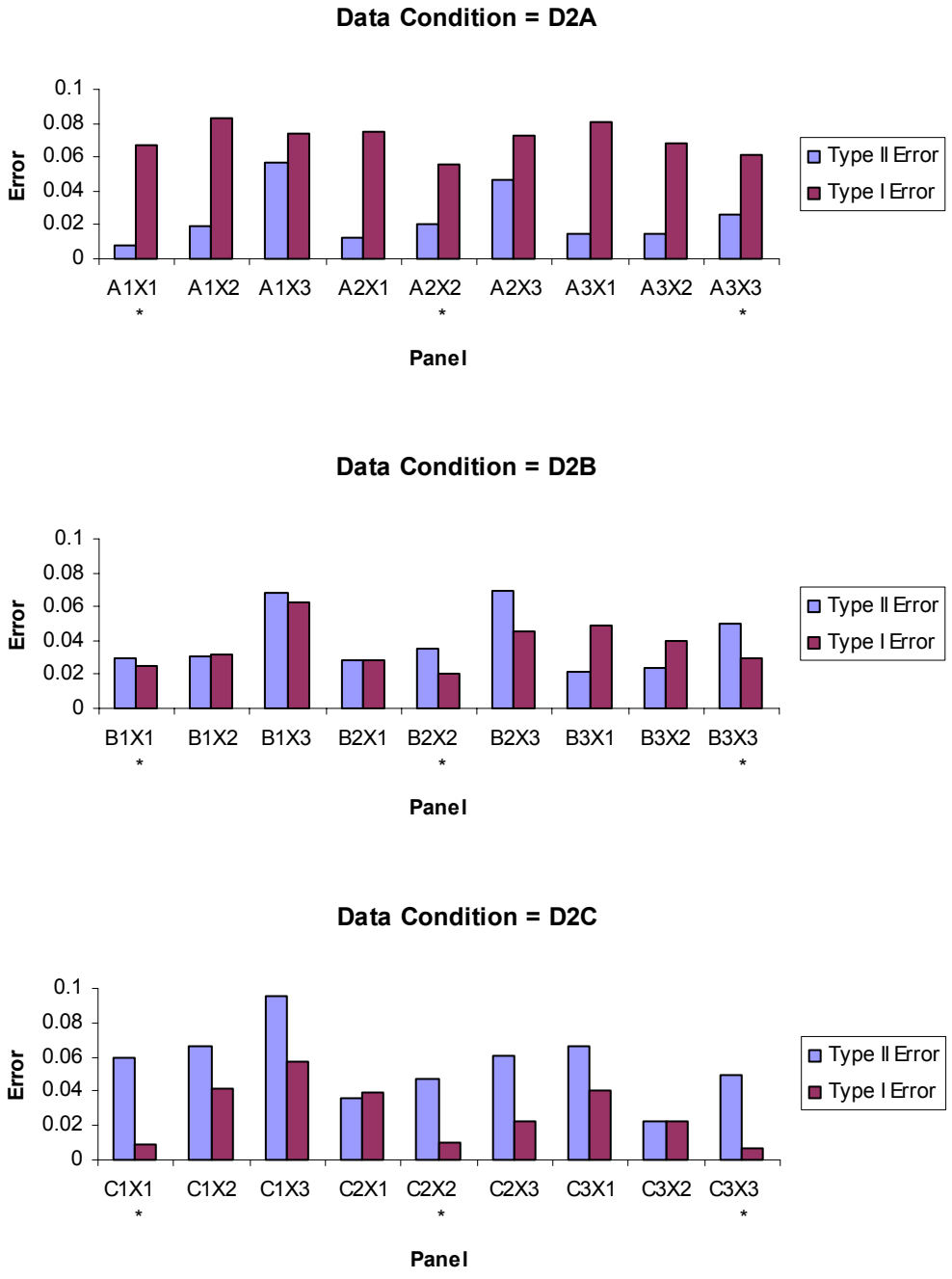
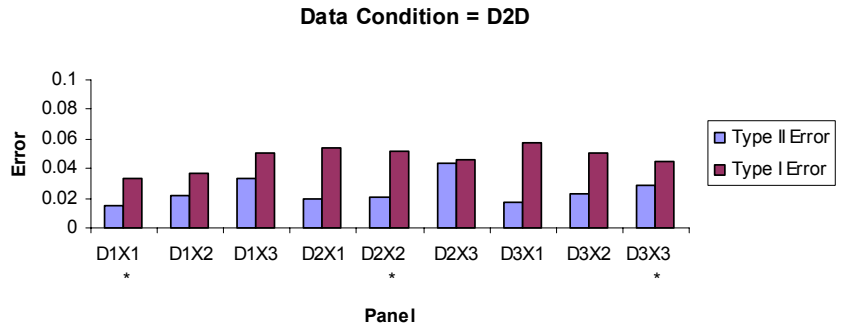
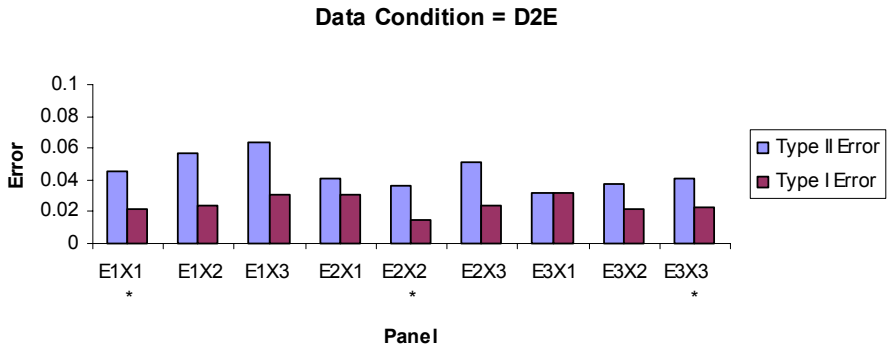


Figure 13
Type I and Type II errors in pass-fail decisions for the two-dimensional data ($\rho_{\theta_1\theta_2} = 0.3$)

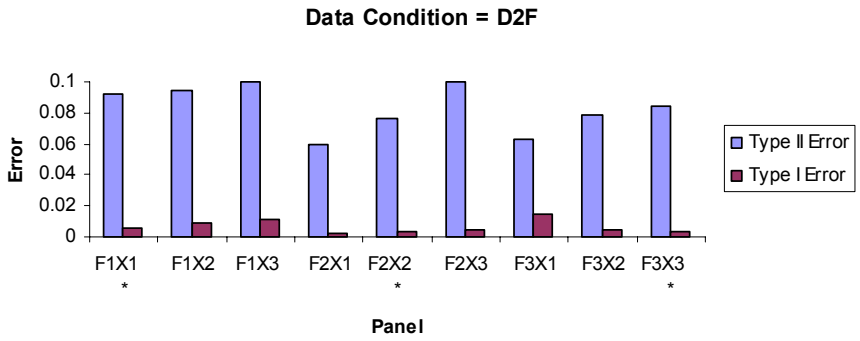
Note: The panel with asterisk has the perfect match between content and dimension for that data condition.



(a)



(b)



(c)

Figure 14
Type I and Type II errors in pass-fail decisions for the two-dimensional data ($\rho_{\theta_1\theta_2} = 0.7$)

Note: The panel with asterisk has the perfect match between content and dimension for that data condition.