Grid Multi-classification Adaptive Classification Testing

with Multidimensional Polytomous Items

A THESIS

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Zhuoran Wang

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Advisor: Chun Wang
Co-Advisor: Niels Waller

August 2019
Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in graduate school. The following is only an incomplete list of those I have met along the way.

Firstly, I would like to express my sincere gratitude to my advisor Dr. Chun Wang for the continuous support of my Ph.D. study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D. study.

In addition to my advisor, I would like to thank the rest of my thesis committee: Dr. David Weiss, Dr. Niels Waller, and Dr. Mark Davison, for their insightful comments and encouragement, but also for their thought-provoking courses and support on my multiple course projects.

I thank my colleagues on the sixth floor of Elliott Hall, Lauren Berry, Allie Cooperman, Kayle Donner, Leah Feuerstahler, Shengyu Jiang, Justin Kracht, Hoang Nguyen, Alec Nyce, Chaitali Phadke, Gretchen Saunders, Matthew Snodgress, King Yiu Suen, and Ziming Zhou, for the stimulating discussions and for all the fun we have had in the last five years.
Last but not least, I would like to express my deepest gratitude to my family and friends. This thesis would not have been possible without their warm love, continued patience, and endless support.

This research was supported by the Eunice Kennedy Shriver National Institutes of Child Health and Human Development of the National Institutes of Health under Award Number R01HD079439 to the Mayo Clinic in Rochester Minnesota through subcontracts to the University of Minnesota and the University of Washington. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.
Abstract

Adaptive classification testing (ACT) is a form of computerized adaptive testing (CAT) that was developed to efficiently classify examinees into multiple categories based on predetermined classification cutoff scores. All existing multidimensional ACT studies handle multidimensional classifications in a unidimensional space by performing classification on a composite of multiple traits. However, classification along separate dimensions is sometimes preferred because it provides clearer information regarding a person’s relative standing along each dimension. This type of classification is referred to as grid classification, as each examinee is classified into one of the grids encircled by cutoff scores (lines/surfaces) on different dimensions. Complications arise when there is more than one cutoff score along each dimension. In order to perform grid classification using ACT, two termination criteria, sequential probability ratio test (SPRT) and confidence interval (CI) were adopted from one-dimensional classification in the between-item multidimensional test. In addition, two new termination criteria for grid multi-classification ACT were developed, namely, grid classification generalized likelihood ratio (GGLR) and simplified generalized likelihood ratio (SGLR). A new item selection rule, i.e., posterior weighted D-optimal on cutoff points (PWCD-optimal), was also proposed.

Three simulation studies were conducted to evaluate the performance of ACT for multidimensional grid classification. The three-dimensional multidimensional graded response model (MGRM) with four response categories was used. The item bank contained 300 between-item multidimensional items with 100 items loading on each
dimension. Examinees were classified into four groups along each dimension, resulting in $4^3 = 64$ classification grids in total. The minimum and maximum test length were fixed at 7 and 60 items. Three item selection methods (D-optimal, PWCD-optimal, and multidimensional mutual information (MMI)) and four termination criteria (GGLR, SGLR, CI, and SPRT) were applied in the grid multi-classification ACT.

The first study compared ACT to the two-step measurement CAT-based classification. In the latter scenario, a variable-length multidimensional CAT was conducted, followed by a post-hoc classification. The D-optimal item selection method and the compound termination criteria were used in the two-step approach. The cutoffs for each termination criterion in the two approaches were selected to carefully yield similar classification accuracy, such that the resulting average test length (ATL) was a useful indicator of test efficiency. Results showed that, when D-optimal and PWCD-optimal item selection methods were used, ACT resulted in up to 20% shorter ATLs than the two-step approach. In this way, ACT was more efficient than the two-step approach. Among the four termination criteria for ACT, SPRT, and CI outperformed the two new termination criteria.

The second study further explored the influence of true latent trait location on classification accuracy and test length using grid multi-classification ACT. Instead of simulating discrete $\theta$ points, $\theta$ distributions with various mean vectors and variance-covariance matrices were used to represent true latent traits at different locations. One, two, or three dimensions were manipulated. For the manipulated dimensions, six $\theta$ mean levels and two $\theta$ standard deviation levels were considered. Generally, classification was
more difficult when examinees were closer to the cutoff scores. PWCD-optimal and D-optimal lead to stable test length and classification accuracy. As SPRT and CI resulted in lower classification accuracy for examinees that were close to the cutoff points, thus were difficult to be classified, the overall high efficiency of SPRT and CI in Study 1 can be largely attributed to the large proportion of examinees that were far from the cutoff points in the normally distributed population. As stable classification accuracy across $\theta$ distribution is generally desired in ACT, SGLR and GGLR were found to be preferable.

The third study utilized real item parameters from a health measurement bank containing 324 Likert-scale items and 366 real examinee parameters to compare the performance of ACT and the two-step measurement CAT-based approach in terms of grid classification. Due to the influence of item bank quality, the test lengths were much longer than those in Study 1. All the conditions using MMI resulted in test length longer than the maximum test length (60 items). When D-optimal and PWCD-optimal item selection methods were used, ACT still outperformed the two-step approach and saved up to 20% of items as in Study 1. However, the superiority of PWCD-optimal did not always hold. CI and SPRT still led to shorter ATL than the other two termination criteria.
Table of Contents

1. Introduction ........................................................................................................................................... 1

2. IRT Models in ACT .................................................................................................................................. 6
   2.1 Unidimensional Models .................................................................................................................. 6
   2.2 Multidimensional Models .............................................................................................................. 9

3. Termination Methods in ACT ........................................................................................................... 12
   3.1 Termination in Unidimensional ACT ........................................................................................... 12
   3.1.1 Sequential Probability Ratio Test (SPRT) .................................................................................. 13
   3.1.2 Truncated SPRT (TSPRT) ........................................................................................................ 15
   3.1.3 SPRT in Multi-classification ACT ............................................................................................ 16
   3.1.4 Curtailed SPRT (CSPRT) ......................................................................................................... 18
   3.1.5 Stochastic Curtailed SPRT (SCSPRT) ....................................................................................... 20
   3.1.6 SPRT with Predictive Power (PPSPRT) .................................................................................... 23
   3.1.7 Generalized Likelihood Ratio (GLR) ......................................................................................... 24
   3.1.8 Confidence Interval (CI) ........................................................................................................... 26
   3.2 Termination in Multidimensional ACT with a Composite Score ............................................... 27
   3.2.1 Constrained SPRT (C-SPRT) and Projected SPRT (P-SPRT) ..................................................... 28
   3.2.2 SPRT with Reference Composite ............................................................................................... 30
   3.2.3 Between-item Multidimensional SPRT for an Entire Test ......................................................... 31
3.2.4 Multidimensional GLR (MGLR) ................................................................. 33
3.2.5 Weighted GLR (WGLR) ........................................................................ 34
3.2.6 Wita hin-item Multidimensional Confidence Interval ....................... 35

3.3 Termination in Multidimensional ACT with Grid Classification .......... 36
3.3.1 Grid Classification GLR (GGLR) .............................................................. 36
3.3.2 Simplified Grid Classification GLR (SGLR) ......................................... 39
3.3.3 Between-item Multidimensional CI ...................................................... 40
3.3.4 Between-item Multidimensional SPRT ................................................. 41

4. Item Selection Methods in ACT ................................................................. 43
4.1 Item Selection Methods in Unidimensional ACT .................................. 43
4.1.1 Maximum Fisher Information ............................................................... 43
4.1.2 Maximum Weighted Fisher Information (WFI) ............................... 45
4.1.3 Kullback-Leibler Information (KL) .................................................... 47
4.1.4 Weighted Log Odds (WLO) ................................................................. 50
4.1.5 Expected Log Likelihood Ratio (ELR) .................................................. 51
4.1.6 Mutual Information (MI) .................................................................... 52

4.2 Item Selection Methods in Multidimensional ACT .............................. 53
4.2.1 Maximum Determinant of the Fisher Information Matrix (D-optimal) .. 53
4.2.2 Maximum Trace of Inverse Fisher Information Matrix (A-optimal) .... 56
4.2.3 Maximum Fisher Information of Composite Ability (C-optimal) ................................................. 56
4.2.4 Smallest Eigenvalue of Fisher Information Matrix (E-optimal) .............................................. 57
4.2.5 D-optimal with Posterior Weights on Cutoff Points (PWCD-optimal) ................................ 58
4.2.6 Multidimensional KL Information (MKL) .................................................................................. 59
4.2.7 PWKL with Posterior Weights on Cutoff Points (PWCKL) ...................................................... 61
4.2.8 Multidimensional Mutual Information (MMI) ........................................................................... 62
4.2.9 MMI with Posterior Weights on Cutoff Points (PWCMMI) ..................................................... 62
4.3 Summary ...................................................................................................................................... 64
5. Study 1: ACT and Measurement CAT Classification Efficiency ............................................... 66
5.1 Method ........................................................................................................................................ 66
5.1.1 Item Bank and IRT Model ........................................................................................................ 66
5.1.2 Classification Settings .............................................................................................................. 67
5.1.3 Latent Trait Distribution ........................................................................................................... 69
5.1.4 Item Selection Methods ........................................................................................................... 70
5.1.5 Termination Criteria ................................................................................................................ 73
5.1.6 Interim Ability Estimation ....................................................................................................... 74
5.1.7 Evaluation Criteria ................................................................................................................... 75
5.1.8 Overall Conditions .................................................................................................................. 76
5.2 Results ......................................................................................................................................... 76
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 2</td>
<td>Conditional Classification Efficiency of ACT</td>
<td>82</td>
</tr>
<tr>
<td>Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Bank and Classification Settings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent Trait Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Selection Methods and Termination Criteria</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Study 3</td>
<td>Evaluation of ACT and Measurement CAT with Hybrid Simulation</td>
<td>96</td>
</tr>
<tr>
<td>Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Bank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent Trait Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion and Conclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Necessity of Grid Classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advantages of ACT over the Two-step Approach in Grid Classification</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evaluation of Grid Classification Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion and Conclusion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item Bank</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latent Trait Distribution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparison among Termination Criteria</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
8.6 Conditional Classification Efficiency ................................................................. 111
8.7 Classification with ACT or Diagnostic Classification Modeling (DCM) .......... 112
8.8 Conclusion and Future Studies ........................................................................... 113

References .................................................................................................................. 115
List of Tables

Table 1. The two-step approach and ACT classification efficiency ................................ 78
Table 2. Latent trait distributions ......................................................................................... 87
List of Figures

Figure 1. Multi-classification GLR ................................................................. 26
Figure 2. Relevant cutoff points ................................................................. 37
Figure 3. GGLR for grid classification ................................................... 38
Figure 4. SGLR for grid classification .................................................... 39
Figure 5. Possible standard error by fixing $\theta_2$ and $\theta_3$ to 0 ............ 69
Figure 6. Classification and estimation accuracy with varying item selection methods . 71
Figure 7. Classification and estimation accuracy with varying number of quadrature points ................................................................. 73
Figure 8. Average test length with varying item selection methods and termination criteria ................................................................. 80
Figure 9. Determinant of item bank Fisher information by fixing $\theta_2$ and $\theta_3$ to 0 ....... 83
Figure 10. Determinant of possible test Fisher information by fixing $\theta_2$ and $\theta_3$ to 0 .... 84
Figure 11. Average test length with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions ..................... 90
Figure 12. CCR of dimension 1 with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions .......... 91
Figure 13. CCR of dimension 2 with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions .......... 92
Figure 14. CCR of dimension 3 with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions .......... 93
Figure 15. Dimensional information of the real item bank ...................... 98
Figure 16. Ability distribution on each dimension of the real examinees ..................... 99

Figure 17. Average test length with varying item selection methods and termination criteria ........................................................................................................................................ 101
1. Introduction

Classification testing refers to the family of tests in which examinees are classified into two or more groups based on predetermined classification cutoff points. A classification test determines whether an examinee meets the requirement for a particular purpose (Norcini & Guille, 2002). If the goal of a test is to classify people, it is unnecessary to obtain very precise ability/latent trait estimate on a continuous scale. Instead, the coarser-grained classification testing is more efficient. Classification testing is widely used in educational (Eggen & Straetmans, 2000), occupational, and medical (Smits & Finkelman, 2013) areas. For example, it is common to classify examinees into multiple fluency levels in the language testing domain. Interagency Language Roundtable uses 0, 0+, 1, …, 3 to characterize spoken-language use. Occupational certificate and license tests essentially categorize examinees into passing or failing groups, then grant the passing group the corresponding certificate or license. Clinical questionnaires also classify examinees. Beck’s Depression Inventory classifies examinees from “These ups and downs are considered normal” to “Extreme depression”. Classification testing further facilitates subsequent treatment (after clinic diagnosis) or teaching (after competence-based class grouping). Within the treatment, multiple participants can work together to promote treatment as well as provide more targeted feedback for the plan to improve (Welch & Frick, 1993).

To administer the classification test in a more efficient manner, computerized adaptive testing (CAT) can be adopted as the testing format. CAT has gained great
popularity because it shortens test length while maintaining high precision of ability estimates. To avoid ambiguity, in this thesis, CAT integrated with classification testing is called Adaptive Classification Testing (ACT), whereas CAT used to accurately estimate examinees’ abilities is referred to as measurement CAT. Measurement CAT can also be used for classification purpose. After it is stopped by either precision-related termination criteria or fixed test length, the ability estimate is compared to the predetermined cutoff point so as to arrive at a classification result. This is a two-step procedure, while ACT directly arrives at the classification decision. The advantage brought by adaptive algorithms tends to be even more compelling in ACT than in the measurement CAT-based two-step approach (referred to as two-step approach in the rest of this thesis), as the examinees, who are located far from the cutoff points, can be classified with a few items and relatively rough ability estimate. In these cases, ACT can sometimes be cut short. In order to reach the same level of classification accuracy, test lengths vary among examinees. That is, variable-length tests can fully exert the advantage of adaptiveness in classification testing. Hence, the ACTs are of variable length. Lewis and Sheehan (1990) found that ACT reduced the test length up to 50% on average without sacrificing classification accuracy.

Dimensionality is an unavoidable issue in the testing area. Traits are essentially associated, thus some of the tests are, by nature, multidimensional through using items measuring multiple traits. Furthermore, combining several unidimensional tests measuring different dimensions takes advantage of the information brought by other dimensions, thus it should result in a more accurate ability estimate or higher
classification accuracy than using multiple unidimensional tests separately (Frey & Seitz, 2009; Seitz & Frey, 2013; Luecht, 1996; Segall, 1996). As a result, multidimensional testing has gained great popularity recently. This trend is also true in ACT, reflected by the increasing number of studies on multidimensional ACT (van Groen, 2014; van Groen, Eggen, & Veldkamp, 2016; Nydick, 2013; Seitz & Frey, 2013). However, all of the existing multidimensional ACT studies simplify the multidimensional classification problems to unidimensional ones, hence classification explicitly on each dimension is not viable. On the other hand, classification on each of the original dimensions has more practical meaning and can provide guidance to the following treatment and teaching. This kind of cross-classification in multidimensional classification testing is named grid classification in this thesis, as examinees are classified into one of the grids encircled by cutoff scores (lines/surfaces) on different dimensions.

Moreover, the number of categories also has a sizeable influence on the classification procedure. ACT started as mastery testing, which classifies examinees into only two categories: master and non-master, or pass and fail (Lewis & Sheehan, 1990; Spray et al., 1997; Weiss & Kingsbury, 1984). Occupational certification testing often employs this kind of ACT. Later on, multi-classification tests (examinees are classified into more than two categories) emerged in areas such as language testing and physical evaluations (Glas & Vos, 2009; van Groen, 2014; van Groen, Eggen, & Veldkamp, 2014; Seitz & Frey, 2013). The majority of the recent ACT studies focus on mastery testing, while ignoring the more general, yet more complicated multi-classification cases.
Last but not least, polytomous items need to be promoted in ACT construction. Polytomous items have been broadly used in a variety of exams due to the prevalence of Likert-scale items and partial credit items. In addition, there are several strengths from a psychometric perspective. One of these strengths is that polytomous items provide more information than dichotomous items over a wider range (Samejima, 1976; Birenbaum & Tatsuoka, 1987; Donoghue, 1994). In this way, polytomous items could reduce the test length, while achieving the same effects as dichotomous items, particularly under a CAT context (De Ayala, 1989; De Ayala, Dodd, & Koch, 1992). An additional benefit is that item exposure is rather balanced even without deliberate control (van Rijin et al., 2002). It is also believed that polytomous items measure concepts and skills at greater depth than dichotomous items (Ercikan et al., 1998). However, polytomous items are underused in ACT (Smits & Finkelman, 2013; Thompson, 2007) and they have never been used in multidimensional ACT.

In order to address the above issues and enrich the multidimensional ACT research literature, this thesis focuses on designing efficient multidimensional multi-classification ACT with polytomous items. From a practical point of view, tests entailing items measuring exactly one dimension each (between-item multidimensionality) are much more common than tests based on an item pool with within-item multidimensionality (Seitz & Frey, 2013). Thus, this thesis only addresses ACT with between-item multidimensionality. In the following sections, the prevalent IRT models as well as the current literatures on ACT item selection methods and termination criteria are reviewed. A few new item selection methods and termination criteria, specifically
developed for the grid multi-classification ACT, using polytomous items are proposed. Finally, three simulation studies are implemented to compare ACT to the two-step approach, in terms of classification efficiency, as well as evaluate the performance of item selection methods and termination criteria in ACT.
2. IRT Models in ACT

Selecting an appropriate IRT model for the data is an important precondition of ACT administration. Please note, ACT and measurement CAT can share the same IRT model as well as the item bank. Thus, the models listed below can be used in both ACT and the two-step approach. As a large proportion of ACT studies were conducted using unidimensional models and multidimensional dichotomous models, all these models as well as multidimensional polytomous models, which are the core of this thesis, are reviewed here.

2.1 Unidimensional Models

The most popular unidimensional IRT model is the three-parameter logistic (3PL) model (Birnbaum, 1968). Its popularity spans across both linear testing (all examinees respond to the same test irrespective of their latent trait levels) and adaptive testing, and both measurement CAT and ACT (Lin & Spray, 2000; Thompson & Ro, 2007; Weissman, 2007; Bartroff, Finkelman, & Lai, 2008; Thompson, 2010; Lin, 2011). It is a dichotomous model. That is, for each item there are only two scoring options: 1 and 0, indicating correct and incorrect. The probability of examinee $i$ correctly responding to item $j$ is

$$P_j(\theta_i) = c_j + \frac{1 - c_j}{1 + \exp(-D a_j(\theta_i - b_j))} \quad (1)$$
where $\theta_i$ is the ability of examinee $i$. As this study focuses on adaptive testing, each examinee is assumed to finish the test independently. From here on, the examinee index $i$ is dropped. $a_j$, $b_j$ and $c_j$ are the discrimination, difficulty, and pseudo-guessing parameters of item $j$, respectively. When $c_j = 0$, the 3PL model becomes the two-parameter logistic (2PL) model. 2PL model is also used in numerous ACT studies (Eggen, 1999, 2009; Eggen & Straetmans, 2000; Wonda & Eggen, 2009). Moreover, when the $a_j$s are equal across the entire test, the one-parameter logistic (1PL) model is obtained.

Different from dichotomous IRT models, polytomous IRT models are scored on more than two options, from 1 to $R$ ($R > 2$). Generally, polytomous items provide wider-spread and higher item information which potentially can shorten the test length in ACT and measurement CAT while keeping accurate classification and estimation. The conventional polytomous IRT models include the partial credit model (PCM; Master, 1982), the graded response model (GRM) (Samejima, 1968), the generalized partial credit model (GPCM; Muraki, 1992), the nominal response model (NRM; Bock, 1972) and the rating scale model (RSM; Andrich, 1978; Muraki, 1990). Although all polytomous models can deal with the same format of responses, they have different assumptions about the response process. According to Dodd, De Ayala, and Koch (1995), the GRM, GPCM, and PCM can be adopted for data ordered to represent various degrees of the latent trait measure. The PCM model is specifically fit for items in mathematics, physics, and chemistry because points are awarded for the completion of steps leading to the correct answer. The GPCM is a similar version of the PCM model with the exception
of having varying slope parameters in different items. The GRM model is frequently used in analyzing Likert-scale item responses. As both the GPCM and GRM models have been used in ACT studies (Lau & Wang, 1998, 1999; Thompson, 2007; Gnambs & Batinic, 2011; Smits & Finkelman, 2013), they are reviewed in detail.

The GPCM model is formulated on the assumption that the probability of choosing option $r$ over option $r-1$ in item $j$ is governed by the logistic response model

$$C_{jr} = \frac{P_{jr}(\theta)}{P_{j,r-1}(\theta) + P_{jr}(\theta)} = \frac{1}{1 + \exp\left(-a_j(\theta - b_{jr})\right)}$$  \hspace{1cm} (2)

It can then be written as

$$P_{jr}(\theta) = \frac{C_{ijr}}{1 - C_{ijr}}P_{j,r-1}(\theta) = \exp\left(a_j(\theta - b_{jr})\right)P_{j,r-1}(\theta)$$ \hspace{1cm} (3)

Note that $C_{ijr}/(1 - C_{ijr})$ is the odds of choosing the option $r$ instead of option $r - 1$, given these two available choices. After normalizing each $P_{jr}(\theta)$ within an item such that $\Sigma P_{jr}(\theta) = 1$, the GPCM model is written as

$$P_{jr}(\theta) = \frac{\exp\left(\sum_{v=1}^{r} a_j(\theta - b_{jv})\right)}{\Sigma_{m=1}^{R} \exp\left(\sum_{v=1}^{m} a_j(\theta - b_{jv})\right)}$$ \hspace{1cm} (4)

where $a_j$ is the slope parameter, and $b_{jv}$ is the $v^{th}$ threshold parameter of item $j$. For an item with $R$ options, only $R - 1$ threshold parameters can be identified. For this reason, it is common to arbitrarily define $b_{j1} \equiv 0$.

The GRM model is a generalization of the 2PL to multiple scoring options. It requires the scoring options to be ordered, such as in Likert-scale items that are
extensively used in clinical assessment and questionnaires. GRM is an “indirect” model of option response. \( P_{jr}(\theta) \), the probability of scoring in option \( r \) on item \( j \) is the difference between scoring in option \( r \) or higher and option \( r + 1 \) or higher. That is,

\[
P_{jr}(\theta) = P_{jr}^*(\theta) - P_{j(r+1)}^*(\theta)
\]

(5)

where

\[
P_{jr}^*(\theta) = \frac{1}{1 + \exp(-D a_j(\theta - b_{jr}))}
\]

(6)

\( a_j \) is the slope parameter, and \( b_{jv} \) is the \( v^{th} \) boundary parameter of item \( j \). \( P_{jr}^*(\theta) \) is the probability of selecting option \( r \) or higher. \( P_1^*(\theta) \equiv 1 \) and \( P_{R+1}^*(\theta) \equiv 0 \) always hold. In this way, an item with \( R \) options, only \( R - 1 \) intercept parameters can be identified (\( b_{j1} \) is defined by \( P_1^*(\theta) \equiv 1 \)).

2.2 Multidimensional Models

Multidimensional IRT (MIRT) is a generalization of unidimensional IRT, when an examinee in the former has a vector of latent traits while an examinee in the latter has a single latent trait. Therefore, MIRT assumes that a set of \( K \) abilities account for the examinee’s response to an item.

The compensatory multidimensional 2PL model (M2PL; Reckase, 1985) is the most widely used MIRT model. It is the multidimensional counterpart of the
unidimensional 2PL model. Its compensatory feature allows high ability on one
dimension to make up for low ability on other dimensions. The item response function is

\[ P_j(\theta) = \frac{1}{1 + \exp \left( -D \sum_{k=1}^{K} a_{jk} (\theta_k - b_j) \right)} \]  

(7)

where \( P_j(\theta) \) is the probability of correctly answering item \( j \). Each item has multiple
discrimination parameters but only one difficulty parameter. \( a_{jk} \) is the discrimination
parameter of item \( j \) along dimension \( k \). \( b_j \) is the difficulty parameter of item \( j \). \( \theta \) is the
multidimensional ability which has \( \theta_k \) as the \( k^{th} \) element.

The multidimensional GRM (MGRM; Ferrando & Chico, 2001) is a
multidimensional generalization of the GRM. Same as the GRM, it is used to analyze
Likert-scale data; \( P_{jr}(\theta) = P_{jr}^*(\theta) - P_{j(r+1)}^*(\theta) \) is still valid. The probability of selecting
option \( r \) on item \( j \) is

\[ P_{jr}^*(\theta) = \frac{1}{1 + \exp \left( -D \sum_{k=1}^{K} a_{jk} (\theta_k - b_{jr}) \right)} \]  

(8)

where \( b_{jr} \) is the boundary parameter of category \( r \) and \( a_{jk} \) is the discrimination parameter
along dimension \( k \) in item \( j \). \( P_1^*(\theta) \equiv 1 \) and \( P_{R+1}^*(\theta) \equiv 0 \) still hold. As the MGRM is
widely used in the clinical area (Forero et al., 2013; Hsieh, Eye, & Maier, 2010) where
multi-classification is needed, it is used in this thesis as a representative of
multidimensional polytomous models. Referring back to the unidimensional IRT model
for comparison with MIRT, it can be seen the scalars ability \( \theta \) and discrimination
parameter \( a \) become \( K \)-dimensional vectors \( \theta \) and \( a \).
3. Termination Methods in ACT

3.1 Termination in Unidimensional ACT

As ACT is a variable-length computerized testing procedure, termination criterion selection is the core of the ACT design. Hence, termination criteria are discussed first. Termination criteria for IRT-based ACT fall into three general categories: (1) sequential decision theory; (2) confidence interval (CI) decision rules; and (3) Bayesian decision approaches. The sequential decision theory algorithms are generally based on Wald’s (1945; 1947) sequential probability ratio test (SPRT), which uses a likelihood ratio test statistic to determine when enough independent and identically distributed data has been collected to choose between one of two simple hypotheses. A very popular group of the existing ACT termination criteria are modifications of the SPRT, so they are reviewed first.

When the CI is used to stop an ACT, it is constructed using the standard error of estimate. If all the cutoff scores are outside the CI, the examinee can be classified (Weiss & Kingsbury, 1979; Nydick, 2013; van Groen, 2014). The Bayesian decision approaches determine, after each step, the posterior expected loss given prior information, classification proportions, and a set of responses. A test is then terminated if the expected loss for making a specific classification is the smallest (Lewis & Sheehan, 1990; Rudner, 2009). As Bayesian decision rules have been used only to terminate mastery testing and they are difficult to be generalized to the multi-classification scenario, the specific methods are not reviewed here.
3.1.1 Sequential Probability Ratio Test (SPRT)

SPRT as a statistical test was originally developed by Wald (1947) to perform classification based on a fixed cutoff point. SPRT was later referred to as a sequential classification test by Ferguson (1969). In this sequential test, the items are selected randomly or in a fixed order. For each examinee, the test is terminated when enough items are administered to ensure the SPRT-based classification is accurate. Reckase (1983) later introduced SPRT to adaptive testing as a termination criterion, and since then, SPRT and its derivatives become the state-of-art termination criteria in ACT. This is mainly due to its statistical efficiency and well-controlled decision errors. In the rest of this thesis, SPRT is referred to as a termination criterion.

SPRT was first used in mastery ACT (Eggen, 1999; Spray et al., 1997), in which there are two categories: master and non-master. Let $\theta_c$ be the cutoff point that separates masters from non-masters, then the hypotheses are specified as

$H_0: \theta \leq \theta_c - \delta$;

$H_1: \theta \geq \theta_c + \delta.$

where $\delta > 0$ is the half width of the indifference region, within which classification cannot be made. The test statistic $LR$ is

$$LR = \frac{L(\theta_c + \delta|x_S)}{L(\theta_c - \delta|x_S)}$$

(9)

where $L(\theta|x_S)$ is the likelihood examinee with ability $\theta$ has response pattern $x_S$. $S$ is the current test length. The test statistic $LR$ is then compared to two thresholds A and B to
make classification decisions. When $LR > B$, the test is stopped and the examinee is classified as a master; when $LR < A$, the test is stopped and the examinee is classified as a non-master; when $A \leq LR \leq B$, no confident classification decision can be made so another item is to be administered.

A and B, the two classification thresholds, are determined using type I and type II error rates $\alpha$ and $\beta$. Let $R_1 = \{x: LR > B\}$ and $R_0 = \{x: LR < A\}$. Then,

$$1 - \beta = \int_{R_1} L(\theta + \delta|x_S)dx = \int_{R_1} \frac{L(\theta + \delta|x_S)}{L(\theta - \delta|x_S)}L(\theta - \delta|x_S)dx$$

$$= \int_{R_1} LR \times L(\theta - \delta|x_S)dx \geq A \int_{R_1} L(\theta - \delta|x_S)dx = A \times \alpha \tag{10}$$

and

$$\alpha = 1 - \int_{R_1} L(\theta - \delta|x_S)dx = \int_{R_0} L(\theta - \delta|x_S)dx = \int_{R_0} \frac{L(\theta - \delta|x_S)}{L(\theta + \delta|x_S)}L(\theta + \delta|x_S)dx$$

$$= \int_{R_0} \frac{1}{LR} \times L(\theta - \delta|x_S)dx \geq \frac{1}{B} \int_{R_0} L(\theta + \delta|x_S)dx = \frac{1}{B} \times \beta \tag{11}$$

In this way, $A \leq \frac{\beta}{1-\alpha}$ and $\geq \frac{1-\beta}{\alpha}$. To err on the side of conservatism, $A = \frac{\beta}{1-\alpha}$ and $B = \frac{1-\beta}{\alpha}$. Strictly speaking, the two thresholds $\frac{\beta}{1-\alpha}$ and $\frac{1-\beta}{\alpha}$ do not guarantee both of the desired error rates being achieved exactly. However, they do ensure that if $\alpha'$ and $\beta'$ are the true type I and type II error rates, $\alpha' + \beta' \leq \alpha + \beta$ (Wald, 1947). Sometimes, the log likelihood ratio instead of likelihood ratio is used to facilitate computation. Thus, the test statistic becomes

$$\log LR = \log L(\theta + \delta|x_S) - \log L(\theta - \delta|x_S) \tag{12}$$
and the two thresholds are converted to $C_l = \log \left( \frac{\beta}{1-\alpha} \right)$ and $C_u = \log \left( \frac{1-\beta}{\alpha} \right)$.

### 3.1.2 Truncated SPRT (TSPRT)

Similar to the variable-length measurement CAT, ACT also needs a maximum test length $j_{\text{max}}$ to avoid extremely long tests. The SPRT with a maximum test length is called truncated SPRT (TSPRT). If ACT has not ended by SPRT after administering $j_{\text{max}}$ items, it is stopped, and a classification is made based on the relative location of the ability point estimate $\hat{\theta}$ and the cutoff point $\theta_c$. Alternatively, Bartroff, Finkelman, and Lai (2008) compared $\log \frac{L(\theta_c+\delta|x_{j_{\text{max}}})}{L(\theta_c-\delta|x_{j_{\text{max}}})}$ to 0, where $x_{j_{\text{max}}}$ is the response vector after finishing $j_{\text{max}}$ items. If the log likelihood ratio is larger than 0, then classify this examinee as a master, otherwise a non-master. In addition, $\left[ \log \frac{\beta(1-\beta)}{\alpha(1-\alpha)} \right] / 2$ has also been used to arrive at a classification (Finkelman, 2003, 2008). When $\alpha = \beta$, $\left[ \log \frac{\beta(1-\beta)}{\alpha(1-\alpha)} \right] / 2 = 0$. In this way, it is same as Bartroff et al. (2008) criterion. These classification rules would come up with slightly different classification results unless the likelihood is symmetric about $\theta_c$ and $\alpha = \beta$. There is no final conclusion on which rule is better. This divergence shows that the examinee is very difficult to be classified. It is either that the examinee locates in the indifference region or the item bank cannot provide enough good items.
3.1.3 SPRT in Multi-classification ACT

In addition to mastery ACT, SPRT is also used to conduct multi-classification. Since a single hypothesis test only compares two hypotheses to each other, multi-classification requires a series of tests. Therefore, an overall test decision is needed. Two approaches were developed to generalize SPRT in multi-classification by comparing between either the adjacent categories or all the possible pairs of categories.

Sobel and Wald’s (1949) method compares between the adjacent classification categories (Eggen, 1999, 2009; Eggen & Straetmans, 2000; van Groen et al., 2014). If there are $C$ cutoff points and $C + 1$ categories, $C$ SPRT tests are required.

Two hypotheses are formulated for each cutoff point $\theta_c$, $1 \leq c \leq C$:

$$H_{c0}: \theta \leq \theta_c - \delta;$$

$$H_{c1}: \theta \geq \theta_c + \delta.$$

where $\delta$ remains to be the half width of the indifference region. In each hypothesis test, the test statistic $LR_c = \frac{L(\theta_c + \delta | \mathbf{x}_S)}{L(\theta_c - \delta | \mathbf{x}_S)}$ is still compared to $\frac{\beta}{1 - \alpha}$ and $\frac{1 - \beta}{\alpha}$ to decide among accepting $H_{c0}$, accepting $H_{c1}$, and continuing testing. Combining all $C$ tests, if $H_{10}$ is accepted, classify the examinee to the lowest category; if $H_{c1}$ is accepted, classify the examinee to the highest (i.e., $C + 1^{th}$) category; or for $1 \leq c < C$, if $H_{c1}$ and $H_{(c+1)0}$ are both accepted, classify the examinee in the category between cutoff points $\theta_c$ and $\theta_{c+1}$; otherwise, administer another item. If the maximum test length is reached before SPRT makes a classification, the examinee is classified based on the relative location of the
point estimate $\hat{\theta}$ and cutoff points $\theta_1, \theta_2, \ldots, \theta_C$. This method was originally designed for the three-category case (Sobel & Wald, 1949), however, it was also used in the case of more than three categories (van Groen et al., 2014). When more than three categories are considered, the Sobel-Wald approach may not be able to lead to a clear decision (Ghosh & Ghosh, 1970).

To avoid the “unclear decision” outcome, Armitage’s (1950) method compares between all the possible pairs of categories; through this method/approach the potential for an unclear decision is eliminated. The tradeoff is that Armitage’s method requires more tests: $C(C + 1)/2$ instead of $C$ tests need to be performed (Armitage, 1950; Spray, 1993; Seitz & Frey, 2013). For each pair of $m < n \in \{1, \ldots, C + 1\}$, the two hypotheses of one examinee belonging to category $m$ or $n$ are

$$H_m: \theta \leq \theta_m - \delta;$$

$$H_n: \theta \geq \theta_{n-1} + \delta.$$  

The corresponding testing statistic is

$$LR_{m,n} = \frac{L(\theta_{n-1} + \delta|x_s)}{L(\theta_m - \delta|x_s)}$$  (13)

If all $C$ tests containing hypothesis $H_m$ accept it, the overall SPRT accepts hypothesis $H_m$ (Ghosh & Ghosh, 1970). ACT continues until either a consensus or the maximum test length is reached. The performance of the two methods is comparable (Govindarajulu, 1987), while Armitage’s method requires a slightly longer test than Sobel and Wald’s method (Ghosh & Sen, 1991).
Neither the $C$ tests in Sobel and Wald’s method nor the $C(C + 1)/2$ tests in Armitage’s method are independent. Thus, there comes the multiple comparison problem. Although much statistics research explored this issue thoroughly (Ghosh & Sen, 1991), no psychometric paper has adopted the adjustment. A possible reason is that the changes of $\alpha$ and $\beta$ has little effect on the classification accuracy (Eggen, 1999).

3.1.4 Curtailed SPRT (CSPRT)

In addition to multi-classification generalization, researchers also sought to improve the efficiency of SPRT. If it is impossible for further testing to change the classification decision, ceasing the test immediately is optimal. This variant of the SPRT is referred to as Curtailed SPRT (CSPRT; Gordon Lan, Simon, & Halperin, 1982). It stops the test either when a decision can be made, or when the classification will not change with the maximum test length.

In mastery ACT, assume an examinee has finished $S$ items and the maximum test length is $j_{max}$. The two hypotheses are:

$$H_0: \theta \leq \theta_c - \delta;$$

$$H_1: \theta \geq \theta_c + \delta.$$ 

the SPRT test statistic is

$$LR = \frac{L(\theta_c + \delta|x_S)}{L(\theta_c - \delta|x_S)}$$

(14)

When $S < j_{max}$:
Stop testing and accept \( H_0 \), if \( \log LR \leq \log \left( \frac{β}{1-α} \right) \) or if \( \log LR + \sum_{j=S+1}^{j_{max}} \log \frac{P_j(θ_c+δ)}{P_j(θ_c-δ)} < 0; \)

Stop testing and accept \( H_1 \), if \( \log LR \geq \log \left( \frac{1-β}{α} \right) \) or if \( \log LR + \sum_{j=S+1}^{j_{max}} \log \frac{P_j(θ_c+δ)}{P_j(θ_c-δ)} > 0; \)

Continue testing otherwise.

When \( S = j_{max} \):

Stop testing and accept \( H_0 \), if \( \log LR < 0; \)

Stop testing and accept \( H_1 \), if \( \log LR > 0. \)

In addition to the decision rules used in SPRT (\( \log LR \leq \log \left( \frac{β}{1-α} \right) \) and \( \log LR \geq \log \left( \frac{1-β}{α} \right) \)), CSPRT has two secondary decision rules: \( \log LR + \sum_{j=S+1}^{j_{max}} \log \frac{P_j(θ_c+δ)}{P_j(θ_c-δ)} < 0 \) stands for classifying the examinee as non-master at the maximum test length based on the predicted responses of item \( S + 1 \) to \( j_{max} \), while \( \log LR + \sum_{j=S+1}^{j_{max}} \log \frac{P_j(θ_c+δ)}{P_j(θ_c-δ)} > 0 \) stands for classifying the examinee as master at the maximum test length based on the predicted responses of item \( S + 1 \) to \( j_{max} \). Due to the secondary decision rules, CSPRT always results in equal or shorter test length than SPRT, thus is more efficient.
3.1.5 Stochastic Curtailed SPRT (SCSPRT)

To make CSPRT even more aggressive and effective, it can be extended to the case where a change in decision is possible but unlikely. Then it becomes stochastic curtailed SPRT (SCSPRT) (Finkelman, 2003, 2008; Wouda & Eggen, 2009).

“Stochastic” emphasizes the probabilistic nature of this method. Two extra error rates, $\epsilon_1$ and $\epsilon_2$ are used to accommodate the stochastic curtailment.

In a mastery ACT, the maximum test length is still $j_{max}$. The two hypotheses are:

$$H_0: \theta \leq \theta_c - \delta;$$

$$H_1: \theta \geq \theta_c + \delta.$$

After administered $S$ items, the SPRT statistic is

$$LR = \frac{L(\theta_c + \delta|x_S)}{L(\theta_c - \delta|x_S)} \quad (15)$$

If $S < j_{max}$:

Stop testing and accept $H_0$, if $\log LR \leq \log \left(\frac{\beta}{1-\alpha}\right)$ or if $\{D_S = H_0 \text{ and } P(D_{j_{max}} = H_0|x_S) \geq 1 - \epsilon_1\}$;

Stop testing and accept $H_1$, if $\log LR \geq \log \left(\frac{1-\beta}{\alpha}\right)$ or if $\{D_S = H_1 \text{ and } P(D_{j_{max}} = H_1|x_S) \geq 1 - \epsilon_2\}$;

Continue testing otherwise.

If $S = j_{max}$:
Stop testing and accept $H_0$, if $D_S = H_0$;

Stop testing and accept $H_1$, if $D_S = H_1$.

where $D_S$ is the temporary decision after $S < j_{max}$ items. Set $D_S = H_0$ if $\log LR < \left( \frac{\log \beta(1-\beta)}{\alpha(1-\alpha)} \right)/2$, and $D_S = H_1$ if $\log LR > \left( \frac{\log \beta(1-\beta)}{\alpha(1-\alpha)} \right)/2$. Hence, there are four error rates: $\alpha$, $\beta$, $\epsilon_1$, and $\epsilon_2$.

To administer the SCSPRT decision rule, the probability of switching categories by the maximum test length $j_{max}$ has to be derived. A normal approximation to the log likelihood after $j_{max}$ items conditioning on $S < j_{max}$ already administered items is as follows:

$$P_{\tilde{\theta}}(D_{j_{max}} = H_0 | x_S) = 1 - P_{\tilde{\theta}}(D_{j_{max}} = H_1 | x_S) \approx \Phi \left( \frac{\log \beta(1-\beta)}{\alpha(1-\alpha)} - E_{\tilde{\theta}}(\log LR(x_{j_{max}}) | x_S)}{\sqrt{Var_{\tilde{\theta}}(\log LR(x_{j_{max}}) | x_S)}} \right)$$

where

$$E_{\tilde{\theta}}(\log LR(x_{j_{max}}) | x_S) = \log LR(x_S) + \sum_{j=S+1}^{j_{max}} E_{\tilde{\theta}}(\log LR(x_j))$$

$$Var_{\tilde{\theta}}(\log LR(x_{j_{max}}) | x_S) = \sum_{j=S+1}^{j_{max}} Var_{\tilde{\theta}}(\log LR(x_j))$$

$\tilde{\theta}$ is the assumed ability under which the expectation and variation of the log likelihood at the maximum test length are evaluated. $\Phi(\cdot)$ is the CDF of the standard normal distribution. These probabilities can be calculated as long as the remaining $S + 1$ to $j_{max}$
items are known in advance (e.g. when the next item is selected to maximize information at the cutoff point in the mastery ACT). Finkelman (2003) tried two $\bar{\theta}$ values. For the first one, $\bar{\theta}$ adopted the value of $\hat{\theta}$. For the second one, $\bar{\theta}$ is the endpoint of an appropriate asymptotic one-sided confidence interval. When $\hat{\theta} < \theta_c$, $\bar{\theta} = \hat{\theta} + \frac{1}{\sqrt{\sum_{s=1}^{s=S} I(\hat{\theta}, w_s)}} z_{1-\xi}$; and when $\hat{\theta} > \theta_c$, $\bar{\theta} = \hat{\theta} - \frac{1}{\sqrt{\sum_{s=1}^{s=S} I(\hat{\theta}, w_s)}} z_{1-\xi}$. Here, $\xi$ is the significance level.

To prevent misclassification caused by imprecise $\hat{\theta}$ at the early stage of ACT, a minimum test length $j_{min}$ is usually assigned, which should be completed before the extra termination criteria of SCSPRT take effect. The SCSPRT was shown to significantly shorten the average test length with a minimal increment of type I and type II error rates (Finkelman, 2003, 2008).

As the secondary decision rules are based on the predicted responses, the selection of the $S + 1$ to $j_{max}$ items are critical to the CSPRT. Since cutoff point-based item selection methods are often employed in mastery ACT, the test is not fully adaptive. All the selected items are only determined by the item bank and the cutoff point. As a result, item responses do not affect “future item” selection. However, when SCSPRT is generalized into multi-classification ACT (Wouda & Eggen, 2009), it is impossible to know the series of the prospective items. One solution is to calculate the optimal descending ordering of item Fisher information after every administered item and to plug the first $j_{max} - S$ items into the SCSPRT as the “future items”. Other solutions could be to select the items with highest information around the cutoff point which is nearest to the
current $\theta$ estimate, or to select items which have the highest information at the middle of the two cutoff points (for the three-category condition). In short, the prospective items are updated after each item administration. Sobel and Wald’s (1949) method was employed to conduct multi-classification.

3.1.6 SPRT with Predictive Power (PPSPRT)

The performance of SCSPRT relies on the prospective responses, which in turn is determined by the ability estimate after $S$ items. To account for the uncertainty in $\hat{\theta}$, Finkelman (2010) used the posterior distribution of $\hat{\theta}$ to weight $P(D_{j_{\text{max}}} = H_0|x_S)$ and $P(D_{j_{\text{max}}} = H_1|x_S)$ in the secondary rules of SCSPRT. This weighted sum is called predictive power by Jennison and Turnbull (1999).

The posterior distribution of $\theta$ is

$$P(\theta|x_S) = \frac{\pi(\theta)L(\theta|x_S)}{\int \pi(\theta)L(\theta|x_S)d\theta}$$

(19)

where $\pi(\theta)$ is the prior distribution of $\theta$ and $L(\theta|x_S)$ is the likelihood given response pattern $x_S$. Then the predictive power can be defined as

$$P_\theta(D_{j_{\text{max}}} = H_0|x_S) = \int \theta P(\theta|x_S)P(D_{j_{\text{max}}} = H_0|x_S)d\theta$$

(20)

and

$$P_\theta(D_{j_{\text{max}}} = H_1|x_S) = \int \theta P(\theta|x_S)P(D_{j_{\text{max}}} = H_1|x_S)d\theta$$

(21)
After substituting \( P(D_{j_{max}} = H_0|\mathbf{x}_S) \) and \( P(D_{j_{max}} = H_1|\mathbf{x}_S) \) in the secondary rules of SCSPRT with \( P_\theta(D_{j_{max}} = H_0|\mathbf{x}_S) \) and \( P_\theta(D_{j_{max}} = H_1|\mathbf{x}_S) \), the classification decision can be made as in SCSPRT. PPSPRT results in even smaller type I and type II error rate inflation than the SCSPRT, which makes it by far the best termination criterion in the SPRT family (Finkelman, 2010; Nydick, 2013). The SCSPRT and PPSPRT have only been used in mastery ACT. Sobel and Wald’s method as well as Armitage’s method can be used to generalize them in the multi-classification ACT. However, as these two methods are already very complicated, their multi-classification generalizations would add to that complexity.

### 3.1.7 Generalized Likelihood Ratio (GLR)

GLR is a modification of SPRT, aiming at reducing the arbitrariness of choosing the indifference region and increasing the power of the test (George & Berger, 1990; Thompson, 2009b, 2010; Thompson & Ro, 2007). Same as SPRT, the likelihood ratio in GLR is compared to the two decision criteria \( \frac{\beta}{1-\alpha} \) and \( \frac{1-\beta}{\alpha} \) to make classification decisions. GLR tries to eliminate the subjectivity in \( \delta \) (half width of the indifference region) selection through using more meaningful values in the likelihood ratio. If the ability estimate \( \hat{\theta} \) is outside of the indifference region, it can replace the end point on the same side in the likelihood ratio computation, so the achieved GLR is the most powerful test (George and Berger, 1990; Huang, 2004). Thompson (2009b) proposed to use \( \hat{\theta}_{MLE} \) (the MLE estimate of \( \theta \)) as the ability estimate. The two likelihoods are adjusted to be
\[ L(\theta_1|x_S) = \max_{\theta > \theta_c + \delta} L(\theta|x_S) \]  

(32)

and

\[ L(\theta_2|x_S) = \max_{\theta < \theta_c - \delta} L(\theta|x_S) \]  

(33)

When \( \hat{\theta}_{MLE} \) is outside of the indifference region, based on the unimodal likelihood distribution (Brown & Croudace, 2014), one of \( \theta_1 \) and \( \theta_2 \) is the \( \hat{\theta}_{MLE} \) and the other is the end point of the indifference region on the opposite side of the cutoff point. When \( \hat{\theta}_{MLE} \) is inside of the indifference region, the GLR becomes SPRT.

An even more aggressive version of GLR would be to always use \( \hat{\theta} \) irrespective of its location and the end points of the indifference region on the opposite side of the cutoff point. A pilot study showed that this version of GLR has a very similar performance as the one Thompson (2009b) used.

An additional advantage of GLR is its simplicity in multi-classification. GLR always uses \( \hat{\theta} \) to construct the likelihood ratio, so in its generalization to multi-classification ACT, only the \( \hat{\theta} \)-involved classifications are considered. Hence, the increment of \( C \) barely affects the complexity of GLR. If \( \hat{\theta} \) falls in the extreme categories (such as \( \hat{\theta}_1 \)), only one classification (around \( \theta_1 \)) is considered; otherwise (such as \( \hat{\theta}_2 \)), two classifications (around \( \theta_1 \) and \( \theta_2 \)) are included.
On the other hand, when using multi-classification SPRT, a $C + 1$-category classification requires $C$ (using Sobel and Wald’s (1949) method) or $C(C + 1)/2$ (using Armitage’s (1950) method) SPRT tests. Thus, the complexity of classification highly relies on the magnitude of $C$.

### 3.1.8 Confidence Interval (CI)

Kingsbury and Weiss (1979) proposed use of the confidence interval of the ability estimate to make a classification decision in mastery ACT. This confidence interval is constructed around the current ability estimate using the standard error of the estimate. If the cutoff point is outside of the confidence interval, a confident classification can be made. Otherwise, another item is to be administered. The classification decision is made as follows:

- Classify examinee as below $\theta_c$, if $\hat{\theta} + \gamma \cdot se(\hat{\theta}) < \theta_c$;
- Classify examinee as above $\theta_c$, if $\hat{\theta} - \gamma \cdot se(\hat{\theta}) > \theta_c$;
- Administer another item, if $\hat{\theta} - \gamma \cdot se(\hat{\theta}) < \theta_c < \hat{\theta} + \gamma \cdot se(\hat{\theta})$. 
where $\theta_c$ is the cutoff point, $\hat{\theta}$ is the current ability estimate, $se(\hat{\theta})$ is the standard error of estimate computed using the Bayesian posterior standard deviation, and $\gamma$ is the quantile corresponding to the confidence level. If the confidence interval does not include the cutoff point, the examinee is classified into the category in which the ability estimate lies. Although this method was designed for mastery ACT, it can be used in multi-classification ACT as well. In the multi-classification condition, the examinee can be classified into the category in which the ability estimate lies, if the confidence interval does not include any cutoff points. This multi-classification generalization hardly increases the complexity of the CI approach. The ability estimate is compared to one cutoff point if $\hat{\theta}$ locates in the extreme categories and compared to two cutoff points, otherwise. This merit is similar to that of GLR which also keeps simplicity in the multi-classification generalization.

### 3.2 Termination in Multidimensional ACT with a Composite Score

SPRT, all the above-mentioned SPRT-based methods, GLR, and CI can be generalized to multidimensional ACT. All existing termination criteria in multidimensional ACT use a composite score to transform multidimensional traits to a unidimensional score such that the as-usual unidimensional termination criteria can proceed. Three methods are used to construct the composite score in need. They are the constrained method and the projected method by Nydick (2013), and the reference
composite method by van Groen (2014) and van Groen et al. (2016). Moreover, van Groen’s (2014) classification on the entire test using between-item multidimensional items can be considered as a special case of classification based on a composite score.

3.2.1 Constrained SPRT (C-SPRT) and Projected SPRT (P-SPRT)

Nydick (2013) referred to the cutoff points on the composite score as the classification bound, which can be either compensatory or non-compensatory. For instance, a compensatory bound function for a two-dimensional test can be

\[ g(\theta) = 1.5\theta_1 + \theta_2 - .5 \]  (22)

whereas a non-compensatory bound function can be

\[
g(\theta) = \begin{cases} 
\theta_1 - 2 & \text{if } \theta_2 \geq 1, \\
\theta_2 - 1 & \text{if } \theta_1 \geq 2, \\
1 & \text{otherwise.}
\end{cases}
\]  (23)

Based on the bound function, there are two ways to compute the composite score. The constrained maximum likelihood estimate is computed as follows:

\[
\hat{\theta}_c = \arg \max_{\theta \in \Theta_0} \log L(\theta | x_S)
\]  (24)

where \( \Theta_0 = \{ \theta : g(\theta) = 0 \} \) constrains this estimate to lie on the bound.

In contrast to the constrained method, the projected method projects the unconstrained MLE orthogonally onto the classification bound surface to obtain the cutoff composite score. The same bound function \( g(\theta) \) is used. The projected maximum likelihood estimate would be
\[ \hat{\theta}_c \equiv \arg\min_{\theta \in \Theta_0} \| \hat{\theta}_S - \theta \| \]  

(25)

where \( \| \cdot \| \) is the Euclidean distance, \( \hat{\theta}_S \) is the unconstrained maximum likelihood estimate after administering \( S \) items.

For both the constrained method and the projected method, the SPRT hypotheses are

\[ H_0: \theta \leq \hat{\theta}_c - \delta \theta_\delta; \]

\[ H_1: \theta \geq \hat{\theta}_c + \delta \theta_\delta. \]

where \( \theta_\delta \) is a unit-length vector along the line perpendicular to the bound function at \( \hat{\theta}_c \).

In this way, the C-SPRT and the P-SPRT only differ in the \( \hat{\theta}_c \) computation.

The SPRT statistic is computed as

\[ LR = \frac{L(\hat{\theta}_c + \delta \theta_\delta | x_S)}{L(\hat{\theta}_c - \delta \theta_\delta | x_S)} \]  

(26)

This \( LR \) statistic is still compared to \( \frac{\beta}{1-\alpha} \) and \( \frac{1-\beta}{\alpha} \) as in the unidimensional SPRT to determine whether one examinee is a master, a non-master, or another item should be administered.

In addition to the SPRT, the constrained method and the projected method were also used to generalize the SCSPRT and PPSPRT to multidimensional ACT, thus increasing test efficiency (Nydick, 2013). The only difference between unidimensional SPRTs and this multidimensional version is the way the indifference region is constructed. The decision rule and benchmarks are the same.
3.2.2 SPRT with Reference Composite

Reference composite (RC) (van Groen, 2014; van Groen et al., 2016) is another way to derive the composite score. Different from the composite score defined based on an arbitrarily selected bound function in the constrained method and projected method, RC is determined by the item bank. Let \( \mathbf{a} \) be the item discrimination parameter matrix of the item bank, then RC has the same direction as the largest eigenvalue-related eigenvector of the \( \mathbf{a} \mathbf{a}' \) matrix (Reckase, 2009), which is the direction that the item bank provides the most information. The \( k^{th} \) element of this eigenvector is the direction cosine \( \alpha_{\xi_k} \) of the angle between the RC and the dimension axis \( k \). The Euclidean norm (length) of the estimated ability vector \( \hat{\mathbf{\theta}} \) is

\[
EN = \sqrt{\sum_{k=1}^{K} \hat{\theta}_k^2} \tag{27}
\]

The direction cosine between axis \( k \) and the length of ability is \( \cos \alpha_k = \frac{\hat{\theta}_k}{EN} \). The estimated RC is \( \hat{\xi} = EN \times \cos \alpha_{\xi} \), where \( \alpha_{\xi} = \alpha_k - \alpha_{\xi_k} \). In this way,

\[
\hat{\xi} = \sqrt{\sum_{k=1}^{K} (\hat{\theta}_k \times \cos \alpha_{\xi})} \tag{28}
\]

That is, the weights in the RC are the same for different dimensions within one examinee. The cutoff point \( \xi_c \) and the corresponding half indifference region width \( \delta_{\xi} \) are both on
the $\xi$ scale. The SPRT is conducted based on the end points transformed back to the $\theta$ scale. $\theta_{\xi c\pm\delta} = \cos \alpha_\xi \times (\xi_c \pm \delta_\xi)$ are the two end points, where $\alpha_\xi = (\alpha_{\xi 1}, \ldots, \alpha_{\xi K})$ are all the angles between RC and the dimension axis. The SPRT statistic becomes

$$LR = \frac{L(\theta_{\xi c+\delta} | \mathbf{x}_S)}{L(\theta_{\xi c-\delta} | \mathbf{x}_S)}$$ (29)

The RC method only works in within-item multidimensional tests. The reason is that the diagonal $\mathbf{a} \mathbf{a}'$ matrix for between-item multidimensional test leads the RC to lie on merely the most discriminating dimension. It is obviously not optimal to classify only based on one dimension in multidimensional ACT. In contrast, although Nydick’s (2013) methods (C-SPRT and P-SPRT) were also developed for within-item multidimensional ACT, both can be applied to between-item multidimensional ACT. Although the RC is determined objectively by the item bank characteristics, its usage in practice may still be limited. This is because the classification criteria should be formulated based on the objective of the test rather than exclusively based on the capacity of the item bank.

### 3.2.3 Between-item Multidimensional SPRT for an Entire Test

Seitz and Frey (2013) proposed using SPRT to conduct classification along each dimension separately in the between-item multidimensional mastery ACT. The null and alternative hypotheses for dimension $k$ are

$$H_{k0}: \theta_k \leq \theta_{ck} - \delta;$$

$$H_{k1}: \theta_k \geq \theta_{ck} + \delta.$$
The corresponding test statistic is:

\[
LR_k = \frac{L(\hat{\theta}_k + \delta|\mathbf{x}_S)}{L(\hat{\theta}_k - \delta|\mathbf{x}_S)}
\]  

(30)

where \(\hat{\theta}_k\) is the cutoff point used to make classification along dimension \(k\). This method and the unidimensional SPRT differ only in the choosing of the cutoff point. The \(k^{th}\) element of the cutoff point \(\hat{\theta}_k\) is the cutoff score along dimension \(k\), while the other elements of \(\hat{\theta}_k\) adopt the current \(\theta\) estimates on the other dimensions. \(\hat{\theta}_{kj} = \hat{\theta}_j, \forall j \neq k;\) and \(\hat{\theta}_{kk} = \theta_{ck}\).

van Groen (2014) extended Seitz and Frey’s (2013) study to make classification decisions on the entire test. As all items have between-item multidimensionality, the likelihood of the entire test is the product of the likelihood on each dimension. The SPRT statistic becomes

\[
LR = \prod_{k=1}^{K} \frac{L(\theta_{ck} + \delta; \mathbf{x}_k)}{L(\theta_{ck} - \delta; \mathbf{x}_k)}
\]  

(31)

where \(\theta_{ck}\) is the cutoff point along dimension \(k\), and \(\mathbf{x}_k\) is the response vector of all items measuring dimension \(k\). \(LR\) is still compared to the two benchmarks \(\frac{\beta}{1-\alpha}\) and \(\frac{1-\beta}{\alpha}\).

Superficially, this method classifies examinees according to each dimension. However, it classifies examinees into only two groups: master all traits and fail all traits. It is essentially unidimensional classification. This is indeed a special case of P-SPRT with equal linear weights in the bound function. This method cannot be generalized to within-
item multidimensional tests, because the likelihood on different dimensions are no longer independent.

### 3.2.4 Multidimensional GLR (MGLR)

Similar to SPRT, the GLR was also generalized to multidimensional ACT. Nydick (2013) used the bound function \( g(\theta) \) to separate between the master region and the non-master region. Let \( \Theta_m \) be the set of points in the master region and \( \Theta_n \) be the set of points in the non-master region. The composite hypotheses are:

\[
H_0: \theta \in \Theta_n; \\
H_1: \theta \in \Theta_m.
\]

The test statistic is constructed as follows:

\[
GLR = \frac{\max_{\theta_1 \in \Theta_m} \left( L(\theta_1 | x_S) \right)}{\max_{\theta_2 \in \Theta_n} \left( L(\theta_2 | x_S) \right)}
\] (32)

The two maximums in the numerator and denominator are easily found using a constrained optimization routine. As the unimodal likelihood distribution holds in the multidimensional case, one of the maximums is found at the \( \hat{\theta}_{MLE} \) while the other is the constrained ability estimate \( \hat{\theta}_c \). Nydick (2013) also pointed out that, in addition to the maximum found on the bound function, the maximum found on the end points of the indifference region can also be used. It is more optimal to use the end point of the
indifference region than the maximum found on the bound function, as it ensures enough distance between $\theta_1$ and $\theta_2$, thus is easier to arrive at a likelihood ratio different from 1.

### 3.2.5 Weighted GLR (WGLR)

Weights can be added to MGLR to account for the prior information of $\theta$. Then the two composite hypotheses are:

$$H_0: \theta \in \Theta_n;$$

$$H_1: \theta \in \Theta_m.$$  

Accordingly, the WGLR is defined as

$$WGLR = \frac{\int_{\Theta_m} w L(\theta|x_S) d\theta}{\mu_{H_1}(w)} \div \frac{\int_{\Theta_n} w L(\theta|x_S) d\theta}{\mu_{H_0}(w)}$$

(33)

where $w$ is the weight which is usually set to the prior distribution of $\theta$. $\mu_{H_0}(w) = \int_{\Theta_n} w d\theta$ and $\mu_{H_1}(w) = \int_{\Theta_m} w d\theta$. This $WGLR$ is usually compared to a pair of pre-specified criteria $T$ and $1/T$ (Jha et al., 2013). Given comparable thresholds, WGLR needs fewer items than a corresponding MGLR.

By the virtue of prior information, both PPSPRT and WGLR perform better than their non-Bayesian opponents SCSPRT and MGLR. However, especially in the multidimensional case, the integration over a prior distribution greatly increases the computational burden. Moreover, the influence of the prior is particularly strong with
short tests. In a variable-length adaptive testing like ACT, short tests are expected to occur. As a result, using these Bayesian methods may bring bias to the classification.

### 3.2.6 Wita hin-item Multidimensional Confidence Interval

If examinees are classified based on a linear composite score, \( \zeta = \lambda' \theta \) or \( \zeta = \sum_{l=1}^{L} \theta_l \lambda_l \) (van der Linden, 1999), the CI is used based on the composite score \( \zeta \). In this way, \( se(\zeta) = \sqrt{\lambda'V(\theta)\lambda} \).

Classify examinee as below \( \zeta_c \), if \( \hat{\zeta} + \gamma \cdot se(\hat{\zeta}) < \zeta_c \);

Classify examinee as above \( \zeta_c \), if \( \hat{\zeta} - \gamma \cdot se(\hat{\zeta}) > \zeta_c \);

Administer another item, if \( \hat{\zeta} - \gamma \cdot se(\hat{\zeta}) < \zeta_c < \hat{\zeta} + \gamma \cdot se(\hat{\zeta}) \).

where \( \zeta_c \) is the cutoff point on the composite score \( \zeta \) scale. The RC van Groen (2014) used is a linear combination, thus van Groen used this method to perform classification for the within-item multidimensional ACT. As linear combination is one of many ways to construct a composite score in multidimensional IRT, this method is also a special case of Nydick’s (2013) projected method. In spite of the fact that van Groen (2014) utilized this method in the within-item multidimensional ACT, it can be used for both within-item and between-item multidimensional tests.
3.3 Termination in Multidimensional ACT with Grid Classification

In order to conduct grid classification, that is to classify examinees into grids encircled by cutoff scores (lines/surfaces) on different dimensions, two new termination criteria are developed in this thesis based on GLR. Moreover, the SPRT (Seitz & Frey, 2013) and CI (van Groen, 2014) criteria are also malleable to grid classification, thus they are adapted and described in this section.

3.3.1 Grid Classification GLR (GGLR)

As mentioned above, \( C \) or \( C(C+1)/2 \) comparisons are used for multi-classification SPRT, whereas GLR only requires one or two tests based on the one or two adjacent cutoff points around the ability estimate. As the number of cutoff points increases with the number of dimensions, this frugal merit of GLR is especially beneficial when grid classification is applied in multidimensional ACT.

In grid multi-classification ACT, regardless of the actual number of cutoff points along each dimension, at most \( 2^K \) (the number of endpoints a hypercube has) cutoff points around \( \hat{\theta} \) are relevant. Figure 2 shows an example with \( K = 2 \) and \( C = 3 \).

When \( \hat{\theta}_1 \) is the ability estimate, all four cutoff points are relevant. However, when \( \hat{\theta}_2 \) is the ability estimate, only two cutoff points \( \theta_{12} \) and \( \theta_{22} \) are relevant. Whereas when \( \hat{\theta}_3 \) is the ability estimate, only one cutoff point \( \theta_{22} \) is relevant. As \( K \) tests are performed around each cutoff point, at most \( 2^K \times K \) pairs of hypotheses around \( \hat{\theta} \) need to be tested.
when GGLR is used. Here, \( \hat{\theta} \) and the end point of the indifference region on the opposite side are used to construct each likelihood ratio.

\[ \begin{align*}
\theta_{11} & \quad \theta_{12} \\
\theta_{21} & \quad \theta_{22} \\
\bar{\theta}_1 & \quad \bar{\theta}_2 \\
\bar{\theta}_3
\end{align*} \]

\textit{Figure 2. Relevant cutoff points}

To be specific, if \( \theta_{c_1 \ldots c_K} \) is one of the relevant cutoff points, two hypotheses are formulated for the test along the \( k \)th dimension (\( \theta_{c_k} \) is the \( k \)th component of \( \theta_{c_1 \ldots c_K} \)):

\[ H_{k,c_k,0}: \theta_k \leq \theta_{c_k} - \delta; \]

\[ H_{k,c_k,1}: \theta_k \geq \theta_{c_k} + \delta. \]

The test statistic is

\[ GGLR_{k,c_1 \ldots c_K} = \frac{L(\hat{\theta} | x_S)}{L(\bar{\theta}_{c_1 \ldots c_K} | x_S)} \] (34)
where $\tilde{\theta}_{c_1\cdots c_K} = \theta_{c_1\cdots c_K} + \delta$. The $k^{th}$ element of $\delta$ is $\delta \times \text{sgn}(\theta_{c_k} - \hat{\theta}_k)$ and all other elements of $\delta$ are 0. Here the sign function is defined as $\text{sgn}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0. \end{cases}$

Figure 3 illustrates the end point of indifference region on the opposite side, $\tilde{\theta}_{c_1\cdots c_K}$ when the test focus is on dimension 1. As $GGLR_{k,c_1\cdots c_K}$ is always larger than 1, it is compared to $\frac{1-\beta}{\alpha}$ to make classification decisions. The test stops only when all GGLR tests result in classification or ACT reaches the maximum test length. If the maximum test length is reached, classify the examinee based on the relative location of $\hat{\theta}$ and the cutoff points.

To take advantage of the correlation between dimensions and to facilitate classification, the maximum a posteriori (MAP) estimate is used as $\hat{\theta}$.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure3.png}
\caption{GGLR for grid classification}
\end{figure}
3.3.2 Simplified Grid Classification GLR (SGLR)

As the likelihood at $\hat{\theta}$ is always larger than that at the opposite end point of the indifference region, the GLR essentially determines whether the examinee can be classified into the grid in which $\hat{\theta}$ lies. Thus, instead of conducting $K$ GLR tests at each cutoff point, one GLR comparing between $\hat{\theta}$ and the end point in the diagonally adjacent category is enough. In Figure 4, the orange dot lies in the diagonal adjacent category of $\hat{\theta}$.

![Simplified Grid Classification SGLR](image)

Figure 4. SGLR for grid classification

Two hypotheses are formulated for each cutoff point $\theta_{c_1\ldots c_K}$:

$H_{c_0}: \theta \in G_{\hat{\theta}}$;

$H_{c_1}: \theta \in G_{da,c_1\ldots c_K}$.  

39
where \( G_\hat{\theta} \) is the grid \( \hat{\theta} \) lies in, and \( G_{da,c_1...c_K} \) is the grid diagonally adjacent to \( G_\hat{\theta} \) through cutoff point \( \theta_{c_1...c_K} \). By diagonal adjacency, we define \( sgn(\hat{\theta}_k - \theta_{c_k}) = -sgn(\theta_k - \theta_{c_k}), \forall \theta \in G_{da,c_k} \) and \( \forall k \in 1, \ldots, K \).

The test statistic is

\[
SGLR_{c_1...c_K} = \frac{L(\hat{\theta} | x_S)}{L(\tilde{\theta}_{c_1...c_K} | x_S)}
\]

where \( \tilde{\theta}_{c_1...c_K} = \theta_{c_1...c_K} + \delta \times sgn(\theta_{c_1...c_K} - \hat{\theta}) \) is the end point in the diagonally adjacent category.

### 3.3.3 Between-item Multidimensional CI

CI can be used to make a classification on each dimension. For each dimension the confidence interval termination method becomes

Classify examinee as below \( \theta_{c_kp} \), if \( \hat{\theta}_k + \gamma \cdot se(\hat{\theta}_k) < \theta_{c_kp} \);

Classify examinee as above \( \theta_{c_kp} \), if \( \hat{\theta}_k - \gamma \cdot se(\hat{\theta}_k) > \theta_{c_kp} \);

Administer another item, if \( \hat{\theta}_k - \gamma \cdot se(\hat{\theta}_k) < \theta_{c_kp} < \hat{\theta}_k + \gamma \cdot se(\hat{\theta}_k) \).

where \( \theta_{c_kp} \) is the \( p^{th} \) cutoff score along dimension \( k \), and \( \hat{\theta}_k \) is the current ability estimate along dimension \( k \). If CI is performed on all dimensions, it is essentially grid classification. Although van Groen (2014) developed this method to conduct
classification in the between-item multidimensional ACT, it can also be used in the within-item multidimensional ACT.

### 3.3.4 Between-item Multidimensional SPRT

Seitz and Frey (2013) proposed using SPRT to conduct classification along each dimension separately in between-item multidimensional mastery ACT. The hypotheses pair that differentiates between category $m$ and $n$ ($m < n$) along dimension $k$ is as follows:

$$H_{km}: \theta_k \leq \theta_{ckm} - \delta;$$

$$H_{kn}: \theta_k > \theta_{ck(n-1)} + \delta.$$

The corresponding test statistic is:

$$LR_{k,n-1,m} = \frac{L(\bar{\theta}_{k(n-1)} + \delta|x_S)}{L(\bar{\theta}_{km} - \delta|x_S)}$$

(36)

where $\bar{\theta}_{kp}$ is the cutoff point used to make a classification along dimension $k$. This method and the unidimensional SPRT only differ in the choosing of the cutoff point. The $k^{th}$ element of the cutoff point $\bar{\theta}_{kp}$ is the $p^{th}$ cutoff score along dimension $k$, whereas the other elements of $\bar{\theta}_{kp}$ adopt the current $\theta$ estimates on the other dimensions. The idea of classification on each dimension separately conforms to the grid classification proposed in this thesis. In this way, this method can be adopted to perform grid classification. ACT stops when all dimensions can be classified, or the maximum test
length is reached. Although this method is designed for between-item multidimensional tests, it is also applicable to within-item multidimensional tests.
4. Item Selection Methods in ACT

In this section, different item selection methods are reviewed. A large proportion of the item selection methods used in measurement CAT are also implemented in ACT. Among them are the Fisher information-based methods, Kullback-Leibler information-based methods and mutual information-based methods. Moreover, some of the item selection methods are modified to account for the classification objective of ACT and the existence of one or more cutoff points.

4.1 Item Selection Methods in Unidimensional ACT

4.1.1 Maximum Fisher Information

Fisher information is defined as the negative expected second derivative of the log-likelihood with respect to $\theta$. Maximum Fisher information is a pivotal item selection method in unidimensional measurement CAT. It, as well as some variants, are also used in ACT. In unidimensional IRT, Fisher information for a 3PL item was defined by Birnbaum (1968) as

$$I_j(\theta) = -E\left(\frac{\partial^2 \log P_j(\theta)}{\partial \theta^2}\right) = a_j^2 \frac{(1 - P_j(\theta))(P_j(\theta) - c_j)^2}{P_j(\theta)(1 - c_j)^2}$$

where $P_j(\theta)$ is the probability of correctly responding to item $j$. $a_j$ and $c_j$ are the discrimination and pseudo-guessing parameters of item $j$. 
Fisher information for a GRM item $j$ can be written as (Dodd, De Ayala, & Koch, 1995; Samejima, 1968):

$$I_j(\theta) = \sum_{r=1}^{R} \frac{[P_{jr}^*(\theta) - P_{j(r+1)}^*(\theta)]^2}{P_{j(r+1)}(\theta)}$$

where $P_{jr}(\theta)$ is the probability of obtaining score $r$ with ability $\theta$, $P_{jr}^*(\theta)$ is the probability of endorsing score $r$ or higher. $P_{jr}^*(\theta)$ is the first derivative of $P_{jr}(\theta)$ with respect to $\theta$. $P_{jr}^*(\theta) = P_{jr}(\theta) \times (1 - P_{jr}^*(\theta))$.

Same as measurement CAT, in ACT the item that provides maximum Fisher information at the current ability estimate can be selected to be administered next (Spray & Reckase, 1994; Eggen & Straetmans, 2000; Thompson, 2007; Thompson & Ro, 2007). This method fits both the mastery ACT and multi-classification ACT, using either dichotomous items or polytomous items. Item $j$ is selected as follows:

$$j \equiv \arg\max_{j \in J} I_j(\hat{\theta})$$

where $J$ is the available item bank and $\hat{\theta}$ is the current ability estimate.

Whereas ACT uses the cutoff points to classify examinees, the item that provides the maximum information at the cutoff points would benefit classification the most. Maximum Fisher information at the cutoff points aims at reaching high estimate accuracy around the cutoff points thus leading to high classification accuracy. For mastery ACT, the item is selected to have maximum Fisher information at the only cutoff point (Eggen, 1999; Lin, 2011; Lin & Spray, 2000; Spray & Reckase, 1994). In multi-classification ACT, there exist more than one cutoff points. One way is to select the item that maximize
Fisher information at the cutoff point which is nearest to current $\hat{\theta}$ (Eggen, 2009; Eggen & Straetmans, 2000; Thompson, 2007; Wouda & Eggen, 2009). Item $j$ is selected as follows:

$$j \equiv \arg\max_{j \in J} I_j(\theta_c) \quad (40)$$

where $\theta_c$ is the cutoff point in mastery ACT or the cutoff point located closest to $\hat{\theta}$ in multi-classification ACT.

When there are three categories, in addition to maximizing item Fisher information at the nearest cutoff point, Wouda and Eggen (2009) proposed selecting the item that maximizes Fisher information at the midpoint of the two cutoff points. Item $j$ is selected as follows:

$$j \equiv \arg\max_{j \in J} I_j \left( \frac{\theta_{c1} + \theta_{c2}}{2} \right) \quad (41)$$

where $\theta_{c1}$ and $\theta_{c2}$ denote the two cutoff points. When there are more than three categories, more than two cutoff points exist; then this method can be generalized to maximize Fisher information at the midpoint of the two nearest cutoff points, or at the mean of all cutoff points.

### 4.1.2 Maximum Weighted Fisher Information (WFI)

In addition to maximizing Fisher information at one point, an item can be selected to maximize a weighted sum of Fisher information at multiple points. Veerkamp and
Berger (1997) proposed selecting the item that provides the largest weighted Fisher information. The weighted Fisher information is as follows:

\[
I_{w_j}(\theta|w) = \int_{\theta} w \cdot l_j(\theta) d\theta
\]  

(42)

where \( w \) is the weight. \( \Theta \) is the set of all \( \theta \) values that are included. \( \Theta \) can either be a set of discrete points or an interval. If an identity function is used to generate the weights, WFI becomes the global programming (GP) method developed by van Groen et al. (2014). Item \( j \) is selected as

\[
j \equiv \arg \max \sum_{c=1}^{C} w_c I_j(\theta_c)
\]  

(43)

where \( I_j(\theta_c) \) is the Fisher information of item \( j \) at cutoff point \( \theta_c \). The corresponding weight \( w_c \equiv 1 \) for all cutoff points.

If only a few cutoff points are of concern (e.g. the nearest two cutoff points), these important cutoff points are assigned non-zero weights, while the rest are assigned zero weights. This is the idea of the global criterion (GC) method from van Groen et al. (2014). Item \( j \) is then selected as follows:

\[
j \equiv \arg \max \sum_{c'=1}^{C'} l_j(\theta_{c'})
\]  

(44)

where \( C' \) of the cutoff points are considered. This procedure is different from Wouda and Eggen’s (2009) method which selects the item that provides the highest Fisher information on the mean of the nearest two cutoff points (if the current ability estimate is
between two cutoff points). It is the Fisher information on one point, even though that
one point is derived from the location of several points. In contrast, the GC utilizes the
summation of Fisher information at multiple points.

van der Linden (1998) proposed selecting the item that has maximum expected
information across the posterior distribution in measurement CAT. As item selection
usually focuses on cutoff points in ACT, Weissman (2007) proposed using the posterior
distribution of $\theta$ as weights, then choosing the item that maximizes the weighted sum of
Fisher information on all cutoff points. Item $j$ is then selected as follows:

$$j \equiv \arg\max_{j \in J} \sum_{\theta_c \in \Theta_C} P(\theta_c | x_S) I_j(\theta_c)$$

(45)

where $\Theta_C$ is the set including all cutoff points and $x_S$ is the interim response vector after
administering $S$ items. This weighted method moderates between GP and maximum
Fisher information at the closest cutoff point. The posterior distribution of $\theta$
demonstrates the importance of each cutoff point. Using this weight prevents the
arbitrariness in determining critical cutoff points in GC.

### 4.1.3 Kullback-Leibler Information (KL)

KL information measures the non-symmetric discrepancy between two
probability distributions $f(z)$ and $f(y)$. It is defined as follows, where the expectation is
taken with respect to $f(y)$,
\[
KL(Z, Y) = E \left( f(z) \log \frac{f(z)}{f(y)} \right).
\] (46)

When KL information is applied in measurement CAT item selection, \( Z \) is the item response probability at \( \theta_1 : f(z) = P(x_j|\theta_1) \), whereas \( Y \) is the item response probability at \( \theta_2 : f(y) = P(x_j|\theta_2) \). In this way, KL information is the expected log likelihood ratio between two points \( \theta_1 \) and \( \theta_2 \). It reflects the extent item \( j \) can distinguish between the two points. Then the KL information for dichotomous and polytomous item \( j \) are as follows:

\[
KL_j(\theta_1|\theta_2) = P_j(\theta_1) \log \frac{P_j(\theta_1)}{P_j(\theta_2)} + \left(1 - P_j(\theta_1)\right) \log \frac{1 - P_j(\theta_1)}{1 - P_j(\theta_2)}
\] (47)

and

\[
KL_j(\theta_1|\theta_2) = \sum_{r=1}^{R} P_{jr}(\theta_1) \log \frac{P_{jr}(\theta_1)}{P_{jr}(\theta_2)}
\] (48)

where \( P_j(\theta) \) is the correct response probability of item \( j \), and \( P_{jr}(\theta) \) is the probability of choosing option \( r \) in item \( j \).

In order to utilize KL information as an item selection method, \( \theta_1 \) and \( \theta_2 \) have to be determined. Chang and Ying (1996) proposed using KL information around \( \hat{\theta} \) as the item selection index. That is \( \theta_1 = \hat{\theta} \) and \( \theta_2 = \theta \). The selected item \( j \) is obtained as

\[
j \equiv \arg\max_{j \in J} \int_{\hat{\theta} - \zeta}^{\hat{\theta} + \zeta} KL_j(\hat{\theta}|\theta) d\theta
\] (49)
where $\zeta$ is a small number that denotes the width of the integrating range around $\hat{\theta}$. In this way, the item that provides the highest KL information around the current ability estimate is selected.

When generalizing the KL information to mastery ACT, Thompson (2009a) directly adopted Chang and Ying’s (1996) method to select the item that maximized KL information at the current ability estimate. On the other hand, the item that provides the maximum KL information around the cutoff point can be selected. In addition to utilizing Chang and Ying’s (1996) method at the cutoff point, Eggen (1999) and Lin and Spray (2000) proposed using the information between the two end points of the indifference region. Thus, $\theta_1 = \theta_u$ and $\theta_2 = \theta_l$. The latter method selects the item which is best at distinguishing between the two endpoints of the indifference region, thus coinciding with SPRT which is the most popular termination criterion in ACT. The simulation study also supported Eggen (1999) and Lin and Spray’s (2000) method which outperformed Chang and Ying’s (1996) method when paired with SPRT (Thompson, 2009a).

In multi-classification ACT, there is more than one indifference region, so Eggen (1999) and Lin and Spray’s (2000) method for mastery ACT needs adjustment. When there are three categories, Eggen (1999) proposed two alternatives. The first is to select an item with the maximum KL information index around the cutoff point closest to $\hat{\theta}$. The second determines $\theta_1$ and $\theta_2$ based on which hypothesis can lead to a decision. That is, if none of the two hypothesis pairs (a pair of hypotheses around cutoff point 1 to distinguish between category 1 and 2, and a pair of hypotheses around cutoff point 2 to distinguish between category 2 and 3) can lead to a decision, the item with maximum KL
information between the two cutoff points is selected. If one pair has led to a decision and the other has not, the item with the maximum KL information around the undecided test-related cutoff point is selected. The second is more appealing theoretically as it takes the classification decision-making procedure into account. However, simulation results showed that these two variations barely made any difference. For this reason, the first version is preferable due to its simplicity.

4.1.4 Weighted Log Odds (WLO)

The weighted log odds ratio between $\theta_1$ and $\theta_2$ for a dichotomous item $j$ is

$$WLO_j(\theta_1|\theta_2) = D_{jr} \log \frac{P_{jr}(\theta_1)}{P_{jr}(\theta_2)} + (1 - D_{jr}) \log \frac{1 - P_{jr}(\theta_1)}{1 - P_{jr}(\theta_2)},$$

(50)

and for a polytomous item $j$ is

$$WLO_j(\theta_1|\theta_2) = \sum_{r=1}^{R} D_{jr} \log \frac{P_{jr}(\theta_1)}{P_{jr}(\theta_2)},$$

(51)

where $D_{jr}$ is the expected response rate of option $r$ in item $j$ over the whole ability distribution or the classical test theory (CTT) difficulty, and $P_{jr}(\theta)$ is the probability of selecting option $r$ on item $j$. $D_{jr}$ is used as the weight. In mastery ACT, as the null hypothesis is the examinee belongs to the non-master group, that is $\theta \in \Theta_n$, the best item to reject the null hypothesis would lead to the largest likelihood ratio between a point in the master region and a point in the non-master region. In this way, the two end points of the indifference region in ACT, $\theta_l$ and $\theta_u$, are usually chosen as $\theta_1$ and $\theta_2$ (Lin & Spray,
WLO is very similar to, yet different from, KL information, in which the item response probability $P_{jr}(\theta_1)$ replaces the CTT difficulty, $D_{jr}$, as the weight.

Nydick (2013) generalized WLO to multidimensional mastery ACT by simply plugging in the multidimensional ability. The cutoff points $\theta_u$ and $\theta_l$ are still obtained using the constrained method or the projected method.

4.1.5 Expected Log Likelihood Ratio (ELR)

Nydick (2013) replaced $D_{jr}$ in WLO with $P_{jr}(\hat{\theta})$ and developed the ELR. The expected log likelihood ratio for the prospective item $j$ is

$$ ELR_j = P_j(\hat{\theta}) \log \frac{P_j(\theta_u)}{P_j(\theta_l)} + (1 - P_j(\hat{\theta})) \log \frac{1 - P_j(\theta_u)}{1 - P_j(\theta_l)} $$

(52)

when it is a dichotomous item, and

$$ ELR_j = \sum_{r=1}^{R} P_{jr}(\hat{\theta}) \log \frac{P_{jr}(\theta_u)}{P_{jr}(\theta_l)} $$

(53)

when it is a polytomous item, where $P_{jr}(\hat{\theta})$ is the probability option $r$ is selected in item $j$ based on the current ability estimate $\hat{\theta}$, and $P_j(\hat{\theta})$ is the correct response probability of item $j$ based on the current ability estimate $\hat{\theta}$. Whether ELR should be maximized or minimized is determined by the location of $\hat{\theta}$. Item $j$ should be chosen to maximize $ELR_j(\hat{\theta})$ when $\hat{\theta} \in \Theta_m$, and item $j$ should be chosen to minimize $ELR_j(\hat{\theta})$ when $\hat{\theta} \in \Theta_n$. 

51
4.1.6 Mutual Information (MI)

For two continuous random variables $Z$ and $Y$, MI measures the amount of information $Z$ provides about $Y$. It is also the information $Y$ provides about $Z$.

$$MI(Z, Y) = \int_{z \in Z} \int_{y \in Y} f(z, y) \log \frac{f(z, y)}{f(z)f(y)}$$  \hspace{1cm} (54)

Distance MI can be seen as a special case of KL information, as it measures the KL between the joint distribution $f(z, y)$ and the product of the marginal distribution $f(z)$ and $f(y)$. When $Z$ and $Y$ are independent, the joint information can provide nothing more than the product of two marginal information, then $f(z, y) = f(z)f(y)$, and the mutual information equals zero. Otherwise, the joint distribution would always provide information in addition to the marginal information of the two distributions. Therefore, MI is always larger than or equal to zero.

As an item selection method, MI developed by Weissman (2007), Mulder and van der Linden (2010), and Wang (2013) is the most general. It integrates over the entire space of the joint distribution, both ability distribution and item response. In the measurement CAT item selection case, $Y$ is the current posterior distribution of $\theta$ based on $S$ items $\pi(\theta|x_S)$, and $Z$ is the predictive response distribution of item $j$ conditioning on the previous responses of $S$ items, $P(x_j|x_S)$. In addition, as the response is discrete in IRT ($x_j = 1, 2, ..., R$ for polytomous items and $x_j = 0, 1$ for dichotomous items), the
integration over the space of $x_j$ becomes a summation. Hence, the MI item selection criterion in measurement CAT is as follows:

$$MI_j = \sum_{x_j=1}^{R} \int_{\theta \in \Theta} f(\theta, x_j|x_S) \log \frac{f(\theta, x_j|x_S)}{\pi(\theta|x_S)P(x_j|x_S)} d\theta$$

(55)

where $X$ is the set of all the possible responses of item $j$, and $\Theta$ is the entire space of $\theta$. Note that $f(\theta, x_j|x_S) = P(x_j|\theta)\pi(\theta|x_S)$, so MI can be simplified as

$$MI_j = \sum_{x_j=1}^{R} \int_{\theta \in \Theta} P(x_j|\theta)\pi(\theta|x_S) \log \frac{P(x_j|\theta)}{P(x_j|x_S)} d\theta$$

(56)

The item maximizing MI between the examinee’s current posterior distribution and the response distribution on the candidate item $j$ should be selected, because this item response $x_j$ provides the most information about $\theta$.

### 4.2 Item Selection Methods in Multidimensional ACT

#### 4.2.1 Maximum Determinant of the Fisher Information Matrix (D-optimal)

In multidimensional tests, Fisher information is a matrix instead of a scalar. Various methods are used to extract a scalar that represents the Fisher information matrix and can be used in finding the optimal item (Frey & Seitz, 2009). The Fisher information matrix of a M2PL item is as follows (Wang & Chang, 2011):
\[ I_j(\theta) = a_j a_j' P_j(\theta)(1 - P_j(\theta)) \tag{57} \]

where \( a_j \) is a column vector containing the discrimination parameters on all dimensions of item \( j \).

As a multidimensional expansion of Fisher information in the GRM derived by Samejima (1969), the Fisher information of the MGRM can be expressed as follows:

\[ I_j(\theta) = a_j a_j' \sum_{r=0}^{R} \frac{[P_{jr}'(\theta) - P_{j(r+1)}'(\theta)]^2}{P_{j(r+1)}(\theta)} \tag{58} \]

where the notation is very similar to those in the GRM and M2PL: \( a_j \) is the discrimination parameter vector; \( P_{jr}(\theta) \) is the probability of obtaining score \( r \) with ability \( \theta \); \( P_{jr}'(\theta) \) is the probability of indorsing score \( r \) or a higher score; and \( P_{j(r+1)}'(\theta) \) is the first derivative of \( P_{jr}(\theta) \) with respect to \( \theta \).

The most commonly used item selection method in multidimensional measurement CAT is D-optimal (D stands for determinant), which selects the item that maximizes the determinant of the test Fisher information matrix at the current ability estimate (Segall, 1996; Mulder & van der Linden, 2009). van Groen et al. (2016) also adopted this method in multidimensional ACT. The determinant of the Fisher information matrix is the inverse of the confidence ellipsoid volume. The item that maximizes the determinant of the Fisher information matrix can greatly decrease the confidence ellipsoid volume, thus efficiently improve ability estimate accuracy. The item \( j \) selected following D-optimal is as follows:
\[ j \equiv \arg\max_{j \in J} \left\{ \det \left( \sum_{s=1}^{S} I_s \left( \widehat{\theta}^{MLE} \right) + I_j \left( \widehat{\theta}^{MLE} \right) \right) \right\} \]  

(59)

where \( I_j \) is the information matrix of item \( j \), \( S \) items are administered, and \( J \) denotes the set of available items. When incorporating the prior information, the Bayesian version of D-optimal can be written as:

\[ j \equiv \arg\max_{j \in J} \left\{ \det \left( \sum_{s=1}^{S} I_s \left( \widehat{\theta}^{MAP} \right) + I_j \left( \widehat{\theta}^{MAP} \right) + \Phi^{-1} \right) \right\} \]  

(60)

where \( \Phi \) is the variance-covariance matrix of the \( \theta \) prior distribution. This criterion leads to the largest decrement in the Bayesian credibility ellipsoid around \( \widehat{\theta}^{MAP} \) (Wang & Chang, 2011; van der Linden, 1999; Luecht, 1996; Segall, 1996). The Bayesian estimate overcomes the potential convergence problem in MIRT, thus is more preferable than the MLE estimate. Therefore, the Bayesian version of D-optimal is usually adopted in multidimensional measurement CAT and ACT (Nydick, 2013).

In unidimensional mastery ACT, the item can be selected to have highest information at the cutoff point. In multidimensional mastery ACT, with the composite score methods proposed by van Groen et al. (2016), an item can be selected to maximize the determinant of Fisher information at the RC cutoff point \( \theta_c \). Then item \( j \) is selected as

\[ j \equiv \arg\max_{j \in J} \left\{ \det \left( \sum_{s=1}^{S} I_s \left( \theta_c \right) + I_j \left( \theta_c \right) \right) \right\} \]  

(61)
4.2.2 Maximum Trace of Inverse Fisher Information Matrix (A-optimal)

Trace is another commonly used way to aggregate a Fisher information matrix (van der Linden, 1999). The trace of the inverse Fisher information matrix is the summed asymptotic variance of $\theta$ across all dimensions. The optimal item $j$ is chosen as:

$$j \equiv \arg\min_{j \in J} \left\{ \text{tr} \left[ \left( \sum_{s=1}^{S} I_s \left( \hat{\theta} \right) + I_j \left( \hat{\theta} \right) \right)^{-1} \right] \right\}$$ (62)

When the prior information $\Phi$ is added to this formula, the Bayesian version of A-optimal (Wang, Weiss, & Shang, 2019) becomes

$$j \equiv \arg\min_{j \in J} \left\{ \text{tr} \left[ \left( \sum_{s=1}^{S} I_s \left( \hat{\theta} \right) + I_j \left( \hat{\theta} \right) + \Phi^{-1} \right)^{-1} \right] \right\}$$ (63)

In mastery ACT, similar to D-optimal, A-optimal at current ability estimates can also be used (Nydick, 2013).

4.2.3 Maximum Fisher Information of Composite Ability (C-optimal)

When the item bank is multidimensional but only an estimate of a specific linear combination of the abilities is studied, an item that brings the most benefit in this particular direction should be selected (van der Linden, 1999; Mulder & van der Linden, 2009). The composite ability can be expressed as $\theta_p = \lambda^T \theta$. $\lambda$ is the weight vector used
to construct this composite ability, so it is restricted by $\forall \lambda_k \geq 0$, and $\sum_{k=1}^{K} \lambda_k = 1$. Then, item $j$ is selected to minimize the variance of $\lambda^T \theta$:

$$j \equiv \arg\min_{j \in J} \{ \lambda^T \text{Var}(\hat{\theta}) \lambda \}$$

$$= \arg\min_{j \in J} \{ \lambda^T \left( \sum_{s=1}^{S} I_s(\hat{\theta}) + I_j(\hat{\theta}) + \Phi^{-1})^{-1} \right) \lambda \}$$

(64)

Comparing to D-optimal and A-optimal, it is more intuitive to adopt C-optimal in multidimensional mastery ACT: items that can best separate masters from non-masters are desired. Thus, information perpendicular to the classification bound is of the most importance. The norm vector $\theta_{\delta}$ in the constrained method, projected method, and RC method can be used to find the direction of interest (Nydick, 2013). Let $\lambda = a \theta_{\delta}$, where $a$ is a positive number which ensures the weight vector $\lambda$ is parallel to the norm vector $\theta_{\delta}$. Accordingly, the item which minimizes the variance of the composite ability at $\theta_c$ is selected.

As scoring in MIRT is generally difficult, especially with multiple dimensions and short test length, the cutoff point $\theta_c$ is affected by estimation error. Nydick (2013) further proposed to use weighted C-optimal across the classification surface $g(\theta) = 0$ as an item selection criterion. The posterior distribution of $\theta$ is used as the weight.

4.2.4 Smallest Eigenvalue of Fisher Information Matrix (E-optimal)

When a test is aimed to measure all dimensions in a balanced way, especially when all dimensions need to be measured to a certain accuracy level, the most inaccurate
dimension needs to be favored in the subsequent item selection. E-optimal maximizes the smallest eigenvalue of the information matrix, or equivalently, minimizes the largest generalized variance of the ability estimate (Mulder & van der Linden, 2009). In spite of the good intention, E-optimal may behave unfavorably. The contribution of an item with equal discrimination parameters to the test information vanishes when the sampling variances of the ability estimators have become equal to each other. This fact contradicts the fundamental rule that the average sampling variance of an MLE should always decrease after a new observation. Therefore, E-optimal results in occasionally inefficient item selection, and its use is not recommended.

4.2.5 D-optimal with Posterior Weights on Cutoff Points (PWCD-optimal)

Comparing the few Fisher information matrix-based item selection methods, D-optimal is the most widely used and it results in robust good performance in various conditions. Thus, D-optimal was modified to be used in grid classification ACT. The A-optimal, C-optimal, and even E-optimal can all be modified in the same way.

By generalizing the posterior weighted Fisher information on cutoff points proposed by Weissman (2007) to the multidimensional condition by using the determinant of Fisher information matrix, the posterior weighted D-optimal on cutoff points is proposed. As more than one cutoff point are of interest, the weighted sum of D-optimal on the cutoff points is a reconciliation between the direct summation of D-
optimal on all the cutoff points and the D-optimal on the nearest cutoff point. Item $j$ is chosen to satisfy

$$j \equiv \arg \max_{j \in J} \left\{ \sum_{\theta \in \Theta_c} \pi(\theta|x_S) \det \left( \sum_{s=1}^{S} I_s (\hat{\theta}) + I_j(\hat{\theta}) + \Phi^{-1} \right) \right\}$$  \hspace{1cm} (65)$$

where $\Theta_c$ is the set of all cutoff points and $\pi(\theta|x_S)$ is the posterior distribution of $\theta$ based on response pattern $x_S$. Due to the good performance of the Bayesian version of D-optimal, the inverse of the prior variance-covariance matrix $\Phi^{-1}$ is added to the formula.

### 4.2.6 Multidimensional KL Information (MKL)

KL information measures the difference between two distributions, thus would always be a scalar irrespective of the dimensionality of the two distributions. In this way, only the integrating range needs to be changed when generalizing the KL information item selection method to multidimensional measurement CAT. The item $j$ that provides largest KL information is selected as follows:

$$j \equiv \arg \max_{j \in J} \int_{\theta} KL_j(\theta_1|\theta_2) \, d\theta$$  \hspace{1cm} (66)$$

where

$$KL_j(\theta_1|\theta_2) = P_j(\theta_1) \log \frac{P_j(\theta_1)}{P_j(\theta_2)} + \left(1 - P_j(\theta_1)\right) \log \frac{1 - P_j(\theta_1)}{1 - P_j(\theta_2)}$$  \hspace{1cm} (67)$$
for dichotomous items and

$$KL_j(\theta_1|\theta_2) = \sum_{r=1}^{R} P_{jr}(\theta_1) \log \frac{P_{jr}(\theta_1)}{P_{jr}(\theta_2)}$$

(68)

for polytomous items.

Similar to unidimensional measurement CAT, Wang and Chang (2011) utilized KL information around $\hat{\theta}$. Thus, the optimal item $j$ is selected as:

$$j \equiv \arg\max_{j\in J} \int_{\tilde{\theta}_1-\zeta}^{\tilde{\theta}_1+\zeta} ... \int_{\tilde{\theta}_K-\zeta}^{\tilde{\theta}_K+\zeta} KL_j(\hat{\theta}|\theta) \partial \theta$$

(69)

where $\zeta$ is still a small number that denotes the width of the integrating range around $\hat{\theta}$, and $\tilde{\theta}_1 ... \tilde{\theta}_K$ are the $1^{st}$ to $K^{th}$ elements of the ability estimate $\hat{\theta}$. Hence, the integrating range is a hypercube with length $2\zeta$.

Because the true ability is unknown, Veldkamp and van der Linden (2002) proposed a Bayesian version of KL, posterior weighted KL (PWKL), by utilizing the posterior distribution of $\theta$ as a weight. Thus, the optimal item $j$ is selected as

$$j \equiv \arg\max_{j\in J} \int_{\theta}\pi(\theta|x_S) KL_j(\hat{\theta}|\theta) \partial \theta$$

(70)

where $\pi(\theta|x_S)$ is the posterior distribution of $\theta$. As it is a Bayesian method, the Bayesian ability estimate $\hat{\theta}_{MAP}$ is used as $\hat{\theta}$.

Nydick (2013) generalized MKL to multidimensional ACT in two ways. The S-KL utilizes the classification surface $\Theta_c$ which satisfies $g(\theta) = 0$ as the integrating range $\Theta$, and adopts Wang and Chang’s (2011) formula. Whereas the L-KL is a generalization
of Eggen (1999) and Lin and Spray’s (2000) method, in which the multidimensional
cutoff points are used to construct KL information. Accordingly, the item that maximizes
$KL_j(\theta_u|\theta_i)$ is selected to be administered next.

4.2.7 PWKL with Posterior Weights on Cutoff Points (PWCKL)

All the KL-based methods and odds ratio-based methods developed for mastery
ACT (Nydick, 2013; van Groen, 2016) are suited for only the two categories mastery
condition, as they essentially compare between the two end points of the only
indifference region. In order to account for more cutoff points in grid multi-classification
ACT, a weighted method as used in D-optimal with posterior weight on cutoff points is
adopted. In addition, although the KL-based methods and odds ratio-based methods were
developed using dichotomous IRT models, they can be easily generalized to be used with
polytomous models.

PWKL (Veldkamp & van der Linden, 2002) takes the entire posterior ability
distribution into account. The integration is usually accomplished using numerical
integration which is proportional to the summation of the function values on quadrature
nodes. As there are usually more quadrature nodes than cutoff points, the newly
developed method which constrains the integrating range to the cutoff points can be seen
as a simplified version of PWKL. The optimal item $j$ is selected as:

$$
j = \arg\max_{j \in J} \left\{ \sum_{\theta \in \Theta_c} \sum_{r=1}^{R} P(x_j = r|\bar{\theta}) \log \frac{P(x_j = r|\bar{\theta})}{P(x_j = r|\theta_i)} \right\}
$$

(71)
where \( \Theta_c \) is the set containing all cutoff points and \( \pi(\theta|x_s) \) is the posterior distribution of \( \theta \) based on response pattern \( x_s \).

### 4.2.8 Multidimensional Mutual Information (MMI)

The multidimensional generalization of MI (MMI) is very intuitive, by replacing the unidimensional \( \theta \) with the multidimensional \( \Theta \).

\[
MMI_j = \sum_{r=1}^{R} \int P(x_j = r | \Theta) \pi(\theta|x_s) \log \frac{P(x_j = r | \theta)}{P(x_j = r | x_s)} d\Theta
\]

(72)

where \( \Theta \) is the set of all possible \( \Theta \). In measurement CAT, \( \Theta \) is taken to be the entire multidimensional space. However, in mastery ACT it is more appropriate to constrain \( \Theta \) to the cutoff point-related values. For that reason, Nydick (2013) proposed using the classification bound \( \Theta_0 \) as the integrating range.

### 4.2.9 MMI with Posterior Weights on Cutoff Points (PWCMMI)

Similar to PWCD-optimal and PWCKL, this newly developed MMI-based method also focuses on the classification cutoff points. Since MMI already contains the posterior distribution \( \pi(\theta|x_s) \), the new method only changes from integrating over the entire multidimensional space to summing over all cutoff points. In this way, the best item \( j \) is selected as
\[ j = \text{argmax}_{j \in J} \left\{ \sum_{x_j \in X} \sum_{\theta \in \Theta_c} P(x_j | \theta) \pi(\theta | x_S) \log \left( \frac{P(x_j | \theta)}{P(x_j | x_S)} \right) \right\} \]  \hspace{1cm} (73)

where \( P(x_j | x_S) \) is the posterior predictive probability, and

\[ P(x_j | x_S) = \int_{-\infty}^{\infty} \pi(\theta | x_S) P(x_j | \theta) d\theta \]  \hspace{1cm} (74)

Numerical integration is used to obtain the integration on the entire multidimensional real space. In this way, the computational formula for \( P(x_j | x_S) \) is

\[ P(x_j | x_S) = w \sum_{q=1}^{Q} \pi(\theta^q | x_S) P(x_j | \theta^q) \]  \hspace{1cm} (75)

where \( \theta^q \) is the \( q^{th} \) quadrature node. Let \( n \) values be selected from \(-3\) to \(3\) in each of the three dimensions, resulting in \( Q = n^3 \) quadrature nodes selected from the entire space.

The volume of each quadrature node is computed as \( w = \frac{(3+3)^3}{Q} = \frac{216}{Q} \).

The posterior ability distribution \( \pi(\theta^q | x_S) \) used in Equation 75 is computed iteratively using the posterior predictive probability from the previous step \( P(x_S | x_{S-1}) \). It is written as

\[ \pi(\theta^q | x_S) = \frac{\pi(\theta^q | x_{S-1}) \times P(x_S | \theta^q)}{P(x_S | x_{S-1})} \]  \hspace{1cm} (76)

This calculation is rather computationally intensive. On the other hand, the posterior predictive probability \( P(x_j | x_S) \) can also be treated as the weighted average of the \( P(x_j | \theta^q) \), while the weight is the posterior distribution \( \pi(\theta^q | x_S) \). In this way,
\[ P(x_j | x_S) = \frac{\sum_{q=1}^{Q} \pi(\theta^q | x_S) P(x_j | \theta^q)}{\sum_{q=1}^{Q} \pi(\theta^q | x_S)} \] (77)

As \( \pi(\theta^q | x_S) \) for different quadrature nodes share the same standardize constant \( P(x_S | x_{S-1}) \), the unstandardized posterior distribution, namely the Bayesian version of likelihood \( L(\theta^q | x_S) \) can replace \( \pi(\theta^q | x_S) \) to simplify computation. As the Bayesian version of likelihood \( L(\theta^q | x_S) \) is just the product of a likelihood and a prior, it is way simpler to compute than the posterior distribution \( \pi(\theta^q | x_S) \). Thus,

\[ P(x_j | x_S) = \frac{\sum_{q=1}^{Q} L(\theta^q | x_S) P(x_j | \theta^q)}{\sum_{q=1}^{Q} L(\theta^q | x_S)} \] (78)

### 4.3 Summary

Various termination criteria in ACT were reviewed. Most methods can efficiently classify examinees while keeping a short yet individually different test length. Current multidimensional classification methods are essentially unidimensional classification based on a composite score. Due to arbitrariness in composite score construction, the classification results are difficult to explain. Grid classification which classifies examinees into one of many hypercubes confined by the cutoff score along each dimension can provide more informative and instructive classification results. However, there lacks a good termination criterion for grid classification. Among the existing methods, CI and SPRT are the only two that can be directly used in grid classification.
Therefore, two new termination criteria, GGLR and SGLR, were specifically developed for grid classification.

Multiple item selection methods both used in ACT as well as in measurement CAT were reviewed. Some item selection methods developed for measurement CAT and adopted from composite score-based multidimensional ACT, such as D-optimal, PWKL, and MMI, can be directly applied to the grid multi-classification ACT. The D-optimal variations which are based on the RC cutoff score cannot be used in grid multi-classification ACT, as there exist multiple cutoff points.

In addition, three new item selection methods, PWCD-optimal, PWCKL, and PWCMMI, were developed particularly for grid multi-classification ACT. All the item selection methods and termination criteria that are suitable for grid multi-classification ACT were compared in the following three studies.
5. Study 1: ACT and Measurement CAT Classification

Efficiency

Two different approaches were explored for classification purposes using adaptive testing. The first approach is to perform a measurement CAT and then perform an ad-hoc classification, referred to as a two-step approach hereafter. The second approach is to conduct an ACT. Smits and Finkelman (2013) compared these two approaches in unidimensional testing and found that ACT was more efficient than the two-step approach in terms of average test length. However, this comparison has not been performed in multidimensional testing. This simulation study was designed to explore the relative performance of ACT in comparison to the measurement CAT-based two-step approach. In order to make the two-step approach comparable to ACT, a variable length measurement CAT was implemented.

5.1 Method

5.1.1 Item Bank and IRT Model

The three-dimensional MGRM ($K = 3$) with four response categories ($R = 4$) was used. The item bank contained $J = 300$ between-item multidimensional (simple structure) items (van Groen et al., 2016; Jiang, Wang, & Weiss, 2016; Seitz & Frey, 2013). One hundred items loaded on each of the three dimensions.
As simple structure was assumed, only one of the three $a$ parameters in each item was non-zero. For this reason, the correlations among different discrimination parameters were not specified. The multidimensional $a$ parameters were randomly sampled from $U[1.1, 2.8]$ (Jiang et al., 2016; Wang et al., 2018).

The boundary parameters $b$s ranged from $-2$ to $2$, and each parameter was uniformly distributed along an equidistant interval within this range for an item. Thus, the three boundary parameters $b$s in each item were sampled from $U[-2, -0.67]$, $U[-0.67, 0.67]$, and $U[0.67, 2]$, respectively (Jiang et al., 2016; Wang et al., 2018). To avoid sparseness of the response matrix, only adjacent boundary parameters with a distance of at least 0.5 apart were retained; otherwise the parameters of an item were generated again.

### 5.1.2 Classification Settings

Examinees were classified into four groups along each of the three dimensions ($K = 3$), so three cutoff points were needed for each dimension ($C = 3$). Altogether, there were $4^3 = 64$ categories (classification grids). The cutoff scores on each dimension were set to be $-0.67$, $0$ and $0.85$ such that approximately 25% of examinees belonged to each category. In practice, different dimensions can have different cutoff scores. The minimum test length was set to be 7 items, while the maximum test length was set to be 60 items. If the termination criterion or criteria set was not reached after administering 60
items, the test was stopped, and the examinee was classified based on the relative location of the final ability estimate and the cutoff scores along each dimension.

The indifference region plays an important role in the GGLR, SGLR, and the between-item multidimensional SPRT. In previous studies, the width of the indifference region was determined arbitrarily, generally between 0.2 and 0.6. As the width of the indifference region does not affect the results of ACT, the width is more reasonably determined based on the psychometric meaning of indifference. A proper reference is the standard error (SE) of the \( \theta \) estimate. The width of the indifference region should be wider than the confidence interval of the majority of the \( \theta \) estimates. In order to obtain information about the confidence interval of the \( \theta \) estimates, a 60-item (maximum test length of the ACT), fixed-length, three-dimensional CAT was first simulated. The above-mentioned simulated item bank and the D-optimal item selection method were used. Two ability settings were tested: ability on all three dimensions was set to the same value; and while the ability of dimension 1 varied, dimension 2 and 3 abilities were fixed to zero. The standard error of estimate varying with ability on dimension 1 was recorded. Both ability settings produced very similar results. Figure 5 is based on the scenario in which \( \theta_1 \) varied and \( \theta_2 = \theta_3 = 0 \). It shows that the SE was smaller than 0.3 with \( \theta \) varying between \(-2.55\) and \(2.55\), which accounted for around 99% of the examinees (based on the standard normal population distribution). In this way, the 95% confidence interval of the \( \theta \) estimate was smaller than 0.6. Thus, the half width of the indifference region \( \delta \) was set to be 0.3.
5.1.3 Latent Trait Distribution

The true $\theta$s were simulated from a multivariate normal distribution. As dimension correlation may influence classification accuracy, three levels of between-dimension correlations were simulated. To maintain generality, instead of defining the exact correlation matrices, eigenvalues were used to control the magnitude of correlations while allowing random variations between correlation matrices between repetitions. In this way, the following three eigenvalue sets were used: eigenvalues (1, 1, 1) resulted in an identity correlation matrix, eigenvalues (1.6, 0.8, 0.6) resulted in a correlation matrix with small correlation among dimensions (varying from 0.2 to 0.4), and eigenvalues (2.2, 0.5, 0.3) resulted in a correlation matrix with medium correlations among dimensions (varying from 0.5 to 0.7). Waller’s (2018) method was adopted to generate correlation

Figure 5. Possible standard error by fixing $\theta_2$ and $\theta_3$ to 0
matrices from eigenvalue sets. High correlations between dimensions are expected to expedite classification, thus resulting in higher correct classification rates and shorter tests. The sample size was fixed at \( N = 500 \).

5.1.4 Item Selection Methods

The first few items are usually selected from items with medium difficulties (e.g. \( b \) between −1 and 1) when employing dichotomous items in adaptive testing. As one polytomous item can provide high information on a wide range of ability levels, the constraint on item difficulty is unnecessary. In this way, each of the first three items was selected randomly from a different dimension. D-optimal was adopted in a measurement CAT-based, two-step procedure from the 4\(^{th} \) item on.

In the grid multi-classification ACT, on which this thesis is focused, there are multiple cutoff scores along each dimension; therefore, the composite score methods do not apply. There are then two options: using the current ability estimate-based item selection methods and developing new item selection methods that are suitable for grid multi-classification ACT. In addition to D-optimal, PWKL, and MMI, all of which enjoy great popularity in measurement CAT, three new methods, the PWCD-optimal, PWCKL, and PWCMMI, were specifically designed for grid multi-classification ACT. Thus, six item selection methods were available for the grid multi-classification ACT.

Before conducting the full-scale simulation study and to save time, a pilot study was first conducted to identify the better performing item selection methods among the
six. Also, for the sake of time, this pilot study utilized only one item bank and 500 examinees (no repetitions). The aforementioned multivariate normally distributed population and item selection method for the first three items were utilized. The classification and estimation accuracy using each item selection method from the 4th item to the 60th item was recorded and shown in Figure 6. Nine quadrature points in each dimension, that is $9^3 = 729$ total quadrature nodes, were used in PWKL and MMI.

![Figure 6](image)

*Figure 6. Classification and estimation accuracy with varying item selection methods*

Grid classification correct rate (GCCR) was used to evaluate the classification accuracy, and mean ability estimate RMSE was used to evaluate estimation precision. It can be concluded from Figure 6 that the PWCKL and PWCMML methods resulted in the
worst classification accuracies and estimation precisions, whereas the other four methods performed similarly well to each other.

Both MMI and PWKL had considerably high accuracy yet they are computationally very time consuming. The time utilized by MMI or PWKL is generally proportional to the number of quadrature nodes used in the numerical integration. To save processing time for MMI and PWKL, another pilot study was conducted to demonstrate the feasibility of using fewer quadrature points along each dimension. Five, seven, and nine quadrature points were used along each dimension, and Figure 7 shows there was no notable difference caused by the number of quadrature points along each dimension. To save computation time, five quadrature points along each dimension were utilized.
Figure 7. Classification and estimation accuracy with varying number of quadrature points

Due to the established excellence of MMI in measurement CAT and its slightly better performance than PWKL in these two pilot studies, MMI was retained in the follow-up studies. In summary, three item selection methods: D-optimal, PWCD-optimal, and MMI, were used in the grid multi-classification ACT from the 4th item onward.

5.1.5 Termination Criteria

Four termination criteria were used in the grid multi-classification ACT. Among the four, two were newly developed specifically for grid classification: the GGLR and SGLR; and two were generalized to grid classification from classification on one dimension: the between-item multidimensional CI and SPRT.

The compound termination criteria (Wang et al., 2019) were used to stop measurement CAT in the two-step approach, where the D-optimal rule is the primary termination criterion and the absolute change in $\theta$ (CT) rule is the secondary termination criterion. With this compound termination criterion, measurement CAT was stopped either when the determinant of the Fisher information matrix exceeded a predetermined threshold, or when the absolute changes of $\theta$s along all dimensions were smaller than the cutoff. Following Wang’s et al. (2019) results, the cutoff value of the D-optimal and CT-rule were set to be 477.25 and 0.01.
To ensure that the two-step approach and ACT had roughly the same level of classification accuracy so that the test length could be compared, for each combination of ACT item selection method and termination criteria, several levels of $\alpha$ were tried. Then, the $\alpha$ level which resulted in the same GCCR as the two-step approach was retained for that item selection method and termination criterion combination.

### 5.1.6 Interim Ability Estimation

The ability estimate was first obtained after the first three items were answered by each examinee, then it was updated after each item administration. Only by incorporating the correlations among $\theta$s in the ability estimation can a between-item multidimensional test be more efficient than multiple separate unidimensional tests (Segall, 1996; Wang & Chen, 2004). Thus, a Bayesian method was used instead of MLE. Due to computational simplicity, MAP was chosen. A vague prior was used in this study. It is a multivariate normal prior with a mean zero vector and a covariance matrix of $10 \times \Sigma$. Without loss of generality, $\Sigma$ was chosen to be 

$$
\begin{bmatrix}
1 & .4 & .4 \\
.4 & 1 & .4 \\
.4 & .4 & 1 \\
\end{bmatrix}
$$

This is not a correlation matrix actually used to generate any population, but it represents moderate correlation between dimensions. The MAP used to update examinee’s ability is shown as follows:

$$
\hat{\theta} = \arg\max_{\theta \in \Theta} \{ \log L(\theta|x) + \log \pi(\theta) \} \tag{79}
$$

where $L(\theta|x)$ is the likelihood function, and $\pi(\theta)$ is the aforementioned prior distribution of $\theta$. 
5.1.7 Evaluation Criteria

Average test length (ATL) was used to evaluate test efficiency. A shorter ATL indicates a more efficient combination of the item selection method and the termination criterion. In this study, as the classification accuracy was controlled, ATL is the only evaluation criteria shown in the results. It is computed as follows:

\[ ATL = \frac{\sum_{n=1}^{N} h_n}{N} \]  

(80)

where \( N \) is the sample size, and \( h_n \) is the test length of examinee \( n \).

Grid correct classification rate (GCCR) was used to measure grid classification accuracy. It is the average correct classification rate (CCR) over all \( K \) dimensions.

\[ GCCR = \frac{\sum_{k=1}^{K} CCR_k}{K} \]  

(81)

where the CCR of dimension \( k \) is computed as follows:

\[ CCR_k = \frac{\sum_{n=1}^{N} (1 - sgn(|\hat{v}_{nk} - v_{nk}|))}{N} \]  

(82)

where \( \hat{v}_{nk} \) and \( v_{nk} \) are the test-based classification and true grouping of examinee \( n \) in dimension \( k \). \( sgn(x) = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases} \). Although the classification accuracy was controlled in the comparison between ACT and the two-step approach, the GCCR is still reported to provide a complete picture of the performance of each approach.


5.1.8 Overall Conditions

Ultimately, there were 3 (dimension correlation) \times 13 (testing procedure: measurement CAT-based two-step approach, and ACT with 3 item selection methods \times 4 termination criteria) = 39 conditions. 20 replications were applied to each condition. The simulation was conducted using basic R (R Core Team, 2018). As item selection using MMI is very time consuming, three R packages: Rcpp (Eddelbuettel & Fancois, 2011), RcppEigen (Bates & Eddelbuettel, 2013), and RcppNumerical (Qiu, Balan, Beall, Sauder, Okazaki, & Hahn, 2018) were used to program MMI so as to accelerate computation.

5.2 Results

The ATL and GCCR using the two-step approach were first implemented. With the cutoff value of D-optimal rule set at 477.25 and that of the CT-rule set at 0.01 (Wang et al., 2019), the two-step approach reached a GCCR of 0.74, irrespective of dimension correlation. This GCCR level then was used as the benchmark to search for the appropriate \( \alpha \) level for each item selection method and termination criterion combination in ACT. The appropriate \( \alpha \) levels for GGLR, SGLR, CI, and SPRT were 0.1, 0.08, 0.67, and 0.35, respectively.

This search procedure revealed that only the \( \alpha \) level applied in each termination criterion affected the classification accuracy represented by the GCCR. The item
selection methods did not influence the final classification accuracy when the $\alpha$ level was fixed.

The results of both the two-step approach and ACT are shown in Table 1. The numbers inside parentheses are the standard deviations over 20 repetitions. The standard deviations were relatively small indicating the comparison between methods was reliable.
Table 1. The two-step approach and ACT classification efficiency

<table>
<thead>
<tr>
<th></th>
<th>Two-step</th>
<th>ACT</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D=477</td>
<td>GGLR with α=0.1</td>
<td>SGLR with α=0.08</td>
<td>CI with α=0.67</td>
<td>SPRT with α=0.35</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CT=0.01</td>
<td>D</td>
<td>PD</td>
<td>M</td>
<td>D</td>
<td>PD</td>
</tr>
<tr>
<td>Terminal</td>
<td></td>
<td>D</td>
<td>PD</td>
<td>M</td>
<td>D</td>
<td>PD</td>
</tr>
<tr>
<td>selection</td>
<td></td>
<td>D</td>
<td>PD</td>
<td>M</td>
<td>D</td>
<td>PD</td>
</tr>
<tr>
<td>ATL</td>
<td></td>
<td>D</td>
<td>PD</td>
<td>M</td>
<td>D</td>
<td>PD</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.23)</td>
<td>(0.50)</td>
<td>(0.52)</td>
<td>(0.61)</td>
<td>(0.39)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.22)</td>
<td>(0.42)</td>
<td>(0.39)</td>
<td>(0.59)</td>
<td>(0.33)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.22)</td>
<td>(0.35)</td>
<td>(0.45)</td>
<td>(0.77)</td>
<td>(0.26)</td>
</tr>
<tr>
<td>GCCR</td>
<td></td>
<td>D</td>
<td>PD</td>
<td>M</td>
<td>D</td>
<td>PD</td>
</tr>
<tr>
<td>no.cor</td>
<td>0.74</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>small.cor</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>medium.cor</td>
<td>0.74</td>
<td>0.75</td>
<td>0.74</td>
<td>0.75</td>
<td>0.75</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
</tbody>
</table>

Note: D stands for D-optimal; PD stands for PWCD-optimal; M stands for MMI.
As GCCR was controlled, only the test lengths needed to be compared. Figure 8 shows the average test length under different item selection and termination criterion combinations.

![Diagram showing average test length with varying item selection methods and termination criteria.]

**Figure 8.** Average test length with varying item selection methods and termination criteria

According to Figure 8, ACT can be more effective than measurement CAT in classification. Except for a few conditions with MMI, ACT always resulted in shorter ATL than the two-step approach. The PWCD-optimal was the best item selection method irrespective of the termination criterion it was paired with. Moreover, the D-optimal also had promising performance. When there was no or little dimension correlation, pairing the D-optimal with the ACT termination criteria always outperformed (had shorter ATL).
pairing it with measurement CAT termination criterion. However, when the dimension correlation was of medium level, ACT using D-optimal and GGLR had slightly longer ATL than the two-step approach. MMI resulted in the longest ATL. When MMI was combined with GGLR, CI, and SPRT, ATL was even longer than that of the two-step approach. There was clear interaction between item selection methods and termination criteria. The best termination criterion was SPRT when paired with D-optimal or PWCD-optimal, while the best termination criterion switched to GGLR when MMI was in use.

While correlation between dimensions had minimal effect when using CI and SPRT termination criteria, larger correlation between dimensions actually deteriorated the efficiency of the GGLR and SGLR, which is contrary to expectation.
6. Study 2: Conditional Classification Efficiency of ACT

In addition to item selection method and termination criterion, examinee’s ability level may also affect the efficiency of ACT. Theoretically, the closer an examinee’s ability is to the cutoff points, the more difficult it is to classify the examinee, resulting in longer test length or lower classification accuracy. In this study, the conditional test length and classification accuracy were explored.

6.1 Method

6.1.1 Item Bank and Classification Settings

The same 300-item three-dimensional MGRM item bank in Study 1 was adopted. The same three cutoff scores (−0.67, 0 and 0.85) along each dimension were also inherited. The minimum and maximum test length were set to be 7 and 60 items, respectively.

Despite the existence of multiple information statistics, Fisher information is still the most commonly used. As the item bank was constructed with same quality items from the three dimensions, only one plot is needed when two dimensions are controlled and only one dimension is allowed to vary. Figure 9 shows the determinant of item bank Fisher information when \( \theta_1 \) varied from −3 to 3 and \( \theta_2 = \theta_3 = 0 \). As the item bank was built identically for each dimension (each dimension was measured by the same number of items of the same quality), there is no need to plot the item bank information with the
\( \theta_2 \) and \( \theta_3 \) variations. In addition to the item bank information, Figure 10 illustrates the ideal test information based on the ideal test with the most informative items at the true ability. The ideal test information was obtained by simulating the same 60-item (maximum test length of ACT) fixed length three-dimensional CAT as the one used to determine the half-width of the indifference region, \( \delta \), in Study 1. Similar to Figure 9, in Figure 10 only the condition with \( \theta_1 \) varying from \(-3\) to \(3\) and \( \theta_2 = \theta_3 = 0 \) was considered. The D-optimal item selection method was used to identify the most informative items at the true ability. Both Figure 9 and Figure 10 demonstrate that the item bank was of relatively high quality and had general flat information.

**Figure 9.** Determinant of item bank Fisher information by fixing \( \theta_2 \) and \( \theta_3 \) to 0
Figure 10. Determinant of possible test Fisher information by fixing $\theta_2$ and $\theta_3$ to 0

6.1.2 Latent Trait Distribution

Instead of simulating discrete $\theta$ points, $\theta$ distributions with various mean vectors and variance-covariance matrices were used to represent examinees at different locations. In order to simulate examinees around a certain cutoff point, the marginal $\theta$ standard deviation should be reduced from 1 to a smaller number. Here, 0.1 and 0.3 were used as the reduced $\theta$ standard deviations. In addition, the number of shrunk dimensions was also manipulated. If only one dimension was manipulated, $\theta_1$ had the reduced standard deviation while $\theta_2$ and $\theta_3$ maintained the standard deviation of 1; if two dimensions were manipulated, both $\theta_1$ and $\theta_2$ had the reduced standard deviation while $\theta_3$ maintained a
standard deviation of 1; if all three dimensions were manipulated, all three dimensions had the reduced standard deviation.

To choose the reasonable $\theta$ distribution, the test information was scrutinized. Both Figure 9 and 10 display the rather flat test information over the range $-2$ to $2$, which covers the span of the cutoff points well. Thus, it was expected that the specific cutoff point around which $\theta$ will be simulated does not affect the results; only the relative distance between the center of $\theta$ distribution and the cutoff point matters. In this way, the mean of $\theta$ was placed around the first cutoff score, $-0.67$, on each dimension. Three distances from the cutoff point were used: 0.02, 0.32, and 0.62. Because moving away from the targeted cutoff score $-0.67$ to the right results in moving close to another cutoff score 0, both $\theta$ distributions to the negative and positive directions of $-0.67$ were considered. In theory, the conditions with $\theta$ mean lager than $-0.67$ would have lower classification accuracy or longer test length than the conditions with same distance from $-0.67$ and $\theta$ mean smaller than $-0.67$. This is because the majority of the examinees with $\theta$ mean lager than $-0.67$ were between two cutoff scores, $-0.67$ and 0 and they are difficult to classify. In this way, six levels of $\theta$ mean were utilized: $-1.29$, $-0.99$, $-0.69$, $-0.65$, $-0.35$, and $-0.05$. Table 2 shows the mean and standard deviation vectors of all latent trait distributions ($6 \theta$ means $\times$ 2 $\theta$ standard deviations $\times$ 3 number of manipulated dimensions $= 36$ conditions). The small dimension correlation in Study 1 was used in this study. Thus, the correlation matrix was generated with eigenvalues 1.6, 0.8, and 0.6. The sample size for each $\theta$ distribution was 500.
Table 2. Latent trait distributions

<table>
<thead>
<tr>
<th>(θ mean vector)</th>
<th>(θ standard deviation vector)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1.29, 0, 0), (0.1, 1, 1)</td>
<td>(-0.99, 0, 0), (0.1, 1, 1)</td>
</tr>
<tr>
<td>(-1.29, 0, 0), (0.3, 1, 1)</td>
<td>(-0.99, 0, 0), (0.3, 1, 1)</td>
</tr>
<tr>
<td>(-1.29, -1.29, 0), (0.1, 0.1, 1)</td>
<td>(-0.99, -0.99, 0), (0.1, 0.1, 1)</td>
</tr>
<tr>
<td>(-1.29, -1.29, 0), (0.3, 0.3, 1)</td>
<td>(-0.99, -0.99, 0), (0.3, 0.3, 1)</td>
</tr>
<tr>
<td>(-1.29, -1.29, -1.29), (0.1, 0.1, 0.1)</td>
<td>(-0.99, -0.99, -0.99), (0.1, 0.1, 0.1)</td>
</tr>
<tr>
<td>(-1.29, -1.29, -1.29), (0.3, 0.3, 0.3)</td>
<td>(-0.99, -0.99, -0.99), (0.3, 0.3, 0.3)</td>
</tr>
</tbody>
</table>
6.1.3 Item Selection Methods and Termination Criteria

Similar to Study 1, each of the first three items was selected randomly from a different dimension. As this study aimed at comparing the classification efficiency of examinees at different locations, only the ACT-related item selection methods and termination criteria were employed. The three item selection methods in use were D-optimal, PWCD-optimal, and MMI, while the four termination criteria were GGLR, SGLR, CI, and SPRT. When one or two dimensions were manipulated, it was expected that classification accuracy along different dimensions would differ. For this reason, recording the CCR of each dimension was more meaningful than integrating them into a single GCCR, and it was not worthwhile to do the GCCR match as in Study 1. Accordingly, the $\alpha$ levels used for termination criteria were adopted from Study 1. Recall the $\alpha$ levels for the GGLR, SGLR, CI, and SPRT were 0.005, 0.03, 0.5, and 0.32, respectively.

6.1.4 Overall Conditions

Ultimately, $2 (\text{reduced } \theta \text{ standard deviation}) \times 3 \text{ (number of manipulated dimension)} \times 6 \text{ (} \theta \text{ mean)} = 36$ conditions were considered in this study. 20 replications were applied to each condition. The simulation was also conducted using basic R (R Core Team, 2018).
6.2 Results

Multiple factors affected the classification accuracy, including the item selection method, termination criterion, number of manipulated dimensions, and mean and standard deviation of $\theta$ distribution. As the interactions among these factors were quite complicated, data visualization, in addition to the ANOVA analysis, is presented. Figures 11 to 14 are the multifaceted figures presenting the influence of the five factors at the same time. The effect of $\theta$ mean was not expected to be linear, so the $\theta$ mean was split into two factors in the ANOVA analysis. They are location relative to the cutoff score (left and right) and distance from the cutoff score (0.02, 0.32, and 0.62). Therefore, six factors and all their interactions were included in the ANOVA analysis.
Figure 11. Average test length with varying item selection methods, termination criteria, \( \theta \) mean, \( \theta \) standard deviation, and number of manipulated dimensions

Figure 11 shows the ATL under different conditions. The vertical black dashed line denotes the location of the first cutoff score (~0.67). As it is very easy to make an effect statistically significant by increasing sample size, effect size \( \eta^2 \) instead of significance was used to evaluate the importance of an effect. \( \eta^2 \) is the proportion of variance an effect can explain. The rule of thumb \( \eta^2 > 0.05 \) was used to select influential effects (Cohen, 1988). Based on this criterion, five factors had critical influence on ATL: item selection method, termination criterion, location relative to the cutoff score, interaction between distance from the cutoff score and location relative to the cutoff score, and interaction between location relative to the cutoff score and number of manipulated dimensions.
dimensions. Specifically, with the increment of $\theta$ mean, the ATL increased. Note that, because the classification accuracy was not controlled in this study, the resulted ATL is not entirely comparable to that in Study 1.

![Figure 12. CCR of dimension 1 with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions](image)

Figure 12. CCR of dimension 1 with varying item selection methods, termination criteria, $\theta$ mean, $\theta$ standard deviation, and number of manipulated dimensions
Figure 13. CCR of dimension 2 with varying item selection methods, termination criteria, \( \theta \) mean, \( \theta \) standard deviation, and number of manipulated dimensions
Figure 14. CCR of dimension 3 with varying item selection methods, termination criteria, \(\theta\) mean, \(\theta\) standard deviation, and number of manipulated dimensions

Figures 12 to 14 display the CCR of the three dimensions, respectively. Among them, it can be seen that there were different CCR trends in the manipulated dimensions and the dimensions that retained a standard deviation of 1. The ANOVA analysis results also support the considerable influence of manipulated dimensions. As dimension 1 was always manipulated, only the interaction between distance from the cutoff score and location relative to the cutoff score was of crucial importance. The interactions that involve number of manipulated dimensions had critical influence on the CCRs of dimension 2 and dimension 3.
The CCR of a manipulated dimension decreased when the $\theta$ mean increased to the first cutoff score ($-0.67$) from the negative end of the $\theta$ scale. Combining this observation with the trend in ATL (ATL increased with the $\theta$ mean), it can be concluded that the classification efficiency was lower when examinees’ abilities were closer to the cutoff score. When the reduced $\theta$ standard deviation was 0.1, with the increase of the $\theta$ mean from the first cutoff score ($-0.67$) to the second cutoff score (0), the CCR of a manipulated dimension first increased then decreased. However, when the reduced $\theta$ standard deviation was 0.3, the aforementioned trend did not hold, and the CCR was relatively low when the $\theta$ mean was $-0.35$. This may be because of the relatively wide examinee ability distribution. A large proportion of examinees were located close to either cutoff scores when the $\theta$ mean was $-0.35$ which leads to a relatively low CCR.

On the other hand, the CCR of the not manipulated dimensions slightly increased with the $\theta$ mean. Theoretically, the classification efficiency would be lower around the two cutoff scores ($-0.67$ and 0). However, the substantially wide examinee ability distribution weakened the influence of cutoff scores. The incremental trend of CCR may just be caused by the escalation of ATL. Furthermore, the influence of reduced $\theta$ standard deviation was omittable for the non-manipulated dimensions.

The influence of item selection method on CCR was essentially small. The difference only appeared when the reduced $\theta$ standard deviation was 0.1 and $\theta$ mean was $-0.35$. In this specific condition, MMI resulted in higher CCR than D-optimal and PWCD-optimal. In contrast, MMI resulted in generally higher ATL then the other two
methods and MMI amplified the influence of the $\theta$ mean. That is, the ATL varied more dramatically depending on the location of the $\theta$ mean when MMI was utilized. The performance of D-optimal and PWCD-optimal were rather similar, with a slightly lower ATL and a slightly smaller influence of the $\theta$ mean on ATL when PWCD-optimal was used.

Among the four termination criteria, the influence of $\theta$ mean on ATL was the smallest when the SPRT was used, thus SPRT resulted in the most robust ATL. The impact of $\theta$ mean on ATL was ordered, from least to most, from SPRT, to CI, to SGLR, and to GGLR, respectively. As for the influence on CCR, the trend is conditional on whether the dimension was manipulated or not. SPRT resulted in the most dramatically changed CCR for the manipulated dimensions, then followed by CI, SGLR, and GGLR. On the other hand, SPRT resulted in the most robust CCR for the not manipulated dimensions. The influence of $\theta$ mean and $\theta$ standard deviation became more evidence from CI, to SGLR, and to GGLR. In other words, SPRT and CI resulted in lower CCR than SGLR and GGLR when examinees were located close to the cutoff points, while SPRT and CI resulted in higher CCR than SGLR and GGLR when examinees were located far away from the cutoff points. As stable classification accuracy across $\theta$ distribution is generally desired in ACT, SGLR and GGLR are more preferable.

Generally, the classification was more difficult when examinees were closer to cutoff scores. PWCD-optimal and D-optimal led to stable test length and classification accuracy. SPRT and CI resulted in the most stable test length and the most dramatically changed classification accuracy when the examinees were close to the cutoff scores.
7. Study 3: Evaluation of ACT and Measurement CAT with Hybrid Simulation

Several item selection methods and termination criteria were developed for grid multi-classification ACT. There were thorough comparisons in Study 1 and Study 2. However, a real data-based simulation study was needed to further provide external validity support. True item parameters and examinee parameters were employed to compare between ACT and the two-step approach in a more practical setting.

7.1 Method

7.1.1 Item bank

The item bank contained 324 Likert-scale items from the Activity Measure for Post-Acute Care Questionnaire which measures three correlated dimensions. Each item loaded on only one of the three dimensions. The Applied Cognition, Daily Activity, and Mobility dimension were measured by 107, 106 and 111 items, respectively. All the items were four-category Likert-scale items. The three-dimensional MGRM model was used for item calibration. In the MGRM model calibration, if an option was seldom endorsed (e.g., it was never selected or it was only selected once), the response of this option, if any, was combined into the responses of the next higher option. Consequently, there were 244 items with four response categories, 79 items with three response
categories, and one item with two response categories. The item parameters were calibrated using a concurrent design that incorporated data from four groups of patients, each completing a unique subset of about 90 items plus a common linking set of 24 items (eight per each of the three scales). Each group of patients was comprised of over 600 hospitalized patients, resulting in a total group of approximately 2,400 response vectors for estimating item parameters onto the common scale provided by the linking items. The estimated item parameters were treated as true item parameters in the following ACT process. Figure 15 presents the total item information per dimension by treating all the items loading on each dimension as a unidimensional test. Generally, the boundary parameters were substantially low, which resulted in highest information around $\theta = -2$. Items from the Mobility dimension were more informative, followed by the Daily Activity domain, whereas items from Applied Cognition were least informative. The test information distributions were drastically different from those in Study 1 and Study 2, which were high and flat centered around $\theta = 0$. 
Figure 15. Dimensional information of the real item bank

7.1.2 Latent Trait Distribution

The factor scores of another 366 first-time questionnaire respondents were used to generate true examinee parameters. Figure 16 shows the $\theta$ distribution of these 366 $\theta$ vectors along each dimension. The population correlation matrix for the three dimensions was

$$
\begin{pmatrix}
1 & 0.47 & 0.38 \\
0.47 & 1 & 0.84 \\
0.38 & 0.84 & 1
\end{pmatrix}
$$

To increase sample size in order to obtain more stable results, jitters were added to the 366 $\theta$ vectors to generate another 1,830 $\theta$ vectors, resulting in 2,196 $\theta$ vectors in total. The jitters were generated from a uniform distribution $ofU(-a, a)$, where $a$ is $\frac{1}{5}$ of the smallest difference between two original $\theta$ values.
Figure 16. Ability distribution on each dimension of the real examinees

The cutoff scores were chosen to be the first quantile, median and the third quantile of the original 366 $\theta$ vectors along each dimension.

7.1.3 Overall Conditions

Ultimately, there were 13 (testing procedure: measurement CAT, and ACT with 3 item selection methods $\times$ 4 termination criteria) conditions. No replication was conducted. The simulation was conducted using basic R (R Core Team, 2018).
7.2 Results

Similar to Study 1, the classification accuracy was matched, so the classification efficiency was reflected by ATL. When the compound measurement CAT termination criteria of D-optimal larger than 477.25 or CT-rule smaller than 0.01 was used, measurement CAT reached a GCCR of 0.75. Then these GCCR levels were used as benchmarks to search for the appropriate $\alpha$ level for each item selection method and termination criterion combination. When the D-optimal and PWCD-optimal method were used to select items, the appropriate $\alpha$ levels for the GGLR, SGLR, CI, and SPRT were 0.005, 0.03, 0.5, and 0.32, respectively. However, when the MMI was the item selection method, the GCCR was only 0.63 at maximum test length. Thus, the ATL of MMI was not comparable to the other two methods. Figure 17 shows the ATL under the item selection method and termination criterion combinations. The results related to MMI were excluded.
Figure 17. Average test length with varying item selection methods and termination criteria

Similar to the trend in Study 1, ACT utilizing D-optimal or PWCD-optimal as the item selection method outperformed the two-step approach in terms of classification efficiency. The MMI resulted in longer ATL than the two-step approach. However, the advantage of the PWCD-optimal did not always hold. When using the GGLR and CI, D-optimal resulted in even shorter ATL than the PWCD-optimal. Based on the ATL of more than 40 items in this study, the discrepancy in ATL between using the D-optimal and PWCD-optimal was minimal. Thus, these two methods could be treated as equally well performing in this real-data based simulation.

When either the D-optimal or the PWCD-optimal was used, SPRT resulted in the shortest ATL, followed by CI and SGLR, with GGLR resulting in the longest ATL. This trend was also analogous to that in Study 1.

Generally, the results of this study were quite similar to those of Study 1. However, due to the quality of the real item bank, the ATL in this study (around 50 items) was substantially longer than that in Study 1 (around 20 items), whereas the classification accuracy was similar (GCCR=0.75 in Study 3; GCCR=0.73 in Study 1). As the levels of measurement CAT compound termination criteria were held constant in both Study 1 and Study 3, the relationship between the termination criteria level and classification accuracy was very stable in the two-step approach. On the other hand, in
order to reach the same classification accuracy, ACT termination criteria levels varied between tests.
8. Discussion and Conclusion

Adaptive classification testing (ACT) aims at efficiently classifying examinees into two or more groups based on predetermined classification cutoff points. Various studies with both unidimensional and multidimensional testing demonstrated the high efficiency of ACT in classification. However, previous multidimensional classifications were essentially unidimensional classifications based on a composite score. Due to the arbitrariness in composite score construction, the classification results are difficult to justify. For that reason, grid classification, which classifies examinees into one of the many hypercubes confined by the cutoff score along each dimension, was proposed in this thesis. Grid classification was expected to provide more informative and instructive classification results. This thesis developed several item selection methods (PWCD-optimal) and termination criteria (GGLR and SGLR) specifically for grid multi-classification ACT using multidimensional polytomous items. Along with a few item selection methods adopted from measurement CAT (D-optimal and MMI) and termination criteria from one-dimensional classification in multidimensional ACT (SPRT and CI), the performance of grid multi-classification ACT was compared to the measurement CAT-based two-step approach in terms of classification accuracy and test length. Simulation studies based on both simulated item banks and a real item bank indicated that when D-optimal or PWCD-optimal was used, ACT was more efficient than the two-step approach. Moreover, the classification was more difficult when examinees were closer to cutoff scores in grid multi-classification ACT. The implications of these results and further studies are discussed in the following sections.
8.1 Necessity of Grid Classification

Due to the existence of correlations among traits, it is more efficient to perform one multidimensional ACT that is constructed with between-item multidimensional items than several unidimensional ACTs. However, it is impossible to split a multidimensional ACT into separate unidimensional ACTs when within-item multidimensional items are involved. Overall, it is optimal to conduct multidimensional ACT when an item bank contains items measuring different dimensions. However, previous multidimensional classifications were essentially unidimensional classifications based on a composite score. However, composite score construction is rather arbitrary, either based on pre-knowledge of the importance of each dimension (Nydick, 2013; van der Linden, 1999), or based on the item bank properties (van Groen, 2014; van Groen et al., 2016). For this reason, the classification results are difficult to justify. Moreover, many classification tests are designed to facilitate treatment following testing, especially in the education and clinical assessment areas. For example, after providing a diagnosis to a patient, doctors need to design and apply treatment to cure the patient. In order to do so, the levels this patient belongs to on each dimension are more important and informative than an indistinct general diagnosis as “sick.” In this way, classification along separate dimensions is sometimes preferred. This type of classification is referred to as grid classification, as each individual is classified into one of the grids encircled by cutoff scores (lines/surfaces) on different dimensions. This thesis focused on this less explored topic, grid classification ACT, which is essential in multidimensional classification. The
comparison with the measurement CAT-based two-step approach, and the new item selection methods and termination criteria development were all based on the grid classification scenario.

8.2 Advantages of ACT over the Two-step Approach in Grid Classification

Smits and Finkelman (2013) proved that in unidimensional testing, ACT was more efficient than the measurement CAT-based two-step approach. Study 1 further demonstrated that ACT generally required shorter test length than the two-step approach in order to reach the same classification accuracy in the grid multi-classification case. This result strengthens the necessity of developing grid classification ACT. When the testing goal is classification, utilizing ACT can shorten the test length and further elevate the test efficiency, which is in accordance with the primary goal of adaptive testing. As an example, one important application context of classification testing, clinical assessment, particularly requires short test length to alleviate the burden patients may suffer from when filling out the assessment questionnaires.

The most prominent difference between ACT and measurement CAT lies in the termination criteria. Two types of termination criteria are used in measurement CAT (Dodd, Koch, & De Ayala, 1993), namely, the standard error (SE) rule (Weiss & Kingsbury, 1984) and the minimum information rule (Maurelli & Weiss, 1981). The SE rule terminates a test when a predetermined observed standard error of the $\theta$ estimate has
been satisfied (Boyd, Dodd, & Choi, 2010). The minimum information rule, in contrast, terminates the test when there are no more available items able to provide a predetermined minimum level of information. Unlike the SE rule, the minimum information rule prevents the needless administration of low-informative items, but at the cost of delivering less accurate measurement precision (Dodd, Koch, & De Ayala, 1989). While both the SE rule and the MI rule have pros and cons, Wang et al. (2019) proposed a compound termination criterion that combines the ideas of both. By far, this was the best performing termination criterion and was adopted in Study 1 and Study 3 to represent measurement CAT termination criteria. The results show that when the same item selection method, the D-optimal, was used, ACT was always more efficient than the measurement CAT-based two-step approach in classification, irrespective of the ACT termination criterion in use. Moreover, when a more suitable item selection method, the PWCD-optimal, was applied, the performance of ACT further improved.

In this thesis, two termination criteria (GGLR and SGLR) and three item selection methods (The PWCD-optimal, PWCKL, and PWCMMI) were developed. The PWCKL and PWCMMI were not included in the full-scale studies due to the unpromising performance in the pilot study.

### 8.3 Evaluation of Grid Classification Accuracy

CCR for each dimension was used to evaluate the classification accuracy of a single dimension in Study 2, and the GCCR (i.e., the CCR averaged over all dimensions)
was employed to evaluate the overall classification accuracy in Studies 1 and 3. CCR is the most frequently used classification accuracy evaluation criterion, thus it was adopted in this study. However, previous studies generally focused on mastery ACT while this thesis concentrated on multi-classification. The current CCR, as defined, treats all types of misclassification equally. However, in multi-classification, a misclassification to an adjacent category would be less problematic than misclassification to a category that is further away from the true group. For this reason, a more appropriate evaluation criterion could be to down-weight the misclassification close to the true group and upweight the misclassification far from the true group. A weighted kappa coefficient could be a suitable candidate evaluation criterion for this reason.

Moreover, although the item selection methods and termination criteria were developed or adopted for grid classification, the classification accuracy evaluation criteria in use do not directly evaluate if the grid classification is accomplished well. The GCCR defined in this thesis is the average of CCR along all dimensions, which evaluates if an individual is classified into the correct category per dimensions, not if an individual is classified into the correct grid (or hypercube) defined collectively by the cutoffs along all dimensions. Possible criteria that reflect grid classification accuracy more authentically are either the proportion of individuals that are correctly classified into the true grid, as well as the Hamming distance (Hamming, 1950) between the examinee’s true group membership vector (that is used to define the grid) and the estimated group membership vector. The latter criterion takes into account the severity of misclassification.
8.4 Gains from Posterior Weights on the Cutoff Points

As ACT focuses on the classification around the cutoff points, the majority of the previously developed ACT item selection methods select items that provide maximum information at the cutoff points. In grid classification, there are multiple cutoffs along each dimension. Using the posterior distribution on the cutoff points to weight the information is a reasonable compromise between information average over all cutoff points and information at the closest cutoff point. On one hand, it inherits the cutoff points-related item selection tradition in ACT. On the other hand, it takes into account the different importance of cutoff points, which is reflected by the posterior density. For instance, if the posterior weight on one cutoff is low, that implies that the individual’s true $\theta$ is far away from this cutoff. Thus, the amount of information an item provides at this cutoff is less important.

This modification was applied to the D-optimal, PWKL, and MMI, all of which are well known for their excellent performance in measurement CAT. In this way, the resulting PWCD-optimal, PWCKL, and PWCMMI were all expected to exhibit good performance in the ACT. However, only the PWCD-optimal generalization was successful as it outperformed its prototype, D-optimal. PWCKL and PWCMMI had even worse performance than the original PWKL and MMI. This differential influence of posterior weights on different methods may be due to the essence of the three types of information. D-optimal represents the Fisher information, which is local information,
while the KL information and MI are usually referred to as global information. For instance, in the construction of MI, the entire posterior distribution of $\theta$ is already taken into account. Adding another layer of “posterior weight” in PWCMMI is not helpful and it may be harmful if the interim posterior distribution is far away from the ideal posterior (i.e., probability mass at true $\theta$).

Moreover, the amount of extra information that the posterior weights on the multiple cutoff points brings to the item selection method may also be used to explain the difference between PWCD-optimal and PWCKL and PWCMMI. PWCD-optimal is a weighted sum of D-optimal at all cutoff points, and thus provides more add-ons than D-optimal. On the other hand, PWKL and MMI already contain the posterior density as a weight summed over all quadrature points, while PWCP and PWCM only contain posterior density as a weight summed over all cutoff points. Since there generally were more quadrature points ($5^3 = 125$ in this study) than cutoff points ($3^3 = 27$ in this study), PWCP and PWCM provided less rather than more information compared to PWKL and MMI.

The development of PWCD-optimal was relatively successful. When the simulated item bank was used (Study 1 and Study 2), PWCD-optimal always achieved shorter test length than D-optimal. However, in the hybrid simulation (Study 3), the performance of PWCD-optimal and D-optimal were rather close. This similarity may result from the similar item selection of the two methods in the later stage of ACT. As the posterior ability distribution is fairly peaked at the later stage of ACT, the PWCD-optimal is considerably close to D-optimal at the current ability estimate. That is, the advantage of
PWCD-optimal is only prominent at the beginning of ACT. In this way, the performance of these two item selection methods becomes relatively comparable when the test lengths were generally long, as in Study 3.

The performance of MMI was not as optimal as expected in ACT. Although it is a promising item selection method in measurement CAT and has similar or even better performance than D-optimal, it was considerably less efficient than D-optimal in ACT. This may be because the global information MMI measure does not focus on the cutoff points.

8.5 Comparison among Termination Criteria

There was an interaction between the item selection methods and the termination criteria, and this interaction was consistent across the three studies. The termination criteria performed differently when pairing with different item selection methods. When either the D-optimal or the PWCD-optimal was used, SPRT was the most efficient termination criterion, followed by CI and SGLR, and GGLR was the least efficient. On the other hand, when MMI was used for item selection, SGLR preformed the best, but the rank of GGLR, CI, and SPRT varied between studies. Based on the results in Study 1 and Study 3, SPRT and CI were more efficient than the other two methods when classifying normally distributed populations. However, as shown in Study 2, SPRT and CI resulted in lower classification accuracy for examinees that were close to the cutoff points and, thus, are difficult to classify. The overall high efficiency of SPRT and CI in Studies 1 and
3 can be largely attributed to the large proportion of examinees that are far from the cutoff points in the normally distributed population. As stable classification accuracy across $\theta$ distribution is generally desired in ACT, SGLR and GGLR are more preferable.

### 8.6 Conditional Classification Efficiency

Test efficiency is reflected by both the average test length and accuracy. In addition to evaluating the general test efficiency of the general population, the conditional test efficiency for examinees at different ability levels can also be evaluated. In measurement CAT, the conditional test efficiency is affected by the item bank quality, item selection method, and termination criterion. If there are enough high informative items around the true ability level, with an appropriate item selection method and termination criterion, the test would be highly efficient. In ACT, along with the factors that influence measurement CAT, the relative location to the cutoff points also has a major effect on the conditional test efficiency. Generally, the closer an examinee is to the cutoff points, the less efficient ACT is, which may be reflected by a lower classification accuracy, a long test length, or both.

Compared to the other two item selection methods, the MMI amplified the influence of relative location on both test length and classification accuracy. Among the four termination criteria, the influence of the relative location on the test length was the smallest when the SPRT was used, thus SPRT resulted in the most robust test length. The
impact of relative location on the test length was stronger moving from CI, to SGLR, and to GGLR, respectively.

8.7 Classification with ACT or Diagnostic Classification Modeling (DCM)

Two general types of psychometric models can be used for classification: latent trait models and latent class models. These two kinds of models differ in the definition of the latent variable. The latent trait models assume that the latent variable is continuous. IRT models are representatives of latent trait models (Embretson & Reise, 2000). When IRT models are used for classification, predetermined cutoff points are required. On the other hand, latent class models presume that the latent variable has discrete categories. Diagnostic classification models (DCM; Rupp & Templin, 2008) are a group of latent class models. When the DCMs are used for classification, all examinees are naturally classified to one of the mutually exclusive latent classes along each dimension. In this way, the classification result based on DCMs coincides with the grid classification developed in this thesis.

Despite the difference in the assumption of the latent variable, the latent trait models and latent class models can be calibrated using the same responses. Hence, it is impossible to distinguish between the two merely based on the data obtained. Commonly, the model is selected based on the conceptualization of the latent variable. For example, the real data used in Study 3 came from a clinical assessment. Based on the
understanding of the three dimensions, Applied Cognition, Daily Activity, and Mobility, the latent variables were assumed to be continuous rather than discrete, and latent trait models were selected. However, the characterization of latent trait is sometimes vague. It is possible that in some applications a set of traits are treated as continuous variables and IRT models are employed, whereas in other instances the same set of traits are rendered as categorical variables, thus DCMs are applied. For example, McKinley and Way (1992) fitted the TOEFL reading test with a MIRT model, while Kasai (1997) and Scott (1998) fitted it using DCMs. Moreover, as the same responses can be used to fit either model, both sets of models are likely to be used when calibrating existing tests. In addition to improving the understanding of the practical problems and selecting the correct model, it is also necessary to explore the severity of using an incorrect model in terms of classification results.

8.8 Conclusion and Future Studies

This thesis defined a new type of classification in the multidimensional case, namely, grid classification, and compared it with the performance of several ACT item selection methods and termination criteria using a between-item multidimensional polytomous item bank. Generally, ACT was more efficient than the measurement CAT-based two-step approach for grid classification. The newly developed PWCD-optimal was the best item selection method whereas the MMI did not perform well in ACT. Thus, it was suggested that MMI be abandoned from ACT. Paired with the two best item
selection methods, D-optimal and PWCD-optimal, the best termination criterion was SPRT. The classification was more difficult when examinees were closer to cutoff scores. PWCD-optimal and D-optimal led to stable test length and classification accuracy. SPRT and CI resulted in the most stable test length and the most dramatically changed classification accuracy when the examinees were close to the cutoff scores.

Due to the promising performance of ACT, it is the recommended procedure to conduct grid classification. Thus, more studies are needed to further explore the performance of ACT in different scenarios, such as using within-item multidimensional item banks. Only between-item multidimensional item banks were used in this thesis. Although all the item selection methods and termination criteria are applicable in the within-item multidimensional test, it is still imperative to conduct grid classification ACT using within-item multidimensional item banks and explore the potential influence of item bank type.

Moreover, the current item selection methods and termination criteria were either adopted directly or generalized from unidimensional ACT and measurement CAT. More item selection methods and termination criteria specifically designed for ACT are in need.
References


Thompson, N. A. (2010). Nominal error rates in computerized classification testing. In *First annual conference of the international association for computerized adaptive testing, Arnhem, the Netherlands*.


