

出國報告（出國類別：其他—國際研討會）

2007GMAC 電腦化適性測驗 國際學術會議

服務機關：國家教育研究院籌備處

姓名職稱：蘇雅蕙助理研究員

派赴國家：美國

出國期間：96年6月5日至96年6月8日

報告日期：96年6月7日

摘要

96年6月6日至8日在美國 Minnesota 舉行之「2007 GMAC Conference on Computerized Adaptive Testing」之國際研討會，包括會前工作坊、專家演講、研討會議等。本次國際研討會由 Graduate Management Admission Council (GMAC) 籌辦，該會不僅致力於電腦化適性測驗之推動、研究、亦展示實務上控制程序該如何進行，吸引很多電腦化適性測驗領域之測量專家學者與會，發表其研究成果及進行學術交流。研討會並以圓桌討論和論文發表的方式交叉進行，鼓勵學者們分享研究經驗與成果，增加研討會的深度與廣度。為習取國際間電腦化測驗進行之經驗，促使臺灣學生學習成就評量資料庫 (Taiwan Assessment of Student Achievement; TASA) 施測能朝電腦化方向發展，並加強 TASA 與國際接軌及提升題庫建置之品質參加。本人受到大會之邀請，會中以英語口頭發表其論文，題目為 Simultaneous Online Control Over Item Exposure and Test Overlap in Computerized Adaptive Testing for Independent and Testlet-Based Items。除論文發表外，亦與其他學者有深入的交流，收穫甚多。

目次

第一章、目的	1
第二章、參加會議經過	1
第三章、與會心得	2
第四章、考察參觀活動	3
第五章、建議	3
第六章、攜回資料名稱及內容	4
第七章、其他	4
附錄一、大會議程	
附錄二、GMAC 技術報告	
附錄三、大會邀請學者演講之講義	

第一章、目的

為習取國際間電腦化測驗進行之經驗，促使臺灣學生學習成就評量資料庫(Taiwan Assessment of Student Achievement; TASA) 施測能朝電腦化方向發展，並加強 TASA 與國際接軌及提升題庫建置之品質參加，參加 96 年 6 月 6 日至 8 日在美國 Minnesota 舉行之「2007 GMAC Conference on Computerized Adaptive Testing」之國際研討會。研討會包括會前工作坊、專家演講、研討會議等，深入瞭解電腦化適性測驗之推動、研究、亦展示實務上控制程序該如何進行，並與該領域之測量專家學者分享研究經驗與成果。

第二章、參加會議經過

首先感謝國科會對國內專家學者出席國際會議並發表論文所給予的補助，才能讓我有這個機會與來自各國的學者齊聚一堂，進行學術交流與經驗分享。

96 年 6 月 6 日至 8 日在美國明尼蘇達州舉行之「2007 GMAC Conference on Computerized Adaptive Testing」國際研討會，包括會前工作坊、專家演講、研討會議等。本次國際研討會由 Graduate Management Admission Council (GMAC) 籌辦，該會不僅致力於電腦化適性測驗之推動、研究、亦展示實務上控制程序該如何進行，吸引很多電腦化適性測驗領域之測量專家學者與會，發表其研究成果及進行學術交流。研討會並以圓桌討論和論文發表的方式交叉進行，鼓勵學者們分享研究經驗與成果，增加研討會的深度與廣度。由於前一次會議是在 25 年前，因此也吸引相當多的教育界、學術界的學者前往參加，與會人士討論亦是踴躍。會場是在 University of Minnesota 校園內的 Radisson Hotel 二樓之會議室，除了大會邀請之專家演講外，同一個時間內至多有三個會場同時進行論文發表，無論是哪一種型態都是完全開放的，任何一個與會者均可任意參加，並在得到主持人的同意下進行發問。

這次與我一同前往的有國立臺灣師範大學教育心理與輔導學系陳柏熹助理教授、國立教育研究院籌備處陳清溪主任秘書等，原預計還有國立中正大學心理學系陳淑英副教授也要一同前往，但因淑英老師最後考量經費有限，而臨時請大會撤銷她的論文發表場次。陳主任秘書清溪為臺灣學生學習成就評量資料庫(Taiwan Assessment of Student Achievement; TASA) 副召集人，為習取國際間電腦化測驗進行之經驗，促使 TASA 施測能朝電腦化方向發展，並加強 TASA 與國際接軌及提升題庫建置之品質，逕行自費前往。柏熹學長跟我都是國立中正大學心理學研究所王文中教

授指導過的學生，他是以國科會計畫項目下支付出席國際會議之補助，我則是獲得國科會補助國內專家學者出席改次國際學術會議。

由於在五月中旬便已收到大會的議程，所以在出發前已稍微看過在發表之餘，預計前往參加的場次，所以我在整個會議的過程中，諸多事宜也就顯得相當從容。我在會中以英語口頭發表我的博士論文，題目為 *Simultaneous Online Control Over Item Exposure and Test Overlap in Computerized Adaptive Testing for Independent and Testlet-Based Items*。這篇論文是研究在電腦化適性測驗中，只是控制試題曝光是不能夠完全達到測驗保密；如果接受測驗的考生有幾百萬，最大曝光率訂在 0.2，看到試題的人數也可能達到幾十萬，所以並不能完全控制測驗的保密性；尤其是當模擬曝光參數的分配與施測的考生能力分配不同時，測驗保密更是受到質疑，我的論文發展出不需要先模擬試題曝光參數，線上同時控制試題曝光和測驗重疊的電腦化適性測驗的流程，的確可以有效控制試題曝光和測驗重疊的情形。論文發表後，受到不少的迴響，如 Wim J. van der Linden, Bernard P. Veldkamp, Mark D. Reckase, Charles Lewis 等大師們交換心得，陸續被這些大師稱讚研究作的不錯，覺得十分開心，然而這些討論也引發我下一個研究的方向。另外，由於時間限制我們也留下名片以便日後繼續聯繫。

第三章、與會心得

有關我個人口頭報告的部分，由於先前內容已與指導老師王文中討論不下數十次，而且也以口頭報告的方式練習過很多次，因此講稿並不陌生。另外，因為我的發表是在會議的第一天，所以我們在會議前一天抵達會場就先行探查會場；且在該場次開始之前先測試電腦及投影設備，故在當天報告時非常順利。事前作充分的準備與練習，可讓當天的報告較不緊張，也會比較順利。另外，因為我先前已經有八次國際研討會以英語口頭發表論文之經驗，所以我在整個會議的進行中，相當順利。我的論文發表之後，受到不少大師的迴響，然而這些討論也引發我下一個研究的方向與思考。

因為我在就讀博士期間，曾經拿到國科會千里馬計畫補助前往荷蘭 University of Twente 與 Cees A. W. Glas 教授進行一年的研究，所以這次出國講英文，對我來說不是什麼難事，甚至有學者問我是不是在美國唸書或長大的。這次會議也有見到荷蘭 University of Twente 的 Wim J. van der Linden, Bernard P. Veldkamp 等教授們，我帶了一些臺灣名產鳳梨酥送給他們，加強與臺灣的外交，他們十分開心。回國後，也收到荷蘭老闆 Cees A. W. Glas 教授回信感謝，並且提及未來研究的

可能性。

大會第一天晚上安排晚宴，讓與會者人士，能夠邊用一些簡單的餐點，以繼續會議中的議題，我覺得用意非常好，果然也達到這效果。與一些學者聊到，如何進入這門領域，這才發現原來都出自不同領域，可能是因為工作上的需要、或者研究上的需要，但是共同點都是對於電腦化適性測驗的熱愛，我個人深受感動。因為國內學術研討會通常只是論文發表，交換完意見就結束，似乎很少有這樣的機會交談。

第四章、考察參觀活動

大會並無安排考察參觀活動。

因王文中老師目前人在柏克萊大學作交換學者，因此在王文中老師熱情的邀請之下，陳主任秘書、柏熹學長和我在舊金山轉機，特去拜訪王文中老師與其家人。其間，文中老師開車來機場接我們，往後幾天更帶我們參觀柏克萊大學、成人學校、小學等等，並說明他們的學制與教學是如何規劃，讓我們這次參訪相當深入，也收穫很多。

第五章、建議

雖然我個人很幸運地獲得國科會補助出國國際會議，但補助補助的預算仍為有限。我個人這次申請只拿到四萬四千元的補助，這額度原本是補助學者專家出席國際會議的機票費用標準，然而今年機票因為燃料稅等等漲幅甚多，所以補助金額連付機票都不夠；開會地點在大學城裡的飯店，與其他飯店相去甚遠，因此只好忍痛荷包住在昂貴的大會指定飯店。另外，淑英老師因為考量其國科會計畫經費補助有限，因此最後決定不去，改去較近程之國際會議（IMPS 2007），所以替她覺得十分可惜，可惜的是論文獲得錄取，卻未能出席發表，增加國際視聽，著實有些可惜。因此，建議如果在經費許可的情況下，應該多開放補助出席國際學術會議的次數與金額，也建議其實也可以採用整個實驗室或者團體補助的方式，由某個教授帶隊出國，不僅可以增加國內研究學術的互動，亦可讓國內的教授們、博士生們能與國際接軌，對我國的學術風氣相信應會有相當大的助益。有鑑於這次會議中自台灣前往與會的人士委實寥寥可數，反觀自中國大陸、港澳地區的人員出國唸書後留在國外測驗機構工作的卻不少，因此若要提高國家知名度，並落實全民外交，就必須培養更多的人才，提升學術競爭力。希望未來能多參加這種會議，為國家爭取學術知名

度。

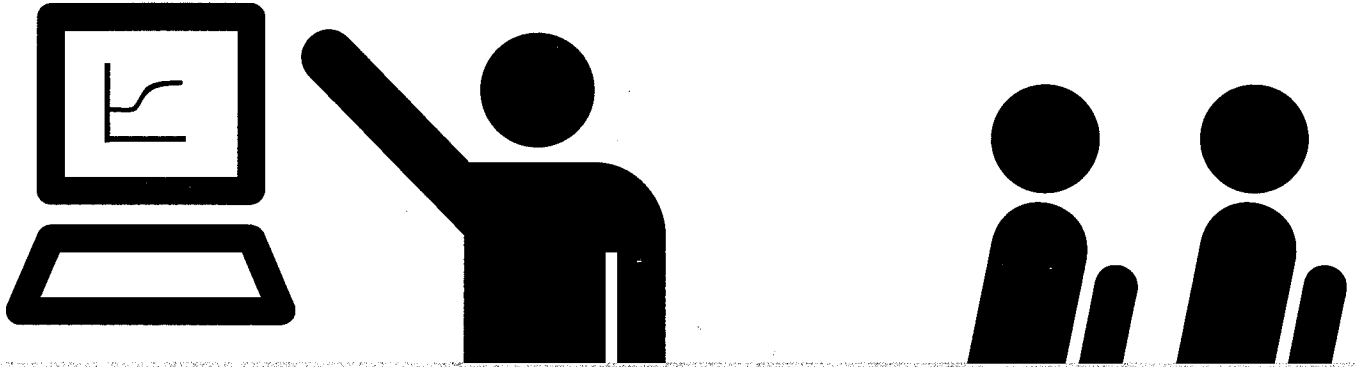
第六章、攜回資料名稱及內容

1. Conference Program: The conference program includes Keynote and Invited Lectures, as well as Invited and Contributed Sessions. (附錄一)
2. Owens, K. M. (2007). Use of the GMAC Analytical Writing Assessment: Past and Present. (GMAC Research Report RR-07-01). McLean, Virginia: GMAC. (附錄二)
3. Handout (附錄三):
 - 3.1. Reckase, M. D. (2007). Designing item pools to optimize the functioning of a CAT.
 - 3.2. Samejima, F. (2007). A nonparametric online item calibration.
 - 3.3. Parshall, C. G. (2007). Designing templates for innovative item types.
 - 3.4. Eggen, T. (2007). Choices in CAT models in the context of educational testing.
 - 3.5. Luecht, R. M. (2007). Emerging topics.
 - 3.6. van der Linden, W. J. (2007). The shadow-test approach: A universal framework of implementing adaptive testing.

第七章、其他

再次感謝國科會的經費補助，才讓我能參加這次在美國明尼蘇達州舉行之學術研討會，與學者專家交換研究心得，謝謝！

附錄一、大會議程



2007 GMAC® Conference on Computerized Adaptive Testing
Radisson University Hotel—Minneapolis • Minneapolis, Minnesota

Thursday, June 7th, and Friday, June 8th, 2007

On behalf of the many individuals at GMAC® and the University of Minnesota who have worked hard to organize this meeting, we welcome you to the 2007 GMAC® *Conference on Computerized Adaptive Testing*. Computerized adaptive testing is one of the more exciting aspects of educational and psychological measurement. Here practitioners and theorists get to apply advanced mathematical models to deliver state-of-the-art assessments that are responsive to the needs of both the test developer and the test taker.

As you look through the program, we are sure you will be struck by the range of topics covered by the 10 keynote presentations and roundtables, 26 invited papers, and three posters. GMAC® and the University of Minnesota have been able to attract many leaders in the field of measurement, including many famous individuals who have laid the foundation for adaptive testing. As conference attendees, we will have an opportunity to reflect on actual computerized adaptive testing practices, learn more about leading programs, benefit from the hard learned lessons of our colleagues, and ponder the possibilities from leading edge research in the field.

We are certain you will also be struck by the international representation at this meeting. We extend a special welcome to our colleagues who have traveled from Australia, Brazil, Canada, Denmark, Hong Kong, Iceland, Korea, Malaysia, Portugal, Russia, South Africa, Spain, Taiwan, The Netherlands, the United Kingdom, and other locations around the world. Clearly, CAT is an important and fascinating topic globally, and we are fortunate that our international colleagues are able to share their work with us.

A special feature of this conference is the opportunity for roundtable exchanges with keynote presenters. The goal is to have rich, and possibly heated, discussions on key topics. Please come to these roundtables ready to share your experiences, discuss your own research, and present your own views. It is the program committee's hope that these discussions will be among the best parts of the meeting.

Again, welcome to the 2007 GMAC® *Conference on Computerized Adaptive Testing*. We sincerely hope this will be a stimulating, productive, and rewarding meeting for us all.

Lawrence M. Rudner, Graduate Management Admission Council®
David Weiss, University of Minnesota at Twin Cities

Featured Keynote Speakers



Theo Eggen is a member of CITO's Psychometric Research Center. He has extensive experience in advising on the methodological aspects of educational research and test development, conducting data analysis, and multidisciplinary cooperation projects. He has worked as a consultant in educational measurement at the university level, at CITO, and internationally. Additionally, Eggen teaches introductory and specialized courses and presents many papers at national and international conferences. He has authored research articles, syllabi, and textbook chapters. Eggen's specializations include IRT, national assessment, and CAT.

Presenting Thursday, June 7: Keynote Roundtables II (2:45-4:15 p.m.)

Choices in CAT Models in the Context of Educational Testing

CITO, the National Institute for Educational Measurement in the Netherlands, employs CATs in a number of educational testing programs. In the presentation, procedures for the calibration of item banks used in CATs and item selection procedures meeting practical constraints will be discussed.



Fanmin Guo is director of psychometric research at GMAC®, owner of the Computer-Adaptive GMAT® (GMAT CAT®) exam. Prior to GMAC®, Guo was a lead measurement statistician and the statistical analysis group leader for both the GMAT® and GRE® exams at ETS, where he worked for eight years (1997 to 2005). His research interests cover practical issues in applications of IRT and operations of CAT.

Before coming to the United States, Guo was an associate professor of English linguistics and vice-dean, faculty of basic studies, China University of Geosciences, Wuhan, China.

Presenting with Lawrence Rudner, GMAC®, Thursday, June 7:
Keynote Roundtables II (2:45-4:15 p.m.)

A Practitioner's Perspective on Computerized Adaptive Testing

There are numerous approaches for obtaining valid scores, especially in a computer adaptive testing environment. Some approaches look good in theory, but fail miserably when one considers the practical issues of costs, security, and program maintenance. This presentation examines alternatives for several key decisions including item selection and administration algorithms, item exposure control, content balancing, item types, test length, rotation of test items, length of testing windows, and test center capacity from a practitioner's perspective. Possible impacts of several alternative decisions will be identified, and suggestions for practice will be offered.

Featured Keynote Speakers



Charles Lewis is director of the psychometrics program and professor of psychology at Fordham University in New York. He also works part time at ETS, where he is a distinguished presidential appointee. Lewis has previously taught at Dartmouth College, the University of Illinois at Urbana-Champaign, and the University of Groningen in The Netherlands. He received his PhD in statistics at Princeton University with John Tukey. His areas of specialization include fairness and validity in educational testing; mental test theory, including IRT and computer-based testing; general(ized) linear models, including multiple comparisons and repeated measures; Bayesian inference, including multilevel modeling; and behavioral decision making.

Presenting Thursday, June 7: Keynote Roundtables I (9:10-11:10 a.m.)

Some Thoughts on Controlling Item Exposure in Adaptive Testing

A particular method for controlling item exposure in a CAT, developed by Martha Stocking and Charles Lewis, is briefly described and taken as a point of reference. The properties of this method of exposure control are considered and critically evaluated relative to the primary goal of adaptive testing: providing valid and reliable measurement with a minimal number of items. They are also evaluated relative to problems of test security that the method is intended to address. The point is made that methods of exposure control should be developed in the context of other test security measures, such as pool rotation, and not in isolation.



Richard Luecht is a professor in the ERM Department at the University of North Carolina at Greensboro, where he teaches graduate courses in applied statistics, measurement theory, and assessment design and technological applications. There he has held simultaneous appointments as the director of the Center for Educational Research and Evaluation and as chair of the ERM Department. Luecht was previously director for computerized adaptive testing research and senior psychometrician at the National Board of Medical Examiners. Luecht also designed software systems and algorithms for large-scale automated test assembly and devised the computer-adaptive multistage testing implementation framework that is used by a number of large-scale testing programs. He previously worked as a research psychometrician with ACT, Inc. He has published numerous research papers, articles, and book chapters on technical measurement issues

and is a technical advisor and consultant for many K-12 state testing agencies and large-scale testing organizations.

Presenting Thursday, June 7: CAT Innovations (5:05-5:35 p.m.)

Emerging Topics

This talk will introduce three future-oriented areas of research related to CBT and CAT. The first area covers multidimensional, adaptive test designs for diagnostic testing. The second area covers a multilevel design framework-amenable to a variety of CAT settings. This multilevel design framework is part of an emerging, integrated measurement discipline called Assessment Engineering. The third area entails data-design and related specifications for implementing adaptive constructed-response, computer-based performance simulations.

Featured Keynote Speakers



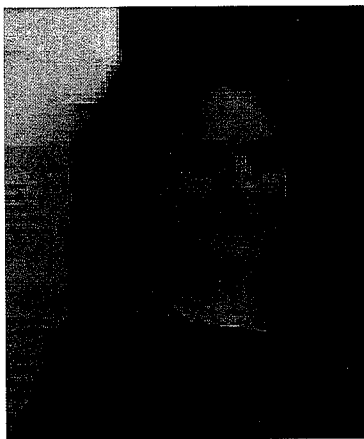
James Olsen is vice president/chief scientist of Alpine Testing Solutions. He has pioneered innovations for computerized, adaptive, e-learning, and performance testing. He has served the Association of Test Publishers (ATP) as certification/licensure chair and secretary, Board member, test standards member, and co-chair for guidelines for computer-based testing, and he participates annually as an invited speaker and program committee member with ATP's Innovations in Testing conference. He also gives technical guidance to GMAC®, the National Assessment Governing Board, the International Digital Communications Standards (as chair), and the Utah Department of Education.

Olsen is an International Test Commission member and Standards Review Board member for the Performance Testing Council, as well as a certification auditor for the American National Standards Institute and principal investigator for multiple computerized assessment grants by NSF and U.S. Department of Education. He has authored key publications on computerized, adaptive, and technology assessment and validity models.

Presenting with C. Victor Bunderson, Edumetrics Institute, Thursday, June 7: Keynote Roundtables II (2:45-4:15 p.m.)

Validity and Operational Considerations in Selecting a CAT Measurement Model

We emphasize the centrality of validity centered design/documentation in CAT systems. Validity centered design/documentation posits a tripartite validation model with facets for design for usability, appeal and positive expectations; design for inherent construct validity; and design for criterion-related validity. Each facet addresses validation argument evidence from three alternative sources. We call for using more complex, integrated, and work model tasks in CAT. We discuss operational issues for selecting a CAT measurement model and implementing CAT systems. Empirical performance test analyses are shown for I&2 parameter logistic IRT calibration, test information, and standard error curve analyses for practicum computer networking certification.



Cynthia Parshall holds a PhD in educational measurement and research and an MEd in instructional computing. Her professional experience includes work as a psychometrician, researcher, and faculty member with ACT, Inc., the University of South Florida, and employment as a measurement consultant. Parshall's support to computer-based testing projects has included the development of CBT-specific usability methods, screen design criteria, and software evaluation guidelines. She has incorporated this focus on usability and interface design into a range of innovative item type development and research projects. In addition, Parshall has conducted research and development work on such CBT topics as: the use of audio, language testing, international test applications, Web-based test design, and the comparability of one test administration mode with another. Parshall is lead author of *Practical Considerations in Computer-Based Testing*.

Presenting Thursday, June 7: Keynote Roundtables I (9:10-11:10 a.m.)

Designing Templates for Innovative Item Types

This presentation will include a discussion of a new taxonomy for innovative item types encompassing: assessment structure, complexity, fidelity, interactivity, media inclusion, response action, and scoring model. The taxonomy levels will be illustrated through the use of operational innovative items. Furthermore, an approach to using these taxonomy levels in the design and development of item templates will be presented. The discussion of item templates will consider implications of the taxonomy levels for test development issues such as test content, psychometrics, programming requirements, examinee computer skill, and cost.

Featured Keynote Speakers



Mark Reckase has been a professor in the College of Education at Michigan State University since August of 1998. He has taught IRT, structural equation modeling, applied and theoretical educational measurement, and basic research methodology, and has worked on numerous projects to improve the quality of applied measurement in teacher training, school accountability, literacy, support of classroom instruction, medical certification, statewide assessment, and national testing programs.

Prior to August 1998, Reckase was the assistant vice president of the assessment innovations area in the development division of ACT, Inc. In that role, he was responsible for the development of new assessment methodology that makes use of advances in technology, education, and cognitive psychology, focusing on the use of non-multiple-choice and computerized methods for assessment. Additionally, staff in the area

conducted research on psychometric topics related to standard setting, performance assessment, and computerized testing.

Presenting Thursday, June 7: Keynote Roundtables I (9:10-11:10 a.m.)

Designing Item Pools to Optimize the Functioning of a CAT

CAT procedures will match their theoretical expectations only if appropriate items are available for selection from an item pool. This paper will describe a process for designing item pools for CATs that will allow their theoretical properties to be attained. The procedures can be used to determine the size and distribution of item characteristics needed for a particular CAT application. Examples of the procedure are provided to show the relationship of item pool design to examinee distributions, exposure control, and details of the CAT procedure.



Lawrence M. Rudner is vice president of research and development at the Graduate Management Admission Council® (GMAC®). His work at GMAC® has included test validation, adaptive testing, professional standards, QTI specifications, test security, data forensics, and contract monitoring for the Graduate Management Admission Test® (GMAT®) exam. He conducted some of the first research on several lasting measurement topics, including the use of IRT to assess item bias, parameter invariance, assessment of person fit, validity of a composite measure, and classification accuracy.

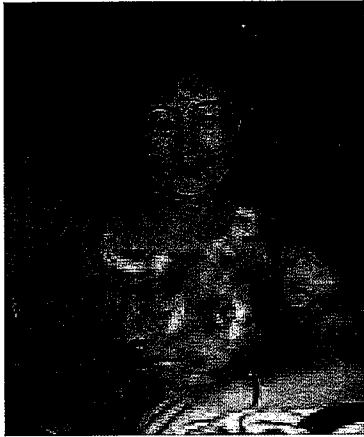
Prior to joining GMAC®, Rudner was director of the ERIC Clearinghouse on Assessment and Evaluation. He is also the founder and co-editor of the online journal Practical Assessment Research and Evaluation, which is now the most widely read journal in the field.

Presenting with Fanmin Guo, GMAC®, Thursday, June 7: Keynote Roundtables II (2:45-4:15 p.m.)

A Practitioner's Perspective on Computerized Adaptive Testing

There are numerous approaches for obtaining valid scores, especially in a computer adaptive testing environment. Some approaches look good in theory, but fail miserably when one considers the practical issues of costs, security, and program maintenance. This presentation examines alternatives for several key decisions including item selection and administration algorithms, item exposure control, content balancing, item types, test length, rotation of test items, length of testing windows, and test center capacity from a practitioner's perspective. Possible impacts of several alternative decisions will be identified, and suggestions for practice will be offered.

Featured Keynote Speakers



Fumiko Samejima is a professor at the University of Tennessee at Knoxville. She began research on item response theory as a PhD student at Keio University when she published a book titled LIS Measurement Scale of Non-verbal Reasoning Ability (in Japanese) with her major professor, Dr. Tarow Indow in 1962.

She subsequently served as visiting research psychologist with ETS and was invited to the Psychometric Laboratory of the University of North Carolina, Chapel Hill before teaching at the University of New Brunswick, Fredericton, NB, Canada. From that research she wrote two popular monographs, "Estimation of latent ability using a response pattern of graded scores" and "A general model for free-response data" in *Psychometrika* Monograph. Samejima has since published on a variety of topics, including proposed continuous response models in the unidimensional and

multidimensional latent space. She has been awarded several multi-year research projects related to IRT and CAT for the Office of Naval Research and the Law School Admission Council, is a past president of the Psychometric Society, and received the 1991 technical contribution award from NCME.

Presenting Thursday, June 7: Keynote Roundtables I (9:10-11:10 a.m.)

A Nonparametric Online Item Calibration

In IRT, development of methods of estimating the operating characteristic (OPC; conditional probability, given the ability level) of each specific item response to the test question is very important. Among such methods, nonparametric estimation methods exceed parametric methods in principle, because it will make us discover the OPC without molding it into a particular mathematical form. The author has been engaged in the development of such nonparametric methods since 1977, and has proposed several methods. The presentation introduces one method, the conditional pdf approach, which has been adjusted to online item calibration in computerized adaptive testing.



Wim J van der Linden is professor of measurement and data analysis, faculty of behavioral sciences, University of Twente, The Netherlands. His research interests include test theory, CAT, optimal test assembly, test equating, modeling response times on test items, and decision theory. His publications have appeared in all major international journals and he is co-editor of *Handbook of Modern Item Response Theory* (New York: Springer, 1997; with R. K. Hambleton) and *Computerized Adaptive Testing: Theory and Applications* (Boston: Kluwer, 2001; with C.A.W. Glas). His latest book is *Linear Models for Optimal Test Design* (Springer, 2005).

Van der Linden has served on the editorial boards of several journals and is a co-editor for the Springer series *Statistics for Social and Behavioral Sciences*. He is also former president of the Psychometric Society, fellow of the Center for Advanced Study in

the Behavioral Sciences, Stanford, CA, and recipient of the ATP and NCME Career Achievement Awards for his work in educational measurement.

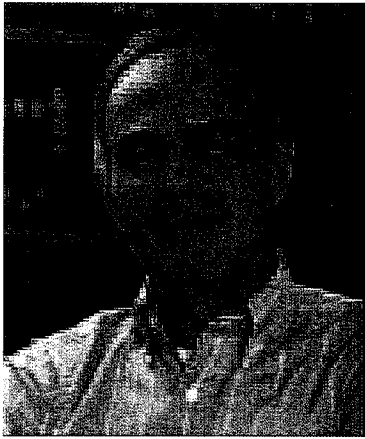
Presenting Friday, June 8: CAT Models (8:00-8:30 a.m.)

The Shadow-Test Approach: A Universal Framework for Implementing Adaptive Testing

In hindsight, early research on adaptive testing was too narrowly focused on the statistical features of adaptive testing, mainly addressing such topics as the efficiency of adaptive testing relative to traditional paper-and-pencil testing, approximate ability estimation methods for use in real-time adaptive testing, and the performances of different item-

Featured Keynote Speakers

selection criteria. However, the first implementations of large-scale adaptive testing programs revealed a multitude of more practical problems related to, for example, the realization of large sets of content specifications, prevention of item compromise, item-pool design, differential speededness, and the equating of scores on adaptive tests to a released linear form for score reporting. We will present the shadow-test approach as a universal framework for solving such problems and illustrate the approach with several empirical examples.



David J. Weiss joined the psychology department at the University of Minnesota as an assistant professor in 1967 and has been the director of its psychometric methods program since 1970. Weiss is internationally recognized as an expert in the field of psychometrics. He was elected fellow of the American Psychological Association for the first time in 1973 and is currently a fellow in both Division 5 (Measurement, Evaluation, and Statistics) and Division 17 (Counseling Psychology). Weiss was the founding editor of *Applied Psychological Measurement* from 1976-2001, one of the most important journals in the measurement field, which focuses on ways to use the most current techniques to address measurement problems in the behavioral and social sciences. He has been actively involved in assessment and testing research for more than 40 years.

Weiss is a long-time proponent of computerized assessment. His primary interest, CAT, began in the early 1970s with research conducted through a small grant from the Graduate School of the University of Minnesota, where he supervised the development of a rudimentary CAT-delivery system on a university mainframe computer. Since that time, he has been closely involved in the design and implementation of at least seven CAT software systems, several versions of CAT research software supported by the U.S. Office of Naval Research (and other agencies of the U.S. Department of Defense), three versions of the MicroCAT Testing System, and most recently the FastTEST Professional Testing System.

Weiss's interest in computer-based testing goes beyond the traditional teaching and advising role for graduate students. He is the co-founder of two influential assessment-related companies. First, he co-founded Assessment Systems Corporation with C. David Vale to create and market software related to computerized testing, particularly CAT. Next, he co-founded (again with C. David Vale) the Insurance Testing Corporation (ITC), which was subsequently sold to The Chauncey Group of Educational Testing Service. Weiss's current projects include the creation and maintenance of CAT Central (<http://www.psych.umn.edu/psylabs/CATCentral>), a global CAT resource Web site, and the continued pursuit of leading-edge research in CAT.

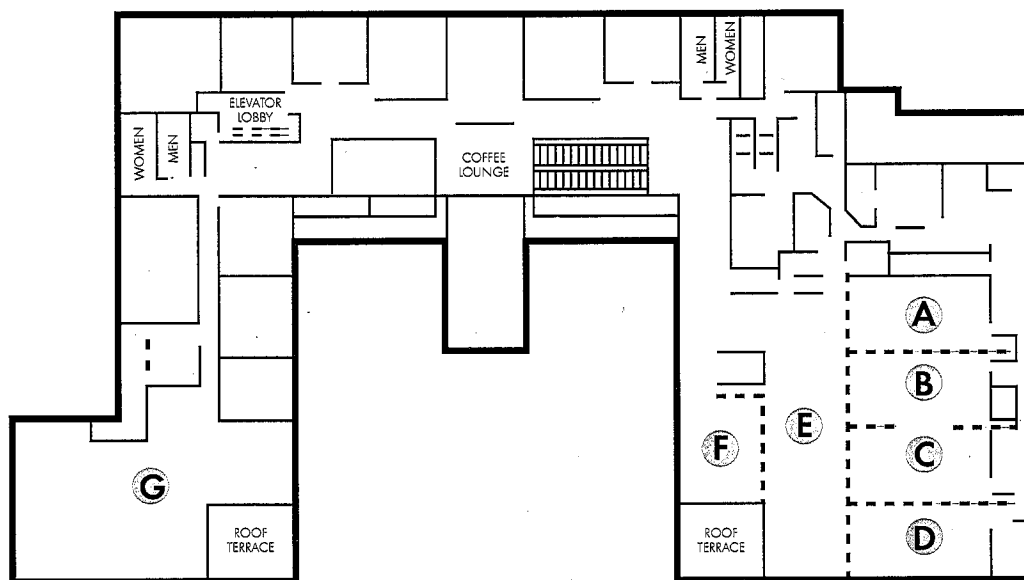
Presenting Thursday, June 7: Keynote Address (8:30-9:00 a.m.)

Computerized Adaptive Testing: Past, Present, and Future

Early developments that have led to the current state of CAT will be reviewed and some issues that were addressed in early research that might require future attention will be identified. The current state of CAT research will be contrasted with early developments. Some emerging issues will also be identified that might affect the future implementation and progress of CAT.

Radisson University Hotel, Second Floor

Radisson University Hotel—Minneapolis • 615 Washington Avenue S.E. • Minneapolis, Minnesota



- A** University Ballroom Section A
- B** University Ballroom Section B
- C** University Ballroom Section C
- D** University Ballroom Section D
- E** Prefunction Area
- F** Faculty Room
- G** Humphrey Ballroom

WEDNESDAY, JUNE 6

8:00 a.m.–7:00 p.m. Registration

Prefunction Area

PRE-CONFERENCE SESSIONS (Separate registration required.)

9:00 a.m.–12:15 p.m. **Introduction to IRT**

*University Ballroom,
Sections C & D*

12:15–1:45 p.m. Lunch

*University Ballroom,
Sections A & B*

1:45–5:00 p.m. **Introduction to CAT**

*University Ballroom,
Sections C & D*

THURSDAY, JUNE 7

7:00–8:30 a.m. Breakfast

*University Ballroom,
Sections C & D*

8:00 a.m.–5:00 p.m. Registration

Prefunction Area

8:30–9:00 a.m.

Computerized Adaptive Testing: Past, Present, and Future

Introduction: *Lawrence M. Rudner, Graduate Management Admission Council®*

**WELCOME & KEYNOTE
ADDRESS**

Keynote: *David J. Weiss, University of Minnesota at Twin Cities*

*University Ballroom,
Sections A & B*

9:00–9:10 a.m. Break

Prefunction Area

9:10–11:10 a.m.

CAT Items: From Pool Design to Item Administration (30 minutes each)

Chair: *Lawrence M. Rudner, Graduate Management Admission Council®*

**KEYNOTE
ROUNDTABLES, I-A**

Mark D. Reckase, Michigan State University

Designing Item Pools to Optimize the Functioning of a CAT

Charles Lewis, Fordham University

Some Thoughts on Controlling Item Exposure in Adaptive Testing

Fumiko Samejima, University of Tennessee at Knoxville

A Nonparametric Online Item Calibration

Cynthia G. Parshall, Measurement Consultant

Designing Templates for Innovative Item Types

Agenda-at-a-Glance

THURSDAY, JUNE 7

11:10–11:20 a.m. Break / Assembly into roundtables

Prefunction Area

11:20 a.m.–12:20 p.m. **CAT Items Roundtables** (two 30-minute sessions)

**KEYNOTE
ROUNDTABLES, 1-B**

Keynote speakers will head roundtable discussions about the preceding CAT Items presentations. Participants will switch tables between sessions.

*University Ballroom,
Sections A, B, & Faculty*

12:20–1:20 p.m. Lunch

*University Ballroom,
Sections C & D*

1:20–2:35 p.m.

Item Exposure

Chair: Charles Lewis, Fordham University

**CONCURRENT PAPER
SESSION I**

Ya-Hui Su, National Academy for Educational Research & Wen-Chung Wang, National Chung Cheng University
Simultaneous Online Control Over Item Exposure and Test Overlap in Computerized Adaptive Testing for Independent and Testlet-Based Items

*University Ballroom,
Section A*

Michael Edwards, The Ohio State University & David Thissen, The University of North Carolina at Chapel Hill
Multi-Stage Computerized Adaptive Testing with Uniform Item Exposure

Thomas O'Neill & Weiwei Liu, National Council of State Boards of Nursing
Detecting Item Compromise Using Item Latency Residuals and Response Probabilities

Lixiong Gu, Educational Testing Service & Mark D. Reckase, Michigan State University
Designing Optimal Item Pools for Computerized Adaptive Tests with Symptom-Hetter Exposure Control

1:20–2:35 p.m.

CAT and Cognitive Structure

Chair: Lawrence M. Rudner, Graduate Management Admission Council®

**CONCURRENT PAPER
SESSION II**

Jiawen Zhou & Mark J. Gierl, University of Alberta
Computerized Attribute-Adaptive Testing: A New Computerized Adaptive Testing Approach Incorporating Cognitive Psychology

*University Ballroom,
Section B*

G. Gage Kingsbury & Ronald L. Houser, Northwest Evaluation Association
ICAT: A Process for Item Selection in Adaptive Tests to Allow the Identification of Idiosyncratic Knowledge Patterns

Jean-Guy Blais, Université de Montréal & Michel Desmarais, École Polytechnique de Montréal
Partial Order Knowledge Structure for CAT Applications

Ying Cheng & Hua-hua Chang, University of Illinois at Urbana-Champaign
The Maximum Dual Information Method for Cognitive Diagnostic CAT

THURSDAY, JUNE 7

<p>1:20–2:35 p.m.</p> <p>CONCURRENT PAPER SESSION III</p> <p><i>Faculty Room</i></p>	<p>Item Calibration and Special Applications Chair: <i>David J. Weiss, University of Minnesota at Twin Cities</i></p> <p><i>Gyenam Kim Kang, Korea Nazarene University & David J. Weiss, University of Minnesota at Twin Cities</i> Comparison of Computerized Adaptive Testing and Classical Methods for Measuring Individual Change</p> <p><i>Rongchun Zhu, ACT, Inc, Jeffrey Douglas, University of Illinois at Urbana-Champaign, & Hua-Hua Chang, University of Illinois at Urbana-Champaign</i> Implementation of Optimal Design for Item Calibration in CAT</p> <p><i>Nathan A. Thompson & Shungwon Ro, Prometric</i> Computerized Classification Testing with Composite Hypotheses</p>
<p>2:35–2:45 p.m.</p> <p><i>Prefunction Area</i></p>	<p>Break</p>
<p>2:45–4:15 p.m.</p> <p>KEYNOTE ROUNDTABLES, II-A</p> <p><i>University Ballroom, Sections C & D</i></p>	<p>CAT Models and Monitoring (30 minutes each) Introduction: <i>Kara M. Owens, Graduate Management Admission Council®</i></p> <p><i>James Olsen, Alpine Testing Solutions & C. Victor Bunderson, Edumetrics Institute</i> Validity and Operational Considerations in Selecting a CAT Measurement Model</p> <p><i>Lawrence M. Rudner & Fanmin Guo, Graduate Management Admission Council®</i> A Practitioner's Perspective on Computerized Adaptive Testing</p> <p><i>Theo Eggen, CITO</i> Choices in CAT Models in the Context of Educational Testing</p>
<p>4:15–4:25 p.m.</p> <p><i>Prefunction Area</i></p>	<p>Break / Assembly into roundtables</p>
<p>4:25–4:55 p.m.</p> <p>KEYNOTE ROUNDTABLES, II-B</p> <p><i>University Ballroom, Sections A, B, C, & D</i></p>	<p>CAT Models and Monitoring Roundtables (one 30-minute session)</p> <p>Keynote speakers will head roundtable discussions about the preceding CAT Models and Monitoring presentations.</p>
<p>4:55–5:05 p.m.</p> <p><i>Prefunction Area</i></p>	<p>Break</p>
<p>5:05–5:35 p.m.</p> <p>KEYNOTE</p> <p><i>University Ballroom, Sections C & D</i></p>	<p>CAT Innovations Introduction: <i>Lawrence M. Rudner, Graduate Management Admission Council®</i></p> <p><i>Richard M. Luecht, University of North Carolina at Greensboro</i> Emerging Topics</p>
<p>6:30–8:30 p.m.</p> <p><i>Humphrey Ballroom</i></p>	<p>Reception</p>

Agenda-at-a-Glance

FRIDAY, JUNE 8

6:00–8:00 a.m.

Breakfast

University Ballroom,
Sections C & D

8:00–8:30 a.m.

CAT Models

Introduction: Lawrence M. Rudner, Graduate Management Admission Council®

KEYNOTE

University Ballroom,
Sections A & B

Wim J. van der Linden, University of Twente, The Netherlands
The Shadow-Test Approach: A Universal Framework for Implementing Adaptive Testing

8:30–8:40 a.m.

Break

Prefunction Area

8:40–9:55 a.m.

New CAT Models

Chair: David J. Weiss, University of Minnesota at Twin Cities

CONCURRENT PAPER SESSION IV

University Ballroom,
Sections A & B

David J. Weiss, University of Minnesota at Twin Cities & Robert D. Gibbons, University of Illinois at Chicago
CAT with the Bifactor Model

Jason Immekus, University of Illinois at Chicago, Robert D. Gibbons, University of Illinois at Chicago, & A. John Rush, University of Texas Southwestern Medical Center at Dallas
Patient Reported Outcomes Measurement and Computerized Adaptive Testing: An Application of a Post-Hoc Simulation of a Diagnostic Screening Instrument

Stephen Stark, University of South Florida & Oleksandr S. Chernyshenko, University of Canterbury, New Zealand
Adaptive Testing with the Multi-Unidimensional Pairwise Preference Model (MUPP)

Kathleen Scalise, University of Oregon & Mark Wilson, UC Berkeley
Bundle Models for Computer Adaptive Testing in E-Learning Assessment

8:40–9:55 a.m.

Applications and Issues

Chair: James Olsen, Alpine Testing Solutions

CONCURRENT PAPER SESSION V

University Ballroom,
Section C

Tony D. Thompson & Walter D. Way, Pearson Educational Measurement
Investigating CAT Designs to Achieve Comparability with a Paper Test

Richard C. Gershon, Center for Outcomes and Research, Northwestern University
Individual Differences in CAT

Jooyong Park, Sejong University & UC Berkeley
A New Delivery System for CAT

Walter D. Way, Scott Davies, & Kelly Burling, Pearson Educational Measurement
Adapting to Assessment Policy: CAT Applications for Statewide K-12 Assessments

FRIDAY, JUNE 8

8:40–9:55 a.m.

IRT Estimation and Polytomous CAT

Chair: Fanmin Guo, *Graduate Management Admission Council®*

CONCURRENT PAPER
SESSION VI

*University Ballroom,
Section D*

Po Hsi Chen, National Taiwan Normal University

The Resolution of Regression Bias of the Bayesian Ability Estimation on Unidimensional and Multidimensional Computerized Adaptive Testing

Yanyan Sheng, Southern Illinois University at Carbondale, Nancy Flournoy, University of Missouri at Columbia, & Steven J. Osterlind, University of Missouri at Columbia
Up-and-Down Procedures for Approximating Optimal Test Designs Using Person-Response Functions

Jean-Guy Blais Université de Montréal, Gilles Raiche, Université du Québec à Montréal, & David Magis, Université de Liège
Adaptive Estimators of Proficiency in Adaptive Testing

Michiel Hol, H. C. M. Vorst, & G. J. Mellenbergh, NOA/Free University
Computerized Adaptive Testing for Polytomous Motivation Items: Administration Mode Effects and a Comparison with Short Forms

9:55–10:05 a.m.

Break

Prefunction Area

10:05–11:20 a.m.

Posters

POSTER SESSION

Prefunction Area

Bernard P. Veldkamp, Iris J. L. Egberink, & Rob R. Meijer, University of Twente, The Netherlands
The Development of a Computer Adaptive Integrity Test

Mick Sumbling, Pablo Sanz, M. Carme Viladrich, Eduardo Doval, & Laura Riera, Universitat Autònoma de Barcelona
Development of a Multiple-Component CAT Measuring Foreign Language Proficiency (SIMTEST)

Marié De Beer, University of South Africa Institution
Use of CAT in Dynamic Testing

10:05–11:20 a.m.

Major CAT Programs

Chair: David J. Weiss, *University of Minnesota at Twin Cities*

*University Ballroom,
Sections A & B*

G. Gage Kingsbury, Northwest Evaluation Association
CAT in the K-12 Schools: 50 Million CATs and Counting

Richard C. Gershon, Jin Shei Lai, Seung Choi Center for Outcomes and Research, Northwestern University, & Jakob Bjorner, Quality Metric Inc.
The Promise of PROMIS for Using CAT in Measuring Health Treatment Outcomes

James R. McBride & Joseph E. Betts, Renaissance Learning, Inc.
Eleven Years of Assessing K-12 Achievement: Some Longitudinal Research Using STAR Reading, STAR Math, and STAR Early Literacy

Agenda-at-a-Glance

FRIDAY, JUNE 8

11:20–11:30 a.m.

Break

Prefunction Area

11:30 a.m.–12:00 p.m.

Final Closing

*University Ballroom,
Sections A & B*

*Lawrence M. Rudner, Graduate Management Admission Council®
David J. Weiss, University of Minnesota at Twin Cities*

**Graduate
Management
Admission
Council®**

Creating Access to Graduate Business Education®

1600 Tysons Boulevard
Suite 1400
McLean, VA 22102
USA
Phone 1-703-749-0131
Fax 1-703-749-0169
gmacmail@gmac.com
www.gmac.com
www.mba.com

Copyright © 2007 Graduate Management Admission Council® (GMAC). All rights reserved.

The Graduate Management Admission Council® is the international, not-for-profit association behind the Graduate Management Admission Test® (GMAT®) used by 220,000 prospective MBA students and about 4,000 programs at 1,700 business schools worldwide.

Creating Access to Graduate Business Education®, GMAC®, GMAT®, Graduate Management Admission Council®, and Graduate Management Admission Test® are registered trademarks of the Graduate Management Admission Council® in the United States and other countries.

GMAC® does not endorse the views of and is not affiliated with any of the speakers of this program, other than those specifically identified as representatives of the Graduate Management Admission Council®.

附錄二、GMAC 技術報告

Use of the GMAT[®] Analytical Writing Assessment: Past and Present

Kara M. Owens

GMAC[®] Research Reports • RR-07-01 • October 7, 2006

Abstract

The recent addition of writing requirements to several admission testing programs implies that higher education institutions believe these skills are essential for college success. To assess whether this interest is true for management education, a survey was conducted to evaluate attitudes toward writing assessments among programs that use the Graduate Management Admission Test[®] (GMAT[®]). This study compared current uses and usefulness of the AWA to advantages originally anticipated in 1993 and previously reported in 1998. Results from 109 respondents indicated that the section was used and found effective for selection and writing deficiency diagnosis. Additionally, programs with high concentrations of non-native English speaking applicants found the section to be more effective than those with lower concentration.

During the past few decades, writing assessments have gained both positive and negative attention from practitioners in need of evaluation strategies beyond standard multiple-choice question formats (Bridgeman & Carlson, 1984; Quellmalz, 1984). There has also been increased interest in the inclusion of writing as a part of admissions testing (Powers & Fowles, 2002), and the SAT and the ACT have recently added writing components. However, other admissions tests have more mature writing assessments, and the Analytical Writing Assessment (AWA) has been a component of the Graduate Management Admission Test[®] (GMAT[®]) for over a decade.

The purposes of the present study were to add to the previous research and reevaluate current uses and usefulness of the AWA for graduate business programs. This included a general evaluation of use and usefulness of the section for several different purposes, including those originally anticipated. The current study also explored differences in uses and usefulness for the AWA section for different subgroups of respondents. Responses were examined separately for different program types and for programs with different concentrations of non-native English speaking applicants. Finally, to determine if the section was still meeting its original purposes of admission

selection and writing deficiency diagnosis, results from the current study were compared to previous research findings on anticipated and early use of the AWA.

This study provides an overview of the past and present AWA. The results of the current survey are presented and different uses for the section are evaluated to determine how the section may be meeting various program needs. Additionally, the inspection of subgroup differences in terms of use and usefulness of the AWA provides information about the needs in different types of programs.

History of the GMAT[®] Analytical Writing Assessment

Perceptions of Analytical Writing

More than 20 years ago, the Graduate Management Admission Council[®] (GMAC[®]) embarked on a journey to determine if analytical writing ability, as measured by a writing task, was a skill deemed necessary to be successful in graduate management education (Bruce, 1984; Bruce, 1992; Bruce, 1993). As a part of this process, exploratory research and surveys were conducted to evaluate institutional needs and potential use of an analytical writing section as an addition to the GMAT[®] exam.

A 1983 survey of management education programs revealed positive support for the concept of a writing assessment. Of the 355 respondents, 88% strongly or moderately approved (Bruce, 1984). The vast majority believed that it would be very or somewhat useful in selecting applicants for admission and determining if students needed additional work on this area, 85% and 88%, respectively. In addition, 87% indicated that they would strongly or moderately encourage the inclusion of a writing assessment. Those who discouraged the inclusion cited their reasons as satisfaction with their current evaluation system, doubt regarding the reliability and validity of the AWA score, and apprehension over increased cost and time commitments required of applicants. Though the overall results were encouraging, there was some trepidation, particularly from top-tier graduate management programs, regarding the usefulness of a writing section.

As a follow-up to the previous survey, telephone interviews were conducted in 1991 to gain specific information on the writing assessment concept and potential design (Bruce, 1992). Respondents were asked questions about attitudes toward essay scoring, issues of pricing, and level of interest in a writing component. Inquiries were made to determine if schools would like GMAC® to score the essay(s) and provide programs with copies of examinee-written essays. Most of the interviewees stated that they would prefer to have both a score for the writing assessment and a copy of the examinee-written essay(s). They believed the combination of both would provide them with a way to personally validate scores using their own criteria, while still providing a consistent, objective score based on trained readers. In addition, when asked if the new writing section should be optional, the majority of interviewees believed that requiring a scored writing section for all examinees taking the GMAT® exam would be the most consistent and fair approach. Ultimately, respondents felt the quality of the writing assessment and the validity of the scores it produced would determine whether or not it would be useful and valued by schools. However, reservations remained regarding the increased cost to the examinee and the adverse impact this cost might have on applications to business school.

In 1993, a final survey was conducted to determine potential uses and attitudes surrounding the addition of a

writing assessment to the GMAT® exam (Bruce, 1993). This proposed assessment would include one or two analytical writing tasks, which would be holistically scored. The essay(s) and the holistic score(s) would be provided to the institution and test takers along with the traditional GMAT® exam Verbal, Quantitative, and Total scores. The results of this survey revealed that respondents, mostly deans and directors from graduate management programs currently using the GMAT® exam, approved of and encouraged the addition of an AWA section. Moreover, they believed it would be useful in diagnosing student deficiencies and selecting students for admission. Respondents further stated that "effective writing skills are needed in order to succeed in both graduate school and business" (p. 32).

In 1994, after years of research that revealed favorable recommendations and encouragement, the AWA was added to the GMAT® exam with the expectation that it would assist in admission selection and diagnosis of writing deficiencies for applicants to graduate management education programs (GMAC®, n.d.). During the AWA, examinees are asked to respond to two questions, each using an essay format. One essay provides an analysis of an issue and the examinee's personal views on a provided topic. The second essay requires examinees to critique a provided argument. Each examinee receives one holistic score based on responses to both questions (GMAC®, n.d.). Both the AWA score and the essays are reported for each examinee.

Actual Uses of the AWA

Several years after the implementation of the GMAT® AWA section, research was conducted to evaluate actual use of the AWA in comparison with the original stated purposes and perceived uses of the section (Noll & Stowers, 1998). The results from 59 respondents revealed that the AWA was used for admissions decisions, but was not as helpful at placing students in appropriate writing courses. Though 86% of respondents said that AWA scores were used to aid admissions decisions, diagnostic uses and usefulness of the scores were less clear. Fewer than 10% of respondents specified that scores were used for placing students in writing development courses or to waive communication course requirements. This was unexpected given the original purposes of the AWA and positive perceptions potential users provided in the first survey

(Bruce, 1993). However, 37% said that they used the section to determine admission if the program was previously uncertain about the applicant's English language skills or believed that these skills could be potentially problematic.

In 2002, Bruce examined satisfaction with the usefulness of the different sections of the GMAT® exam for 288 respondents. The results revealed that, among the four GMAT® exam scores that schools are provided—GMAT® Verbal, GMAT® Quantitative, GMAT® Total, and GMAT® AWA—respondents were least satisfied with the usefulness of the analytical writing scores.

Additionally, a portion of the respondents indicated that they used the AWA section for other purposes. Of this subset, 42%, or 54 respondents, reported that they often use the AWA section to determine if English language deficiencies might require the applicant to do additional work in this area, which was not one of the original perceived uses of the section. Overall, research has suggested that the section provides, "useful information that is not currently available", but that "the AWA only partially meets the expectations of management education" (Rogers & Rymer, 1995, p. 361). Ultimately, respondents were using the AWA for admissions decisions, but not for diagnosing writing deficiencies.

Methods

During a two-week period in 2005, GMAC® conducted a survey of graduate business school usage of AWA scores and essays. A link to the survey was electronically mailed to 417 graduate program faculty and administrators, along with three follow-up reminder e-mails. This yielded a final response rate of 26% (n = 109). The sample of respondents was mostly program directors, admission directors, and assistant deans. In total, respondents represented 104 different institutions, including 15 non-U.S. schools. Program enrollment for these different programs for the 2005 school year ranged between 40 and 1,800 students. The sample of programs that responded to the current study was comparable to those used in previous studies, and the programs represented a range of program types, sizes, and selectivity.

The current survey served several purposes. First, the general findings regarding reported use and usefulness of AWA essays and scores were examined across a variety of potential

uses. This provided information about the use of the section for its original purposes. Moreover, information on the most frequent uses and usefulness of the section for a variety of admission needs were presented. Secondly, responses based on subgroup membership were examined. Specifically, respondents were categorized based on the percentage of non-native English speaking applicants they typically receive as well as the type of program (i.e., full-time, part-time, executive, doctoral) their responses represented. This component of the study provides information about variations in use and usefulness depending on a program's needs. Finally, current use and usefulness of the AWA for its original purposes were compared with anticipated and previously reported use and usefulness. This analysis provides evidence to determine if AWA essays and scores are being frequently and effectively used for admission selection and writing deficiency diagnosis. The specific methods used to evaluate the data for these three purposes are described here.

Evaluating AWA Use and Usefulness for the Current Study

Many of the survey questions required respondents to indicate their frequency of use (always, frequently, sometimes, rarely, and never) and perceived usefulness (extremely useful, very useful, somewhat useful, not very useful, and not at all useful) of AWA scores and essays separately. Responses of "never" or "not at all useful" were assigned a score of "1". Response scores increased incrementally up to a score of "5" for a response of "always" or "extremely useful," as shown in Table I. This scoring system was used to provide information on average responses provided for analyses. However, the majority of the results presented focus on the percentage of respondents selecting specific response options. These percentages provide information on how many respondents used the section and found the section useful for a variety of purposes. As a result, the general response trend of the sample can be determined for each of the purposes examined.

For usefulness questions, respondents were given the option to select "not applicable" instead of indicating the degree of usefulness of the AWA for a specified purpose. Responses of "not applicable" were essentially treated as missing data. Because respondents did not have an opinion regarding the usefulness of the specified purpose, numerical scores were not assigned to "not applicable" responses. Sample sizes for these questions were smaller as a result.

Table I. Response Scores and Meaning

Score	Response	
	Use	Usefulness
1	Never	Not at all useful
2	Rarely	Not very useful
3	Sometime	Somewhat useful
4	Frequently	Very useful
5	Always	Extremely useful

Comparing Current AWA Use and Usefulness across Subgroups

To determine if AWA use and usefulness varied depending on the type of program using the section, respondents were separated into subgroups. First, programs were asked to estimate the percentage of applicants who were non-native English speakers. Programs were then categorized into those with an applicant pool of 25% or less versus those with greater than 25% non-native English speakers. Uses and usefulness for a variety of purposes were compared for the two groups.

Additionally, respondents were asked to select all program types (i.e., full-time, part-time, executive, doctoral) for which their responses were representative. If a respondent selected multiple program types, they were not asked to respond to separate surveys for each program type, rather their responses were represented in the results for each program type they selected. As such, one respondent's opinions may be represented up to four times, once for each program type the respondent selected. The resulting survey responses were representative of: 78 full-time, 61 part-time, 19 executive, and 11 doctoral business programs. Findings for doctoral and executive programs should be cautiously interpreted, as the sample size for these groups are limited and may not be representative of all programs within these specified types. The subgroup analyses allow for comparison of use and usefulness for programs with different needs and purposes.

Comparing Current AWA Use and Usefulness to Previous Findings

In previous research examining the anticipated (Bruce, 1984, 1992, 1993) and actual (Noll & Stowers, 1998)

use and usefulness of the AWA section, the response option formats for survey questions varied across studies. Bruce (1993) used three different response option sets to gauge perceived usefulness, approval, and encouragement for the addition of the AWA section. Specifically relevant to this study was how responses were collected regarding perceived usefulness of the section. Bruce used a four-point response scale including these options: very useful, somewhat useful, not very useful, and not at all useful. On the other hand, Noll and Stowers used a two-point scale to collect responses about usage of the AWA section, yes and no. As mentioned previously, the current study evaluated responses using five-point scales to measure frequency of use (always, frequently, sometimes, rarely, and never) and usefulness (extremely useful, very useful, somewhat useful, not very useful, and not at all useful).

To allow for comparisons between the previous and current research, despite the noted differences in option choices, the percentage of respondents selecting specific options were combined across several response categories. For instance, Bruce (1993) combined the responses very useful and somewhat useful to provide the percentage of respondents who perceived the addition of the AWA to be useful. Similarly, Noll and Stowers response choices only allowed for a dichotomous division of respondents into those who used the AWA section and those who reported that they did not use the section. Following this pattern, the present study combined responses from categories of always, frequently, or sometimes to identify the percentage of respondents who reported using scores or essays. Likewise, responses of extremely useful, very useful, somewhat useful were combined to indicate the percentage of respondents who found scores or essays to be useful. Though percentages combined in this manner

are reported in the text to compare previous and current results, distribution of all response options are also provided for a complete description of the current study findings.

Additionally, the present study examined use and usefulness separately for each component of the AWA section, essays and scores. While this separation provides more detailed information about exactly which component of the AWA section is used more often and effectively, it is difficult to compare the current results to previous section-level findings. As such, the results of the present study are presented using three different methods to allow for a complete comparison of current and previous results.

Specifically, results are presented separately for AWA essays and scores, as well as combined into a category called "either." The inclusion of the either category to evaluate frequency of use compares respondents who indicated that either essays or scores were always, frequently, or sometimes used for a specified purpose versus the percentage who indicated that they rarely or never used scores and essays for that purpose. Similarly, the either category utilized to rate usefulness compares the percentage of respondents who indicated that either essays or scores were extremely useful, very useful, or somewhat useful for a particular purpose versus the percentage who indicated essays and scores were not very useful or not at all useful for this same purpose. The either category therefore compares two groups of respondents, those who found AWA scores or essays to be used or useful for a particular purpose versus those who found that neither the AWA score nor the essays were used or useful. The

multiple groupings of responses allow for various comparisons between the previous research and the current study.

Results

Evaluating AWA Use and Usefulness for the Current Study

Nine potential uses of the AWA sections were identified through a review of the literature: selecting applicants for admission, granting assistantships and scholarships, evaluating applicant English grammar skills, validating essays written as a part of the program application process, diagnosing individual applicant writing deficiencies, planning programs and designing course curriculum, placing applicants in courses, requiring pre-enrollment training, and advising on a career path. Though the original purposes of the AWA section were for applicant selection and deficiency diagnosis, previous research (Bruce, 2002) suggested that alternative uses for the section were common. Respondents were asked to reflect on the frequency of use and usefulness of the section for all of these potential uses. Appendix A provides information on average use and usefulness responses provided for the different purposes examined.

AWA Scores

Table 2 provides the percentage of respondents selecting each of the frequency of use response options. These percentages indicate that AWA scores are being used by the majority of programs to meet the original purposes of the section, but scores are more often being used for admission selection than for deficiency diagnosis.

Table 2. Frequency of Use for AWA Scores for Admission Selection and Deficiency Diagnosis

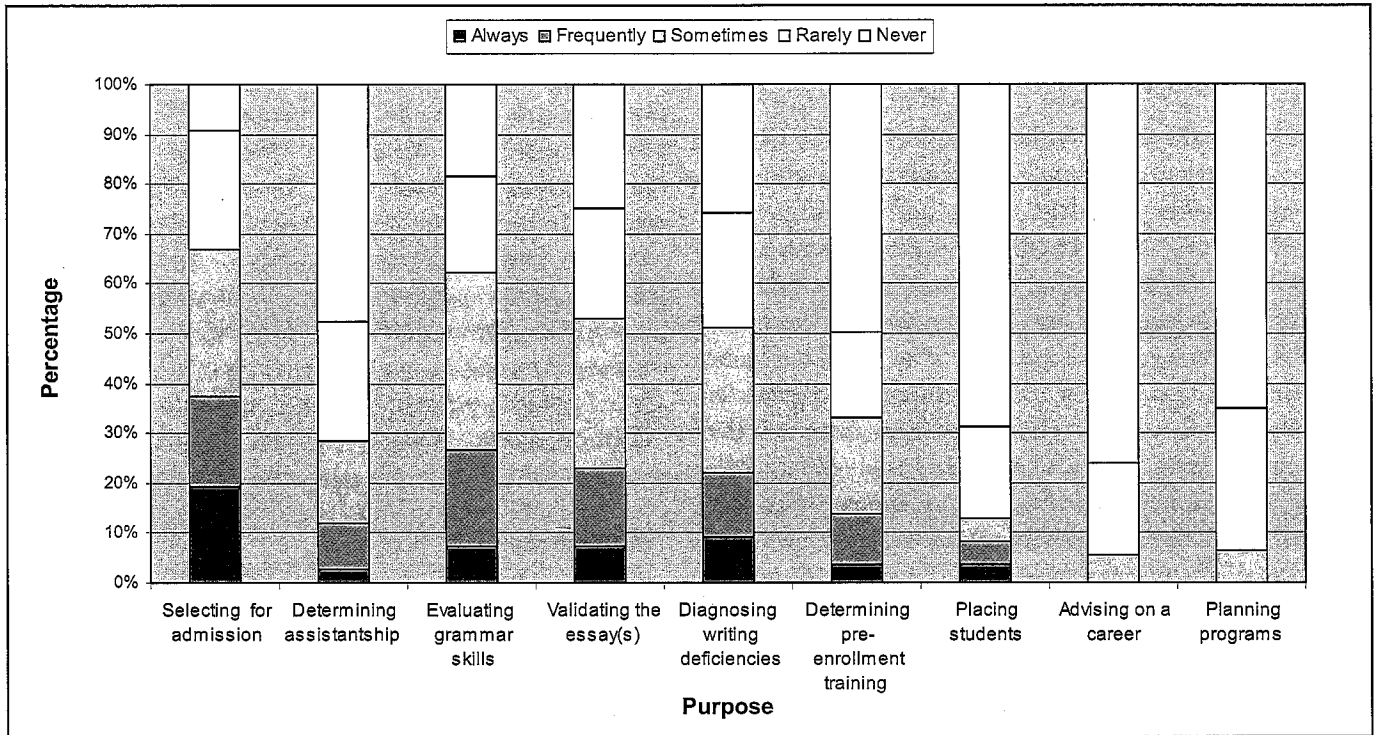
Response categories	Admission		Diagnosis	
	N	%	N	%
Never	10	9.2	28	25.7
Rarely	26	23.9	25	22.9
Sometimes	32	29.4	32	29.4
Frequently	20	18.3	14	12.8
Always	21	19.3	10	9.2

*Note. Percentages may not sum to 100 due to rounding.

Figure I displays the distribution of responses detailing the frequency with which AWA scores were used for a variety of purposes. Given the original purposes of the AWA were to aid in admission selection and deficiency diagnosis, it was hoped that respondents would indicate that scores and essays were especially used and found to be useful for these purposes. When results for AWA scores were examined, the findings revealed that scores were used

most often for selection and to a lesser extent for deficiency diagnosis. When alternative uses were examined, scores were at least sometimes used by a minimum of 50% of respondents for evaluating grammar skills and validating the program application essay. On the other hand, scores were not often used for career advising, student placement, or program planning.

Figure I: Frequency of Use for AWA Scores



The usefulness of AWA scores was also evaluated. Table 3 provides the distribution of responses indicating score usefulness for the original purposes of the AWA. Scores

were similar in terms of their usefulness ratings for selecting applicants for admission and diagnosing deficiencies.

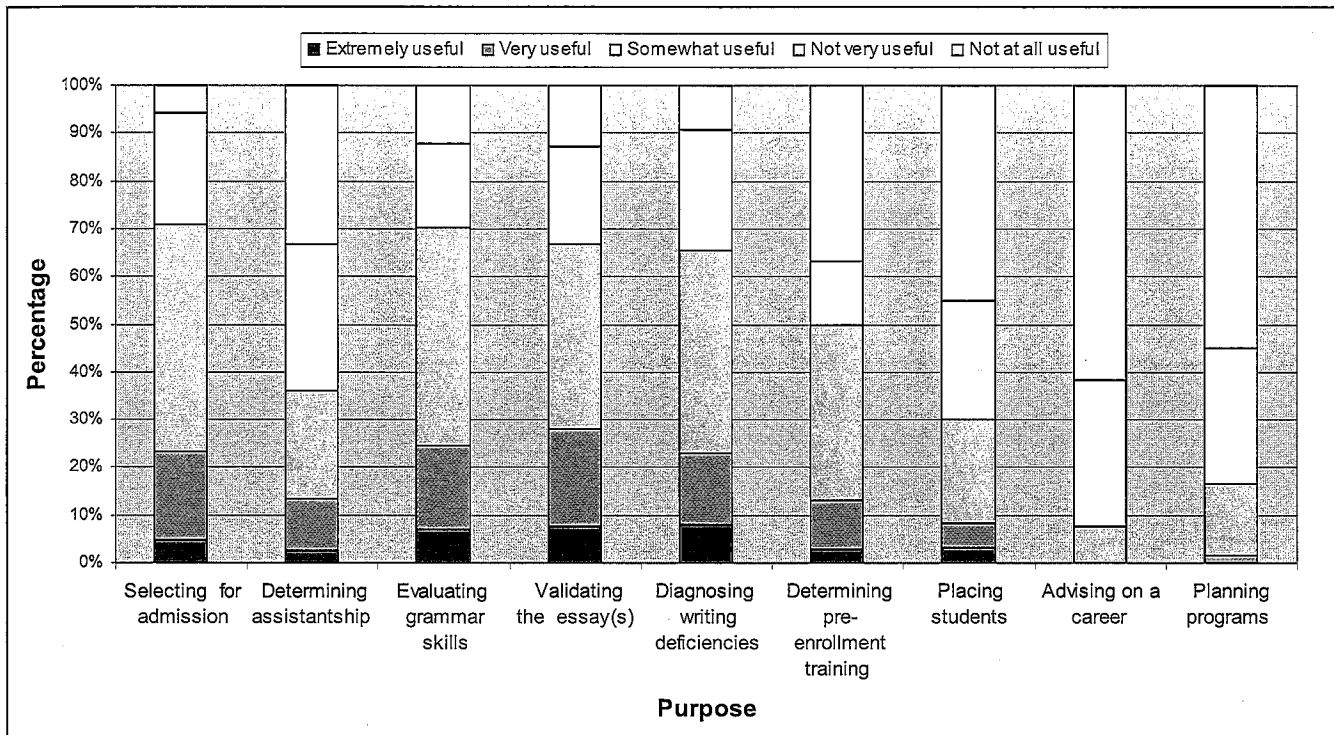
Response categories	Admission		Diagnosis	
	N	%	N	%
Not at all useful	6	5.8	8	9.2
Not very useful	24	23.3	22	25.3
Somewhat useful	49	47.6	37	42.5
Very useful	19	18.4	13	14.9
Extremely useful	5	4.9	7	8.0

Note. Percentages may not sum to 100 due to rounding.

Figure 2 demonstrates the usefulness of AWA scores for nine different purposes, including the original section uses. Very few respondents indicated that AWA scores were extremely useful for any of the purposes examined.

However, over 50% of respondents indicated that scores were at least somewhat useful for admission selection, grammar evaluation, program application essay validation, and deficiency diagnosis.

Figure 2: Usefulness of AWA Scores



AWA Essays

Table 4 provides information on the frequency with which essays were used to fulfill the original purpose of the AWA. As with the results for AWA scores, essays

were used more for admission than they were for diagnosis. However, when score and essay usage are compared, using Tables 2 and 4, it can be seen that scores are used more often for these purposes than essays.

Response categories	Admission		Diagnosis	
	N	%	N	%
Never	25	22.9	44	40.4
Rarely	29	26.6	19	17.4
Sometimes	32	29.4	32	29.4
Frequently	16	14.7	12	11.0
Always	7	6.4	2	1.8

*Note. Percentages may not sum to 100 due to rounding.

Figure 3 depicts frequency of use responses for AWA essays regarding the nine purposes identified earlier. When compared to the AWA score results, respondents indicated that they were less likely to use AWA essays than they were to use AWA scores, though use of essays

was still reasonable. For instance, approximately 50% of respondents indicated that they at least sometimes used essays for admission selection and grammar skill evaluation compared to score usage of 67% and 63%, respectively.

Figure 3: Frequency of Use for AWA Essays

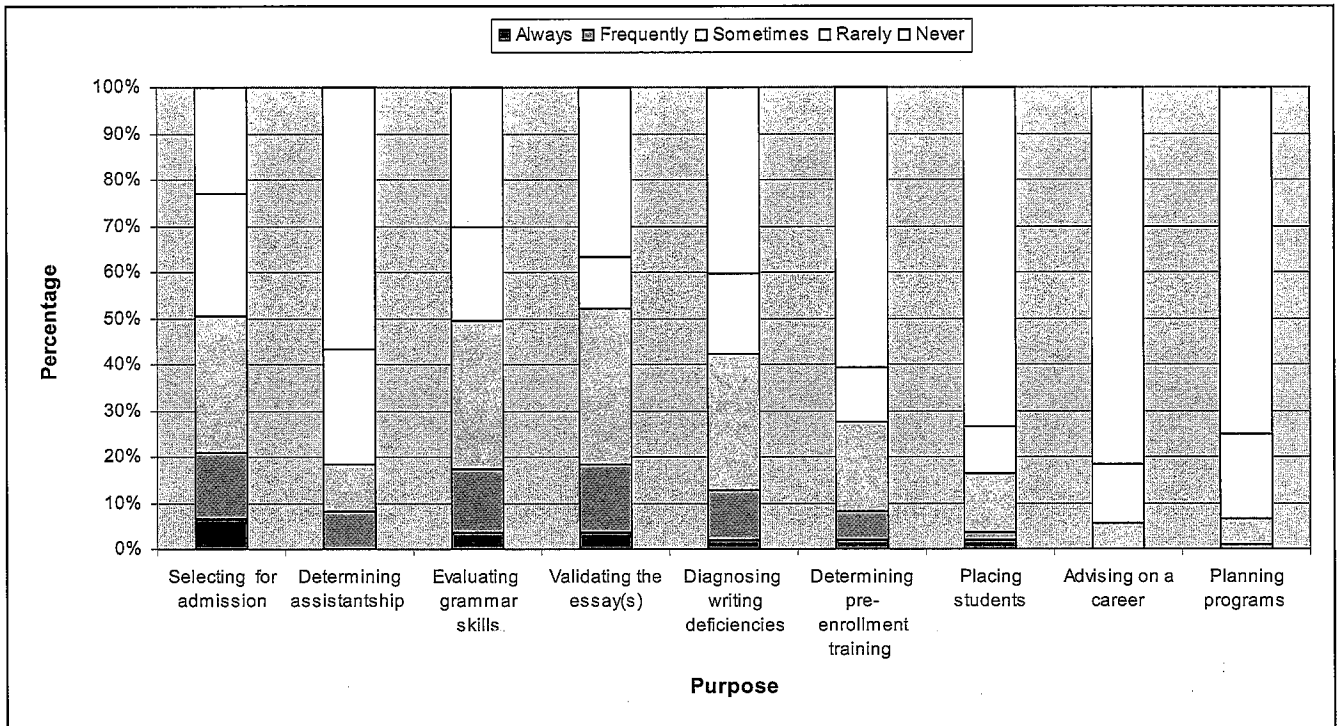


Table 5 presents the distribution of responses indicating the usefulness of essays to meet the original purposes of the AWA. Essays were rated as similarly useful for

admission selection and deficiency diagnosis. Additionally, for these two purposes, reported usefulness was similar for both essays and scores.

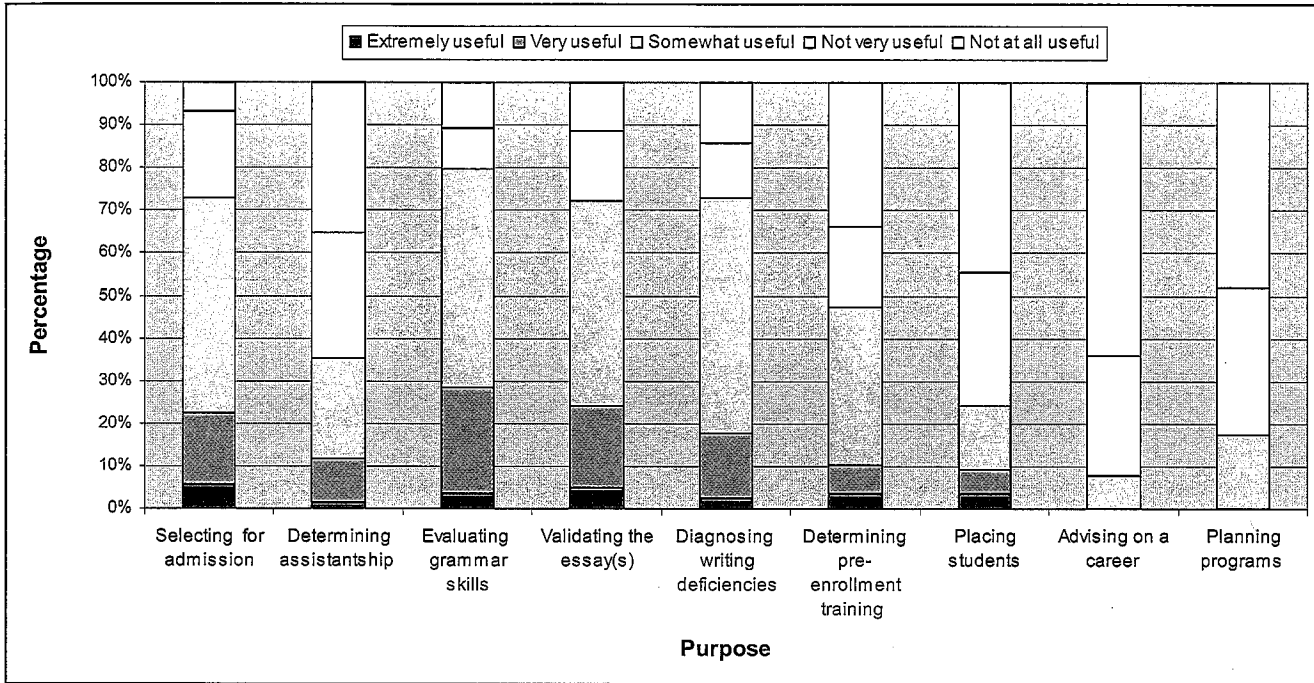
Response categories	Selection		Diagnosis	
	N	%	N	%
Not at all useful	6	6.7	11	14.1
Not very useful	18	20.2	10	12.8
Somewhat useful	45	50.6	43	55.1
Very useful	15	16.9	12	15.4
Extremely useful	5	5.6	2	2.6

*Note. Percentages may not sum to 100 due to rounding.

Figure 4 reveals that the essays were deemed at least somewhat useful by no less than 70% of respondents for the following purposes: admission selection, grammar skill evaluation, program application essay validation, and writing deficiency diagnosis. This was similar to the

findings for AWA scores. Though AWA essays were currently being used to a lesser extent than the scores, both components of the AWA were deemed useful for the original, and some alternative, purposes of the section.

Figure 4: Usefulness of AWA Essays



The results for use and usefulness of AWA scores and essays provides evidence that the different components of the AWA are informative and used to varying degrees, depending on the purposes expected. Though AWA scores were more frequently used than AWA essays for all purposes examined in this study, both were used and useful for meeting the original goals of the AWA section, namely admission selection and deficiency diagnosis. Scores and essays were also effective for evaluating grammar skills and validating the program application essay, which were uses not originally anticipated for the section. While programs do not frequently use the AWA section for purposes of advising and course placement, respondents found it was useful for selecting applicants for admission, evaluating applicant grammar abilities, validating the program application essay, and determining deficiencies. It is also feasible that programs use AWA scores and essays differently depending on their admission needs. As such, it is important to determine if program

differences are related to reported use and usefulness of the AWA.

Comparing Current AWA Use and Usefulness across Subgroups

Use for Non-Native English Speaking Applicants

While there are positive findings regarding the use and usefulness of the AWA section for its original purposes, the section is also being used to accomplish other program needs. Previous AWA research found that scores and essays were being used to evaluate grammar, validate the essay applicants write as a part of a program's admission requirements, and determine if English was a second or potentially problematic language for applicants (Bruce, 2002; Noll & Stowers, 1998).

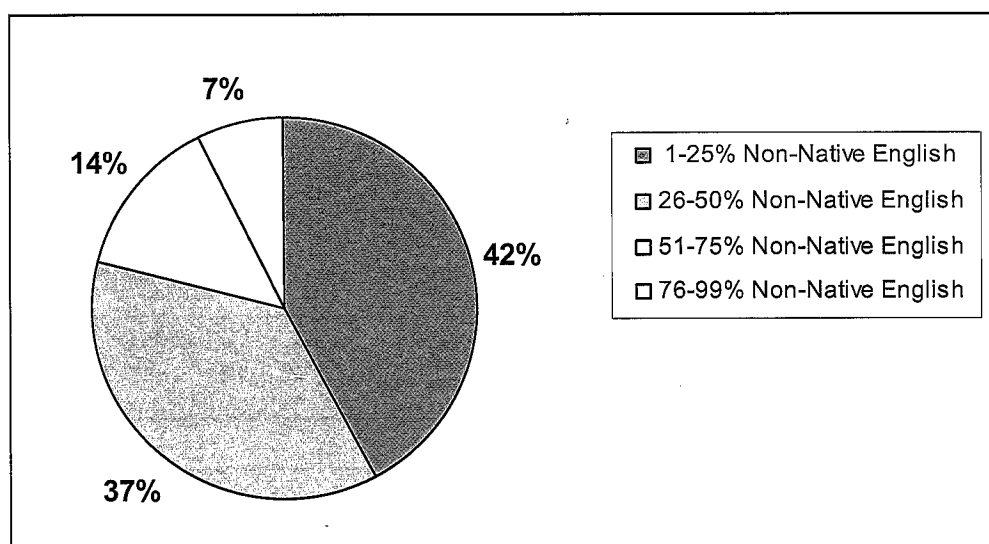
As a part of the current study, respondents were asked to indicate the proportion of native vs. non-native English

speaking applicants for which AWA essays were read prior to admission selection. Respondents indicated that they were more likely to read both or at least one of the AWA essays for non-native English speaking applicants when compared to their native English speaking counterparts. Specifically, 27% of respondents indicated that they read both essays for all or most of their non-native English speaking applicants compared to only 9% of respondents who read both essays for all or most of their native English applicants.

Respondents to the current survey were also asked to select the proportion of applicants to their program who

were non-native English speakers (see Figure 5). Of the respondents, 42% indicated that 25% or less of their applicants were non-native English speakers. An additional 37% indicated that non-native English speakers composed 26–50% of their applicant pool, and the remaining 21% of programs reported that more than 50% of their applicants were non-native English speakers. Thus, programs can be distinguished based on the percentage of non-native English speaking applicants, and perhaps their need for and use of an instrument to identify student writing or grammar deficiencies.

Figure 5: Proportion on Non-Native English Applicants



Figures 6–9 compare the distribution of responses for the group of programs that reported their applicant pool was composed of more than 25% non-native English speakers to a group of programs reporting that less than 25% of their applicants were non-native English speakers. The four categories demonstrated in Figure 5 were divided into two groups, rather than four, to allow for comparisons of similar sample size. These two groups may have unique needs and uses for the AWA section because of the varying concentration of non-native English speaking applicants. Figures 6–9 represent some uses of the AWA scores and essays that demonstrated variability in use and usefulness between the two groups. The most notable distinctions between programs with different

concentrations of non-native English speakers were in the reported uses and usefulness of AWA scores for writing deficiency diagnosis and uses and usefulness of AWA essays for grammar score evaluation. Tables comparing frequency of use and usefulness of the AWA section for these two groups for various purposes can be found in Appendix B.

Figure 6 compares variations in the frequency with which AWA scores were used for diagnosing writing deficiencies. It can be seen from this figure that programs with an applicant pool composed of more than 25% non-native English speakers reported more frequent use of AWA scores for diagnosis of writing deficiencies.

Similarly, programs with a higher concentration of non-native English speaking applicants were more likely to report that AWA scores were extremely or very useful for diagnosing writing deficiencies, as shown in Figure 7.

Programs with fewer non-native English speaking applicants were less likely to use AWA scores and found scores to be less useful for writing deficiency diagnosis.

Figure 6: AWA Score Use for Writing Deficiency Diagnosis by Non-native English Applicant Pool

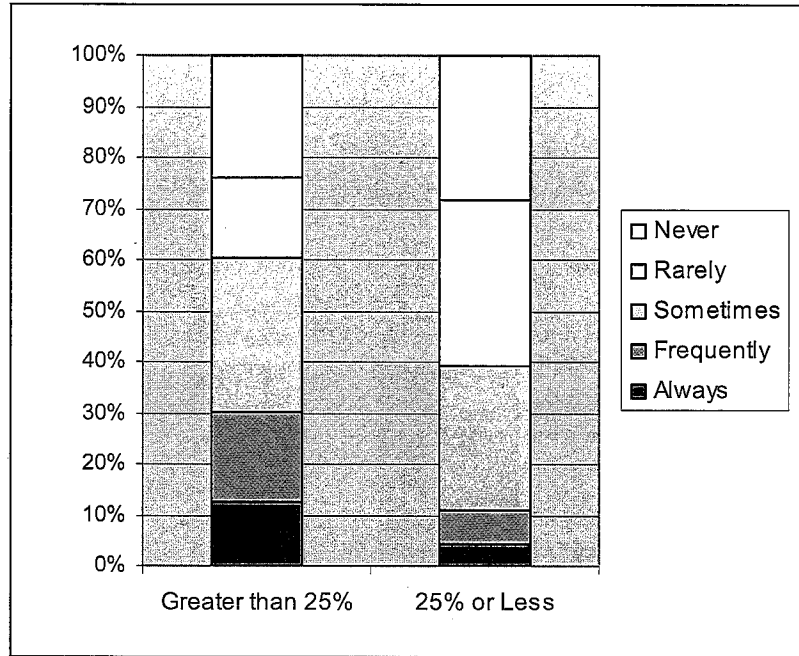
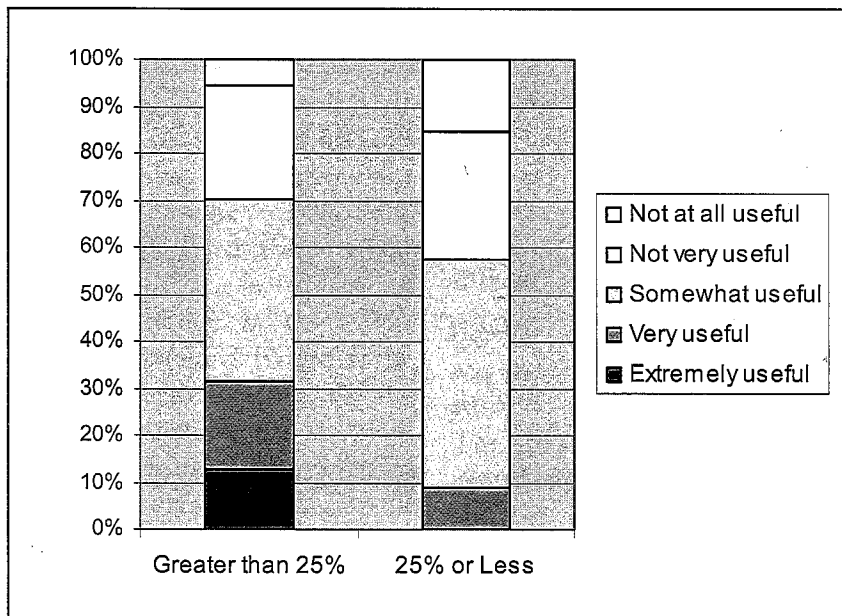


Figure 7: AWA Score Usefulness for Writing Deficiency Diagnosis by Non-Native English Applicant Pool



Figures 8 and 9 provide information on use and usefulness of AWA essays for the evaluation of applicant grammar skills. When responses were examined separately based on concentration of non-native English speaking applicants, differences were found between the two groups. Programs with a greater concentration of non-native English speaking applicants reported greater use and usefulness of AWA essays for evaluation of grammar ability. Overall, programs with greater concentrations of non-native English applicants reported more frequent use of AWA scores for diagnosis and essays for grammar skill evaluation; this distribution of responses varied from that demonstrated for the comparison group. Additionally, a

greater percentage of programs with more than 25% non-native English speaking applicants indicated that scores and essays were very or extremely useful for these two purposes compared to programs with fewer non-native English speaking applicants.

Overall, it appears that some differences exist in terms of perceived use and usefulness of the AWA section based on the number of non-native English speaking applicants a program receives. Figures 6–9 and Appendix B detail these differences for nine specific purposes for both AWA scores and essays. The variation in response patterns were more pronounced for those purposes related to evaluating grammar ability and validating applicant writing samples.

Figure 8: AWA Essay Use for Evaluation of Grammar Skills by Non-Native English Applicant Pool

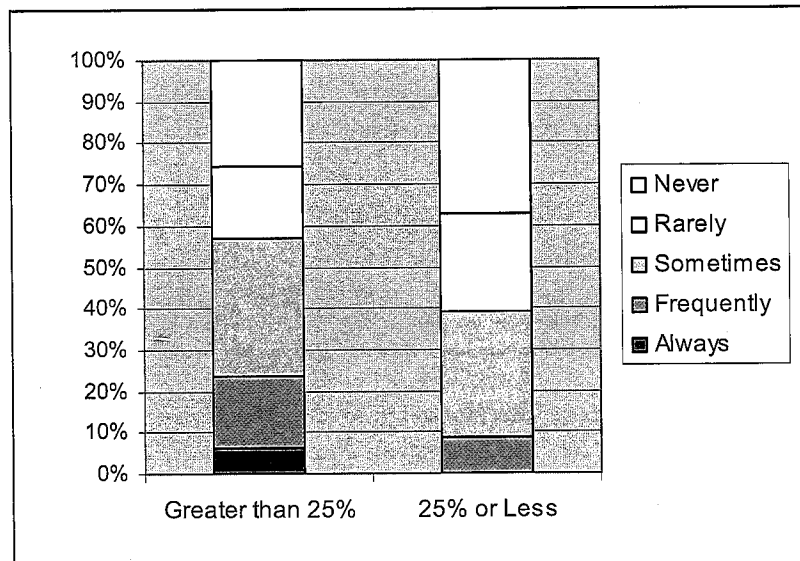
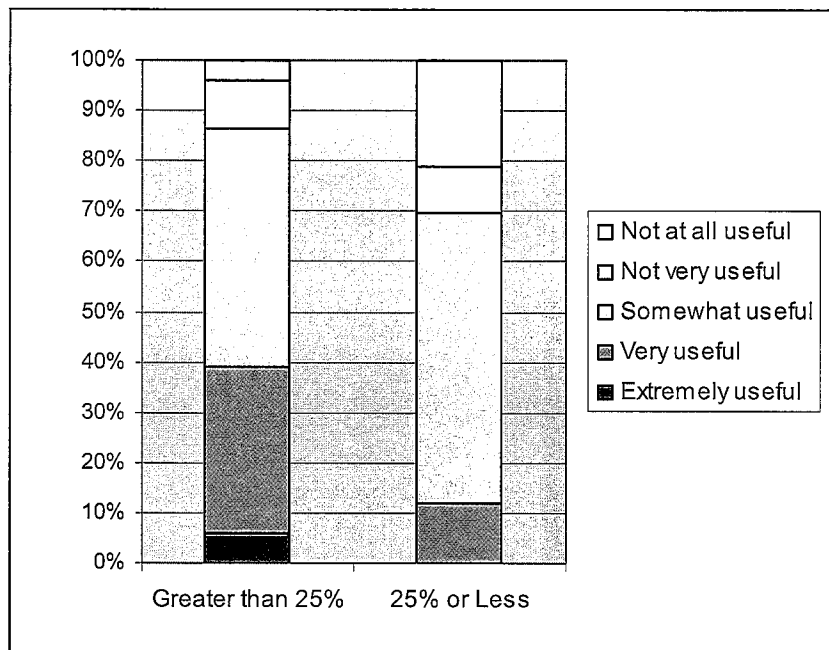


Figure 9: AWA Essay Usefulness for Evaluation of Grammar Skills by Non-Native English Applicant Pool



Use by Program Type

Results were also examined separately for the different program types (i.e., full-time, part-time, executive, and doctoral) and can be found in Appendix C. Differences between the program types in terms of frequency of use and reported usefulness of the AWA section for the variety of purposes described earlier are noted. When comparing response patterns across program types, there was little variability in terms of frequency of use and usefulness of the AWA section for the various purposes examined.

Overall, the distribution of responses regarding use and usefulness of the AWA section were similar across the different program types. One notable difference was that executive MBA programs tended to select the response option extremely useful less often than the other program types when asked about the usefulness of the AWA section for many of the purposes examined. However, the infrequent selection of extremely useful to describe the value of essays and scores for pre-enrollment training, course placement, career advisement, and program planning was consistent across program types. For all program types, scores and essays were most often used for selecting applicants for admission. There were some slight differences among program types in terms of the purposes

for which scores and essays were most useful. For example, full- and part-time programs indicated that scores were most useful for validating the application essays, while executive MBA and doctoral programs found scores most useful for selecting applicants for admission. However, it does not appear that use and usefulness of AWA essays or scores varies greatly depending on program type.

Based on the results from the current study, it appears that AWA scores and essays are meeting their original purposes for admission selection and deficiency diagnosis in addition to alternative purposes. Also, programs with larger concentrations of non-native English speaking applicants find scores and essays especially useful for deficiency diagnosis and grammar skill evaluation.

Discussion

Further perspective on the use and usefulness of the AWA was obtained by comparing the findings from the current study to previously reported perceptions of the section. The following discussion examines changes in the perceived use and usefulness of the AWA for admissions and for diagnosis since the section's addition to the GMAT® exam.

Comparing Current AWA Use and Usefulness to Previous Findings

Use for Admission Selection. Previous research indicated that programs anticipated the addition of an AWA section to the GMAT® exam would be used and useful for selecting applicants for admission into graduate management programs (Bruce, 1984, 1992, 1993). In the present study, actual use and usefulness for admission selection were examined. The results revealed that the majority of respondents always, frequently, or sometimes use AWA scores and essays to select applicants for admission, with combined percentages of 67% and 51%, respectively. Additionally, 69% of respondents indicated that either the scores or essays were always, frequently, or sometimes used for admission selection, as shown in Figure 10. Thus, about two-thirds of respondents indicated use of AWA scores and one-half responded they used AWA essays for admission selection.

Overall, respondents indicated that the AWA is currently being used for admission selection for the majority of the programs represented in this study. In the Noll and

Stowers (1998) study, 86% reported that, yes, they did use the AWA section. However, Noll and Stowers did not differentiate between scores and essays in terms of use for admission selection. In order to make a dual comparison between Noll and Stowers and the current study, responses to scores and essays were combined. This combined category is labeled "Either-2005" in Figure 10. Figure 10 shows that previous reported frequency of use of the section for admission selection in 1998 was higher than in the 2005 study. However, scores and essays are still being used to accomplish this goal by a number of programs.

The perceived and actual usefulness of AWA scores and essays for admission selection were lower in 2005 but comparable to results from Bruce's 1993 study. Figure 11 shows that 71% of respondents reported AWA scores were extremely, very, or somewhat useful as a part of the selection process, and 73% specified that AWA essays were extremely, very, or somewhat useful for selection. Approximately 77% of respondents indicated that either the AWA scores or essays were extremely, very, or somewhat useful for admission selection.

Figure 10: Comparison of 1998 and 2005 Survey Results on Reported Use of the AWA for Admission Selection

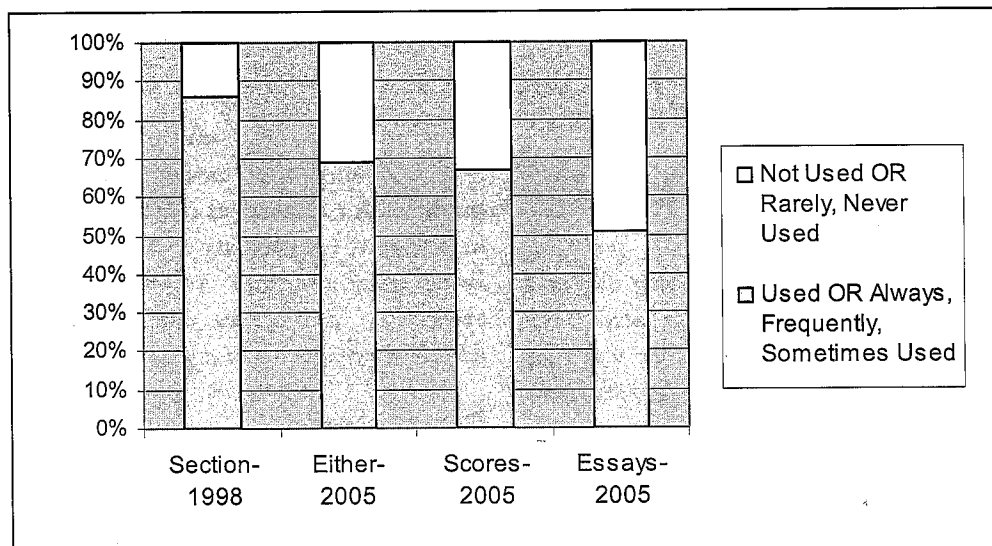
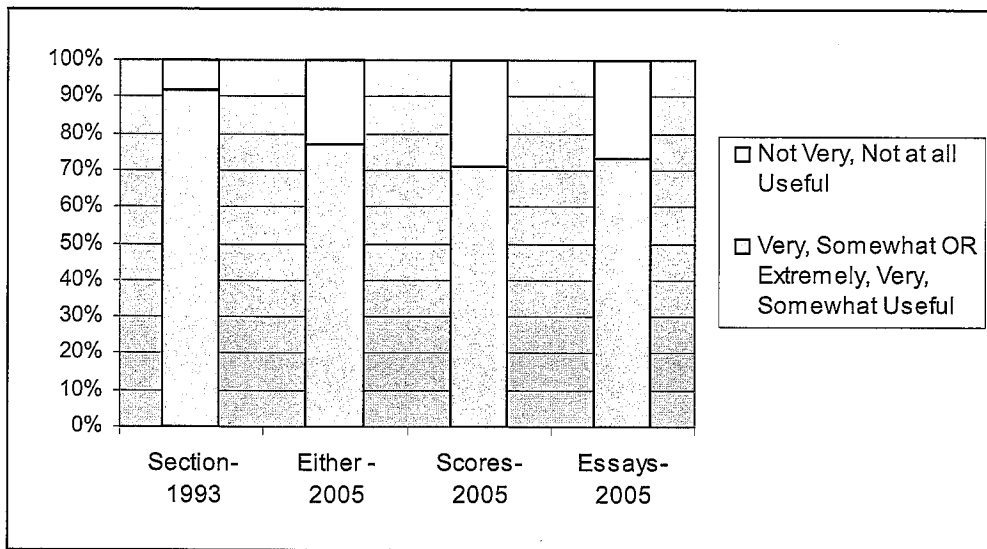


Figure: II: Comparison of 1993 and 2005 Survey Results on Reported Usefulness of the AWA for Admission Selection



Overall, respondents found AWA scores and essays to be similar in terms of usefulness when used to select applicants for admission. Bruce (1993) found that 91% of respondents perceived the addition of the AWA would be very or somewhat useful for admission selection.

In summary, frequency of use for selection was lower in 2005 than previously reported in 1998, but the AWA was still used by the majority of respondents. In addition, a higher percentage of respondents found it to be as useful for selecting applicants for admission, as was previously anticipated in 1993.

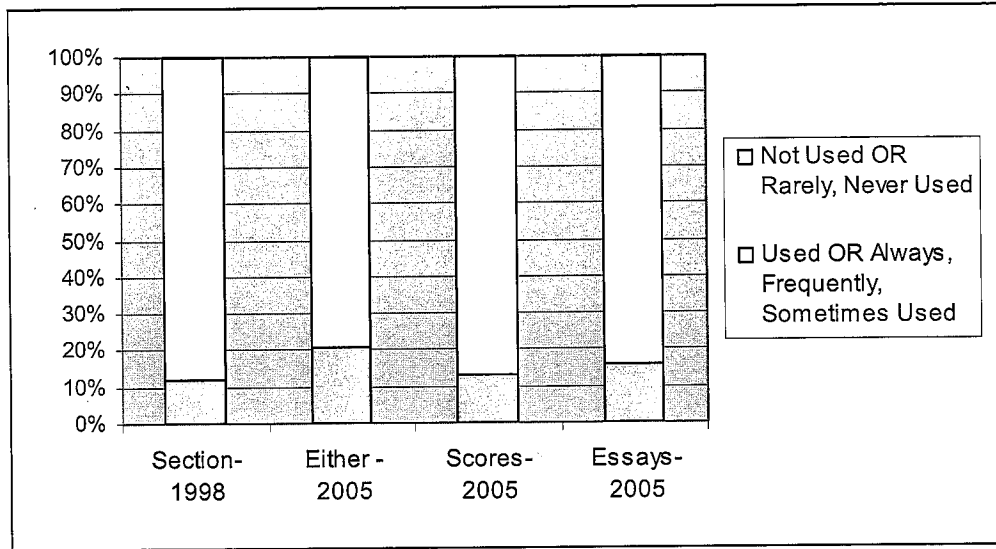
Use for Diagnostic Purposes

To further explore whether the AWA was meeting its original purposes, current use and usefulness of the section for diagnostic purposes were examined. Noll and Stowers (1998) and Bruce (1993) included course placement as a component of deficiency diagnosis. For comparison purposes, the diagnostic results for the current study included analyses for both deficiency diagnosis and course placement.

Overall, respondents to the current study did not use the AWA as frequently for course placement or diagnosis as they did for admission selection. The percentage of respondents indicating that AWA scores were always, frequently, or sometimes used to place students in courses or to diagnose writing deficiencies was 13% and 51%, respectively. In terms of AWA essays, 16% and 42% said that essays were always, frequently, or sometimes used to place students into courses or diagnose deficiencies, respectively.

Similar results for course placement were found by Noll and Stowers (1998); 12% said yes they used the AWA section to place students in writing courses. By examining respondents who indicated that they always, frequently, or sometimes used either the scores or essays for course placement, a better understanding of use of the entire section can be achieved. From this, it was revealed that approximately 21% of respondents reported using either scores or essays for course placement. A comparison of the current results to previous findings of AWA use for course placement can be found in Figure 12.

Figure I2: Comparison of 1998 and 2005 Survey Results on Reported Use of the AWA for Course Placement

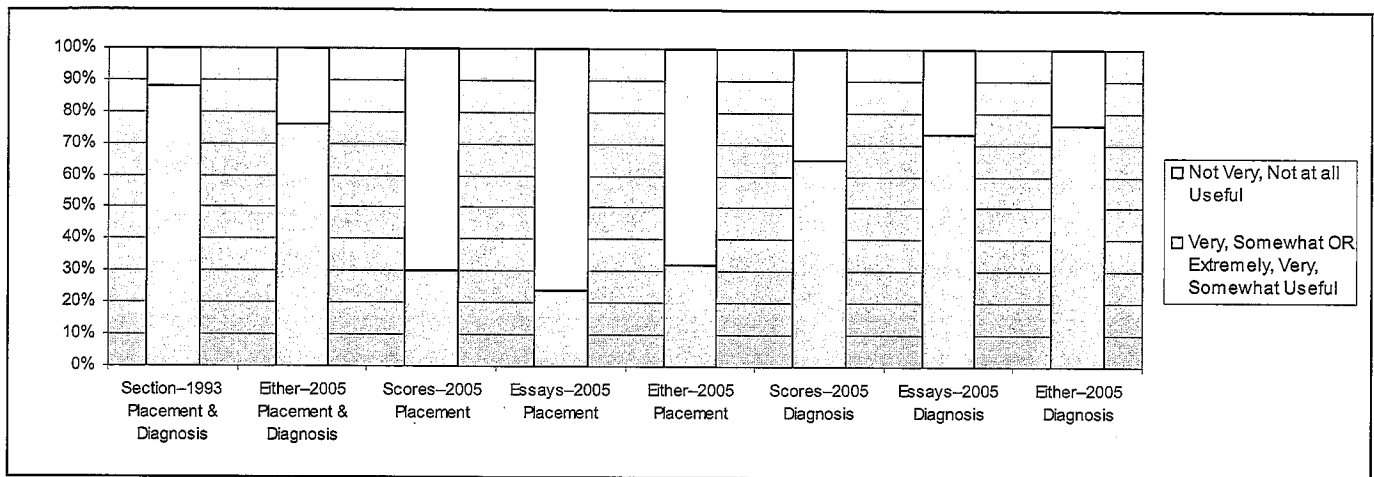


Few respondents felt that the AWA essays or scores were helpful in placing students in courses, but more felt that the section was at least somewhat useful for diagnosing student writing deficiencies. Of those who responded, 30% said scores were extremely, very, or somewhat useful for placing students in courses, and 65% said AWA scores were extremely, very, or somewhat useful in diagnosing writing deficiencies. Only 24% of respondents indicated that essays were extremely, very, or somewhat useful for placing students in courses, whereas 73% indicated that AWA essays were extremely, very, or somewhat useful for diagnosing writing deficiencies. When respondents who indicated that either scores or essays were extremely, very, or somewhat useful were

examined, the percentages increased slightly to 32% and 76%, for course placement and diagnosis, respectively.

Bruce (1993) reported that 88% perceived the AWA would be very or somewhat useful as a diagnostic component or for course placement. A comparison of the 1993 and the 2005 results can be found in Figure I3. For the 2005 study, respondents who indicated that either scores or essays were extremely, very, or somewhat useful for course placement or diagnosis of writing deficiencies were combined into a category labeled "Either-2005-Placement & Diagnosis" in Figure I3. Approximately 76% of respondents indicated that essays or scores were useful for course placement or writing deficiency diagnosis.

Figure I3: Comparison of 1993 and 2005 Survey Results on Reported Usefulness of the AWA for Course Placement and Writing Deficiency Diagnosis



Conclusion and Implications

The present study provided an examination of the AWA section of the GMAT® exam a decade after its original implementation. The findings suggest the GMAT® AWA is fulfilling its original goals. Approximately, 77% of respondents find either the scores or essays useful for admission, and 76% reported them useful for diagnosis. Though the percentage of use and reported usefulness demonstrated that the section was currently meeting programs' needs, fewer programs were using the section than was originally anticipated. The section was also not as used or useful for course placement as was originally expected. This is not surprising given that course placement was not technically one of the original purposes of the AWA section.

This investigation of the AWA use and usefulness revealed that programs with higher concentrations of non-native English speaking applicants found the AWA section to be of greater use. Programs with more non-native English speaking applicants reported that the section was used more often and was more useful in evaluating English grammar skills and diagnosing writing deficiencies. Since programs with more non-native English speaking applicants likely have a greater need to identify writing and grammar deficiencies, it appears that the section is especially meeting the requirements of this group. When results were examined by program type, differences in reported frequency of use and usefulness were not apparent.

As with most research, there were limitations to the current study. First, the sample size was small given the larger number of programs surveyed. As such, the results may not be representative of all graduate business programs. Additionally, one of the purposes of the present study was to compare findings to previous research. Because there were a number of differences between this study and those conducted by other researchers, results were presented and combined in several ways to allow findings to be comparable. However, relevant information can be lost or misinterpreted when categories and groups are combined and separated. Though every effort was made to accurately represent these findings, additional research more closely replicating the previous studies would allow for more exact comparisons.

Finally, future research should also examine ways to enhance the AWA to meet the needs of all types of programs with varying applicant populations. This may include the use of the section for purposes not originally intended. For instance, the AWA may be effective for English grammar skill evaluation and application essay validation if it is modified to more adequately meet these needs. Additionally, a comparison of use and usefulness of the AWA for graduate management programs with findings for other analytical writing tasks could provide additional insight. This would allow for a comprehensive understanding of the usefulness of analytical writing assessments in selecting applicants for admission to higher education institutions. As these results may extend to many

assessment programs, other organizations should consider re-investigating the use and usefulness of their own assessments to determine if they are still meeting their desired purposes.

Contact Information

For questions or comments regarding study findings, methodology or data, please contact the GMAC® Research and Development department at research@gmac.com.

References

- Bridgeman, B., & Carlson, S. B. (1984). Survey of academic writing tasks. *Written Communication*, 1 (2), 247-280.
- Bruce, G. (1984). Reactions to the concept of a GMAT® writing assessment: Results of a survey of institutions using the GMAT®. Santa Monica, CA: GMAC®.
- Bruce, G. (1992). Exploratory marketing research on the addition of an analytical writing assessment to the GMAT®. Santa Monica, CA: GMAC®.
- Bruce, G. (1993). Attitudes toward the addition of an analytical writing assessment to the GMAT®: Results of a survey of institution using the GMAT®. Santa Monica, CA: GMAC®.
- Bruce, G. (2002). Results of the 2002 GMAC® customer feedback survey. McLean, VA: GMAC®.
- Graduate Management Admission Council® (n.d). The GMAT® analytical writing assessment: An introduction. [Brochure]. Santa Monica, CA: Author.
- Noll, C. L., & Stowers, R. H. (1998). How MBA programs are using the GMAT®'s Analytical Writing Assessment. *Business Communication Quarterly*, 61(4), 66-71.
- Powers, D. E., & Fowles, M. E. (2002). Balancing test user needs and responsible professional practice: A case study involving the assessment of graduate-level writing. *Applied Measurement in Education*, 15(3), 217-247.
- Quellmalz, E. S. (1984). Designing writing assessments: Balancing fairness, utility, and cost. *Educational Evaluation and Policy Analysis*, 6(1), 63-72.
- Rogers, P. S., & Rymer, J. (1995). What is the relevance of the GMAT® analytical writing assessment for management education? A critical analysis, part I. *Management*, 8(3), 347-367

Appendix A: Descriptive Statistics

Table A-I: Descriptive Statistics for All Purposes								
Purposes	Use				Usefulness			
	N	M	Med.	SD	N	M	Med.	SD
Scores								
Admission	109	3.15	3.00	1.25	103	2.93	3.00	0.92
Assistantships	109	1.95	2.00	1.13	75	2.19	2.00	1.10
Grammar	109	2.78	3.00	1.17	98	2.90	3.00	1.06
Validation	109	2.59	3.00	1.23	93	2.89	3.00	1.11
Writing	109	2.57	3.00	1.26	87	2.87	3.00	1.04
Programs	109	1.41	1.00	0.61	60	1.63	1.00	0.80
Placement	109	1.56	1.00	1.03	60	1.97	2.00	1.09
Pre-training	109	2.01	2.00	1.20	68	2.29	2.50	1.16
Career	109	1.29	1.00	0.57	52	1.46	1.00	0.64
Essays								
Admission	109	2.55	3.00	1.18	89	2.94	3.00	0.93
Assistantships	109	1.70	1.00	0.96	68	2.13	2.00	1.06
Grammar	109	2.40	2.00	1.16	84	3.01	3.00	0.96
Validation	109	2.38	3.00	1.22	79	2.90	3.00	1.01
Writing	109	2.17	2.00	1.14	78	2.79	3.00	0.96
Programs	109	1.32	1.00	0.62	52	1.69	2.00	0.76
Placement	109	1.49	1.00	0.92	54	1.93	2.00	1.08
Pre-training	109	1.77	1.00	1.09	59	2.27	2.00	1.11
Career	109	1.24	1.00	0.54	50	1.44	1.00	0.64

Appendix B: Use and Usefulness of the AWA by Non-Native English Speaking Applicant Concentration

Table B-1: Use and Usefulness of AWA for Admission Selection				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	6.35	13.04	23.81	21.74
Rarely	17.46	32.61	23.81	30.43
Sometimes	31.75	26.09	30.16	28.26
Frequently	19.05	17.39	15.87	13.04
Always	25.40	10.87	6.35	6.52
Overall statistics				
N	63	46	63	46
Mean	3.40	2.80	2.57	2.52
Median	3.00	3.00	3.00	2.00
SD	1.23	1.20	1.20	1.17
% Usefulness				
Not at all useful	3.33	9.30	3.77	11.11
Not very useful	21.67	25.58	16.98	25.00
Somewhat useful	45.00	51.16	52.83	47.22
Very useful	23.33	11.63	16.98	16.67
Extremely useful	6.67	2.33	9.43	0.00
Overall statistics				
N	60	43	53	36
Mean	3.08	2.72	3.11	2.69
Median	3.00	3.00	3.00	3.00
SD	0.93	0.88	0.93	0.89

Table B-2: Use and Usefulness of AWA for Determining Assistantships				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	46.03	50.00	58.73	54.35
Rarely	26.98	19.57	26.98	21.74
Sometimes	15.87	17.39	9.52	10.87
Frequently	9.52	8.70	4.76	13.04
Always	1.59	4.35	0.00	0.00
Overall statistics				
N	63	46	63	46
Mean	1.94	1.98	1.60	1.83
Median	2.00	1.50	1.00	1.00
SD	1.08	1.20	0.85	1.08
% Usefulness				
Not at all useful	30.95	36.36	43.59	24.14
Not very useful	30.95	30.30	17.95	44.83
Somewhat useful	26.19	18.18	25.64	20.69
Very useful	9.52	12.12	10.26	10.34
Extremely useful	2.38	3.03	2.56	0.00
Overall statistics				
N	42	33	39	29
Mean	2.21	2.15	2.10	2.17
Median	2.00	2.00	2.00	2.00
SD	1.07	1.15	1.17	0.93

Table B-3: Use and Usefulness of AWA for Evaluating Grammar Skills				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	17.46	19.57	25.40	36.96
Rarely	14.29	26.09	17.46	23.91
Sometimes	36.51	34.78	33.33	30.43
Frequently	23.81	13.04	17.46	8.70
Always	7.94	6.52	6.35	0.00
Overall statistics				
N	63	46	63	46
Mean	2.90	2.61	2.62	2.11
Median	3.00	3.00	3.00	2.00
SD	1.19	1.15	1.22	1.02
% Usefulness				
Not at all useful	6.90	20.00	3.92	21.21
Not very useful	13.79	22.50	9.80	9.09
Somewhat useful	46.55	45.00	47.06	57.58
Very useful	22.41	10.00	33.33	12.12
Extremely useful	10.34	2.50	5.88	0.00
Overall statistics				
N	58	40	51	33
Mean	3.16	2.53	3.27	2.61
Median	3.00	3.00	3.00	3.00
SD	1.02	1.01	0.87	0.97

Table B-4: Use and Usefulness of AWA for Validating the Application Essay				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	23.81	26.09	31.75	43.48
Rarely	15.87	30.43	6.35	17.39
Sometimes	31.75	28.26	38.10	28.26
Frequently	17.46	13.04	19.05	8.70
Always	11.11	2.17	4.76	2.17
Overall statistics				
N	63	46	63	46
Mean	2.76	2.35	2.59	2.09
Median	3.00	2.00	3.00	2.00
SD	1.30	1.08	1.25	1.13
% Usefulness				
Not at all useful	9.09	18.42	6.12	20.00
Not very useful	14.55	28.95	12.24	23.33
Somewhat useful	41.82	34.21	53.06	40.00
Very useful	23.64	15.79	20.41	16.67
Extremely useful	10.91	2.63	8.16	0.00
Overall statistics				
N	55	38	49	30
Mean	3.13	2.55	3.12	2.53
Median	3.00	3.00	3.00	3.00
SD	1.09	1.06	0.95	1.01

Table B-5: Use and Usefulness of AWA for Diagnosing Writing Deficiencies				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	23.81	28.26	38.10	43.48
Rarely	15.87	32.61	14.29	21.74
Sometimes	30.16	28.26	28.57	30.43
Frequently	17.46	6.52	15.87	4.35
Always	12.70	4.35	3.17	0.00
Overall statistics				
N	63	46	63	46
Mean	2.79	2.26	2.32	1.96
Median	3.00	2.00	2.00	2.00
SD	1.33	1.08	1.23	0.97
% Usefulness				
Not at all useful	5.56	15.15	8.33	23.33
Not very useful	24.07	27.27	8.33	20.00
Somewhat useful	38.89	48.48	58.33	50.00
Very useful	18.52	9.09	20.83	6.67
Extremely useful	12.96	0.00	4.17	0.00
Overall statistics				
N	54	33	48	30
Mean	3.09	2.52	3.04	2.40
Median	3.00	3.00	3.00	3.00
SD	1.09	0.87	0.90	0.93

Table B-6: Use and Usefulness of AWA for Determining Pre-Enrollment Training				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	46.03	54.35	52.38	71.74
Rarely	11.11	26.09	9.52	15.22
Sometimes	23.81	13.04	26.98	8.70
Frequently	14.29	4.35	9.52	2.17
Always	4.76	2.17	1.59	2.17
Overall statistics				
N	63	46	63	46
Mean	2.21	1.74	1.98	1.48
Median	2.00	1.00	1.00	1.00
SD	1.30	1.00	1.16	0.91
% Usefulness				
Not at all useful	27.91	52.00	26.32	47.62
Not very useful	6.98	24.00	7.89	38.10
Somewhat useful	46.51	20.00	55.26	4.76
Very useful	13.95	4.00	5.26	9.52
Extremely useful	4.65	0.00	5.26	0.00
Overall statistics				
N	43	25	38	21
Mean	2.60	1.76	2.55	1.76
Median	3.00	1.00	3.00	2.00
SD	1.18	0.93	1.11	0.94

Table B-7: Use and Usefulness of AWA for Placing Students into Courses				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	63.49	76.09	71.43	76.09
Rarely	19.05	17.39	7.94	13.04
Sometimes	7.94	0.00	15.87	8.70
Frequently	6.35	2.17	3.17	0.00
Always	3.17	4.35	1.59	2.17
Overall statistics				
N	63	46	63	46
Mean	1.67	1.41	1.56	1.39
Median	1.00	1.00	1.00	1.00
SD	1.08	0.96	0.98	0.83
% Usefulness				
Not at all useful	37.14	56.00	41.18	50.00
Not very useful	22.86	28.00	26.47	40.00
Somewhat useful	28.57	12.00	20.59	5.00
Very useful	5.71	4.00	5.88	5.00
Extremely useful	5.71	0.00	5.88	0.00
Overall statistics				
N	35	25	34	20
Mean	2.20	1.64	2.09	1.65
Median	2.00	1.00	2.00	1.50
SD	1.18	0.86	1.19	0.81

Table B-8: Use and Usefulness of AWA for Advising on a Career Path				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	79.37	71.74	84.13	78.26
Rarely	17.46	19.57	11.11	15.22
Sometimes	3.17	8.70	4.76	6.52
Frequently	0.00	0.00	0.00	0.00
Always	0.00	0.00	0.00	0.00
Overall statistics				
N	63	46	63	46
Mean	1.24	1.37	1.21	1.28
Median	1.00	1.00	1.00	1.00
SD	0.50	0.65	0.51	0.58
% Usefulness				
Not at all useful	62.07	60.87	56.67	75.00
Not very useful	31.03	30.43	36.67	15.00
Somewhat useful	6.90	8.70	6.67	10.00
Very useful	0.00	0.00	0.00	0.00
Extremely useful	0.00	0.00	0.00	0.00
Overall statistics				
N	29	23	30	20
Mean	1.45	1.48	1.50	1.35
Median	1.00	1.00	1.00	1.00
SD	0.63	0.67	0.63	0.67

Table B-9: Use and Usefulness of AWA for Planning Programs and Courses				
Use/Usefulness	Scores		Essays	
	>25%	≤25%	>25%	≤25%
% Use				
Never	66.67	63.04	76.19	73.91
Rarely	26.98	30.43	17.46	19.57
Sometimes	6.35	6.52	4.76	6.52
Frequently	0.00	0.00	1.59	0.00
Always	0.00	0.00	76.19	73.91
Overall statistics				
N	63	46	63	46
Mean	1.40	1.43	1.32	1.33
Median	1.00	1.00	1.00	1.00
SD	0.61	0.62	0.64	0.60
% Usefulness				
Not at all useful	48.57	64.00	38.71	61.90
Not very useful	31.43	24.00	38.71	28.57
Somewhat useful	20.00	8.00	22.58	9.52
Very useful	0.00	4.00	0.00	0.00
Extremely useful	0.00	0.00	0.00	0.00
Overall statistics				
N	35	25	31	21
Mean	1.71	1.52	1.84	1.48
Median	2.00	1.00	2.00	1.00
SD	0.79	0.82	0.70	0.68

Appendix C: Use and Usefulness of the AWA by Program Type

Table C-I: Use and Usefulness of AWA for Admission Selection by Program Type								
Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	10.26	13.11	15.79	0.00	23.08	19.67	15.79	18.18
Rarely	17.95	22.95	26.32	27.27	23.08	29.51	31.58	9.09
Sometimes	35.90	29.51	15.79	36.36	30.77	26.23	26.32	54.55
Frequently	16.67	19.67	26.32	9.09	17.95	14.75	15.79	18.18
Always	19.23	14.75	15.79	27.27	5.13	9.84	10.53	0.00
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	3.17	3.00	3.00	3.36	2.59	2.66	2.74	2.73
Median	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
SD	1.23	1.25	1.37	1.21	1.18	1.24	1.24	1.01
% Usefulness								
Not at all useful	5.41	5.45	15.79	0.00	4.62	6.25	12.50	0.00
Not very useful	21.62	23.64	31.58	18.18	18.46	22.92	18.75	11.11
Somewhat useful	48.65	47.27	26.32	54.55	53.85	45.83	43.75	55.56
Very useful	17.57	18.18	15.79	27.27	18.46	20.83	25.00	0.00
Extremely useful	6.76	5.45	10.53	0.00	4.62	4.17	0.00	33.33
Overall statistics								
N	74	55	19	11	65	48	16	9
Mean	2.99	2.95	2.74	3.09	3.00	2.94	2.81	3.56
Median	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
SD	0.94	0.93	1.24	0.70	0.87	0.93	0.98	1.13

Table C-2: Use and Usefulness of AWA for Determining Assistantships by Program Type								
Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	46.15	47.54	57.89	36.36	51.28	47.54	57.89	54.55
Rarely	24.36	18.03	15.79	27.27	30.77	26.23	15.79	27.27
Sometimes	19.23	18.03	10.53	18.18	10.26	14.75	10.53	9.09
Frequently	7.69	11.48	10.53	18.18	7.69	11.48	15.79	9.09
Always	2.56	4.92	5.26	0.00	0.00	0.00	0.00	0.00
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	1.96	2.08	1.89	2.18	1.74	1.90	1.84	1.73
Median	2.00	2.00	1.00	2.00	1.00	2.00	1.00	1.00
SD	1.10	1.26	1.29	1.17	0.93	1.04	1.17	1.01
% Usefulness								
Not at all useful	29.09	30.43	56.25	14.29	32.69	24.39	46.15	20.00
Not very useful	36.36	26.09	25.00	42.86	28.85	34.15	23.08	40.00
Somewhat useful	21.82	23.91	6.25	28.57	26.92	26.83	23.08	20.00
Very useful	9.09	15.22	12.50	14.29	9.62	12.20	7.69	20.00
Extremely useful	3.64	4.35	0.00	0.00	1.92	2.44	0.00	0.00
Overall statistics								
N	55	46	16	7	52	41	13	5
Mean	2.22	2.37	1.75	2.43	2.19	2.34	1.92	2.40
Median	2.00	2.00	1.00	2.00	2.00	2.00	2.00	2.00
SD	1.08	1.20	1.07	0.98	1.07	1.06	1.04	1.14

Table C-3: Use and Usefulness of AWA for Evaluating Grammar Skills by Program Type

Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	17.95	16.39	15.79	18.18	24.36	31.15	21.05	36.36
Rarely	16.67	21.31	31.58	27.27	19.23	18.03	31.58	18.18
Sometimes	35.90	34.43	31.58	27.27	34.62	31.15	31.58	9.09
Frequently	20.51	16.39	21.05	27.27	17.95	16.39	15.79	27.27
Always	8.97	11.48	0.00	0.00	3.85	3.28	0.00	9.09
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	2.86	2.85	2.58	2.64	2.58	2.43	2.42	2.55
Median	3.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00
SD	1.20	1.22	1.02	1.12	1.16	1.19	1.02	1.51
% Usefulness								
Not at all useful	8.57	11.32	27.78	11.11	4.92	13.04	33.33	14.29
Not very useful	10.00	16.98	22.22	22.22	6.56	10.87	0.00	0.00
Somewhat useful	51.43	43.40	33.33	33.33	52.46	45.65	53.33	42.86
Very useful	21.43	18.87	16.67	22.22	32.79	26.09	13.33	28.57
Extremely useful	8.57	9.43	0.00	11.11	3.28	4.35	0.00	14.29
Overall statistics								
N	70	53	18	9	61	46	15	7
Mean	3.11	2.98	2.39	3.00	3.23	2.98	2.47	3.29
Median	3.00	3.00	2.50	3.00	3.00	3.00	3.00	3.00
SD	1.00	1.10	1.09	1.23	0.82	1.04	1.13	1.25

Table C-4: Use and Usefulness of AWA for Validating the Application Essay by Program Type								
Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	21.79	22.95	31.58	36.36	29.49	36.07	31.58	45.45
Rarely	17.95	21.31	15.79	27.27	11.54	11.48	15.79	0.00
Sometimes	33.33	32.79	31.58	18.18	37.18	29.51	31.58	27.27
Frequently	17.95	13.11	21.05	18.18	17.95	16.39	21.05	27.27
Always	8.97	9.84	0.00	0.00	3.85	6.56	0.00	0.00
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	2.74	2.35	2.42	2.45	2.55	2.46	2.42	2.36
Median	3.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00
SD	1.24	1.08	1.17	0.93	1.20	1.31	1.17	1.36
% Usefulness								
Not at all useful	7.46	9.80	17.65	22.22	5.17	11.63	23.08	14.29
Not very useful	16.42	21.57	29.41	11.11	17.24	13.95	15.38	0.00
Somewhat useful	40.30	33.33	23.53	44.44	46.55	41.86	53.85	42.86
Very useful	26.87	21.57	29.41	22.22	24.14	25.58	7.69	42.86
Extremely useful	8.96	13.73	0.00	0.00	6.90	6.98	0.00	0.00
Overall statistics								
N	67	51	17	9	58	43	13	7
Mean	3.13	3.08	2.65	2.67	3.10	3.02	2.46	3.14
Median	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
SD	1.04	1.18	1.12	1.12	0.95	1.08	0.97	1.07

Table C-5: Use and Usefulness of AWA for Diagnosing Writing Deficiencies by Program Type

Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	21.79	24.59	15.79	36.36	35.90	39.34	26.32	54.55
Rarely	21.79	22.95	47.37	18.18	15.38	16.39	36.84	9.09
Sometimes	32.05	29.51	15.79	27.27	32.05	31.15	31.58	9.09
Frequently	15.38	14.75	5.26	9.09	15.38	11.48	5.26	18.18
Always	8.97	8.20	15.79	9.09	1.28	1.64	0.00	9.09
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	2.68	2.59	2.58	2.36	2.31	2.20	2.16	2.18
Median	3.00	3.00	2.00	2.00	2.00	2.00	2.00	1.00
SD	1.23	1.24	1.31	1.36	1.15	1.14	0.90	1.54
% Usefulness								
Not at all useful	6.25	6.67	12.50	25.00	10.34	11.63	33.33	16.67
Not very useful	20.31	24.44	37.50	12.50	6.90	16.28	6.67	16.67
Somewhat useful	46.88	44.44	37.50	37.50	60.34	51.16	53.33	16.67
Very useful	17.19	13.33	12.50	12.50	20.69	18.60	6.67	33.33
Extremely useful	9.38	11.11	0.00	12.50	1.72	2.33	0.00	16.67
Overall statistics								
N	64	45	16	8	58	43	15	6
Mean	3.03	2.98	2.50	2.75	2.97	2.84	2.33	3.17
Median	3.00	3.00	2.50	3.00	3.00	3.00	3.00	3.50
SD	1.01	1.06	0.89	1.39	0.88	0.95	1.05	1.47

Table C-6: Use and Usefulness of AWA for Determining Pre-Enrollment Training by Program Type								
Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	44.87	54.10	63.16	54.55	55.13	65.57	63.16	54.55
Rarely	15.38	21.31	15.79	18.18	10.26	9.84	15.79	18.18
Sometimes	24.36	13.11	15.79	9.09	24.36	18.03	10.53	9.09
Frequently	12.82	8.20	0.00	9.09	8.97	4.92	5.26	9.09
Always	2.56	3.28	5.26	9.09	1.28	1.64	5.26	9.09
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	2.13	1.85	1.68	2.00	1.91	1.67	1.74	2.00
Median	2.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.20	1.14	1.11	1.41	1.13	1.04	1.20	1.41
% Usefulness								
Not at all useful	30.19	40.00	60.00	28.57	29.79	30.30	66.67	20.00
Not very useful	11.32	20.00	10.00	28.57	14.89	30.30	11.11	20.00
Somewhat useful	43.40	25.71	10.00	28.57	46.81	24.24	11.11	40.00
Very useful	13.21	11.43	20.00	0.00	6.38	12.12	11.11	0.00
Extremely useful	1.89	2.86	0.00	14.29	2.13	3.03	0.00	20.00
Overall statistics								
N	53	35	10	7	47	33	9	5
Mean	2.45	2.17	1.90	2.43	2.36	2.27	1.67	2.80
Median	3.00	2.00	1.00	2.00	3.00	2.00	1.00	3.00
SD	1.12	1.18	1.29	1.40	1.05	1.13	1.12	1.48

Table C-7: Use and Usefulness of AWA for Placing Students into Courses by Program Type

Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	65.38	67.21	78.95	54.55	67.95	68.85	78.95	63.64
Rarely	20.51	19.67	15.79	18.18	11.54	13.11	0.00	9.09
Sometimes	5.13	3.28	0.00	9.09	16.67	14.75	15.79	9.09
Frequently	5.13	4.92	0.00	9.09	2.56	1.64	0.00	9.09
Always	3.85	4.92	5.26	9.09	1.28	1.64	5.26	9.09
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	1.62	1.61	1.37	2.00	1.58	1.54	1.53	1.91
Median	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD	1.06	1.10	0.96	1.41	0.95	0.92	1.12	1.45
% Usefulness								
Not at all useful	40.91	41.67	70.00	28.57	40.48	34.38	55.56	20.00
Not very useful	25.00	30.56	10.00	28.57	30.95	43.75	44.44	40.00
Somewhat useful	25.00	19.44	20.00	14.29	19.05	12.50	0.00	0.00
Very useful	6.82	5.56	0.00	14.29	7.14	6.25	0.00	20.00
Extremely useful	2.27	2.78	0.00	14.29	2.38	3.13	0.00	20.00
Overall statistics								
N	44	36	10	7	42	32	9	5
Mean	2.05	1.97	1.50	2.57	2.00	2.00	1.44	2.80
Median	2.00	2.00	1.00	2.00	2.00	2.00	1.00	2.00
SD	1.08	1.06	0.85	1.51	1.06	1.02	0.53	1.64

Table C-8: Use and Usefulness of AWA for Advising on a Career Path by Program Type								
Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	78.21	72.13	68.42	54.55	82.05	78.69	78.95	63.64
Rarely	19.23	22.95	21.05	27.27	14.10	14.75	15.79	9.09
Sometimes	2.56	4.92	10.53	18.18	3.85	6.56	5.26	27.27
Frequently	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Always	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	1.24	1.33	1.42	1.64	1.22	1.28	1.26	1.64
Median	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD	0.49	0.57	0.69	0.81	0.50	0.58	0.56	0.92
% Usefulness								
Not at all useful	63.16	56.67	63.64	42.86	62.16	56.67	77.78	40.00
Not very useful	34.21	36.67	36.36	28.57	32.43	33.33	22.22	40.00
Somewhat useful	2.63	6.67	0.00	28.57	5.41	10.00	0.00	20.00
Very useful	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Extremely useful	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall statistics								
N	38	30	11	7	37	30	9	5
Mean	1.39	1.50	1.36	1.86	1.43	1.53	1.22	1.80
Median	1.00	1.00	1.00	2.00	1.00	1.00	1.00	2.00
SD	0.55	0.63	0.51	0.90	0.60	0.68	0.44	0.84

Table C-9: Use and Usefulness of AWA for Planning Programs and Courses by Program Type

Use/Usefulness	Scores				Essays			
	FT	PT	EMBA	DOC	FT	PT	EMBA	DOC
% Use								
Never	61.54	63.93	73.68	63.64	70.51	70.49	68.42	63.64
Rarely	32.05	31.15	21.05	9.09	21.79	21.31	26.32	9.09
Sometimes	6.41	4.92	5.26	27.27	6.41	6.56	5.26	18.18
Frequently	0.00	0.00	0.00	0.00	1.28	1.64	0.00	9.09
Always	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall statistics								
N	78	61	19	11	78	61	19	11
Mean	1.45	1.41	1.32	1.64	1.38	1.39	1.37	1.73
Median	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
SD	0.62	0.59	0.58	0.92	0.67	0.69	0.60	1.10
% Usefulness								
Not at all useful	51.11	52.94	70.00	16.67	42.50	46.67	66.67	20.00
Not very useful	31.11	32.35	20.00	50.00	37.50	33.33	22.22	20.00
Somewhat useful	15.56	11.76	10.00	33.33	20.00	20.00	11.11	60.00
Very useful	2.22	2.94	0.00	0.00	0.00	0.00	0.00	0.00
Extremely useful	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Overall statistics								
N	45	34	10	6	40	30	9	5
Mean	1.69	1.65	1.50	2.17	1.78	1.73	1.44	2.40
Median	1.00	1.00	1.00	2.00	2.00	2.00	1.00	3.00
SD	0.82	0.81	0.97	0.75	0.77	0.79	0.73	0.89

© 2007 Graduate Management Admission Council® (GMAC®). All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, distributed or transmitted in any form by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of GMAC®. For permission contact the GMAC® legal department at legal@gmac.com.

Creating Access to Graduate Business Education®, GMAC®, GMAT®, Graduate Management Admission Council®, and Graduate Management Admission Test® are registered trademarks of the Graduate Management Admission Council® in the United States and other countries.

附錄三、大會邀請學者演講之講義

Designing Item Pools to Optimize the Functioning of a CAT

Mark D. Reckase
Michigan State University

2007 GMAC® Conference on Computerized Adaptive Testing



Computerized Adaptive Test

- The basic design of a CAT is well documented and researched.
 - Item Selection Rule
 - Proficiency Estimation Procedure
 - Stopping Rule
- It is clear that an item pool is needed for the CAT to function.
- There is little in the research literature on CAT about the desired characteristics of an item pool.

2007 GMAC® Conference on Computerized Adaptive Testing



Optimal Item Pool

- An optimal item pool is one that always has an item available for selection that matches the characteristics specified by the item selection rule.
 - Maximum information example with the Rasch model
 - An item is always available that is equal to the current proficiency estimate
- The characteristics of the optimal pool are dependent on the item selection rule, the stopping rule, and the examinee population.

2007 GMAC® Conference on Computerized Adaptive Testing



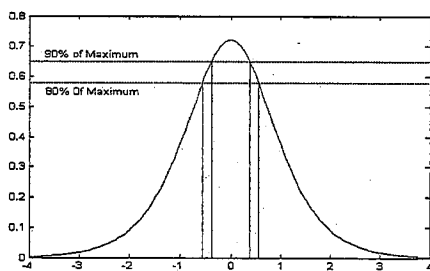
Approximately Optimal Pool

- Optimal item pools are prohibitively large because items are needed for every possible proficiency estimate.
- But, the results from items with b -parameters 1.00 or 1.01 are likely indistinguishable.
- Therefore modify the definition of an optimal pool to be one that has items within a small range of the item requested by the item selection rule.

2007 GMAC® Conference on Computerized Adaptive Testing



Rasch Item Information Function



2007 GMAC® Conference on Computerized Adaptive Testing



Range of b -parameters for the Maximum Information Criterion

- The figure shows the item information function for the Rasch model with $D = 1.7$.
- The approximate width of the θ -scale that has .9 or more of the maximum information is from -.4 to .4.
 - The range for .8 of the maximum is -.6 to .6
 - The range for .95 of the maximum is -.3 to .3
- The approximate optimal pool has items within the selected range of the current θ estimate.

2007 GMAC® Conference on Computerized Adaptive Testing



Optimal Items for One Examinee

$\theta = -1$

- Maximum Information Item Selection
- Maximum Likelihood Ability Estimation
 - Fixed Step size of .7 until both 0 and 1 item scores are present
- Fixed Length 20-Item Test

2007 GMAC® Conference on Computerized Adaptive Testing



Optimal Items for One Examinee

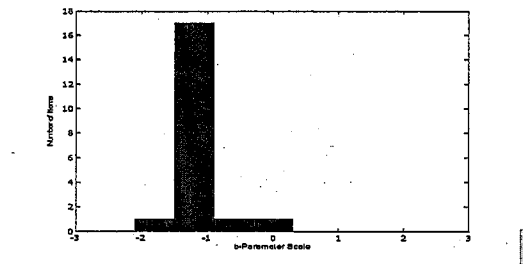
$\theta = -1$

Item Number	b-parameter	Item Score
1	0	0
2	-0.7000	0
3	-1.4000	1
4	-1.1986	1
5	-1.5924	1
6	-1.2914	1
7	-1.0607	1
8	-1.2537	1
9	-1.0178	1
10	-1.2332	1
11	-1.1037	1
12	-1.2204	1
13	-1.1142	1
14	-1.2116	1
15	-1.1217	1
16	-1.2052	1
17	-1.1272	1
18	-1.0938	0
19	-1.1229	0
20	-1.1880	1

2007 GMAC® Conference on Computerized Adaptive Testing



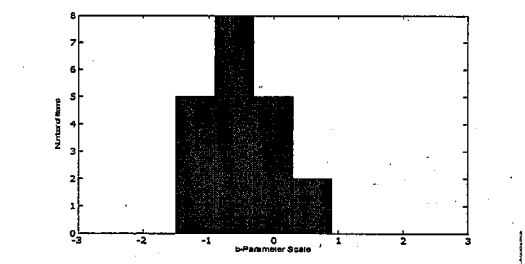
Approximately Optimal Items for $\theta = -1$, Bin Size = .6



2007 GMAC® Conference on Computerized Adaptive Testing



Approximately Optimal Items for $\theta = -.5$, Bin Size = .6



2007 GMAC® Conference on Computerized Adaptive Testing



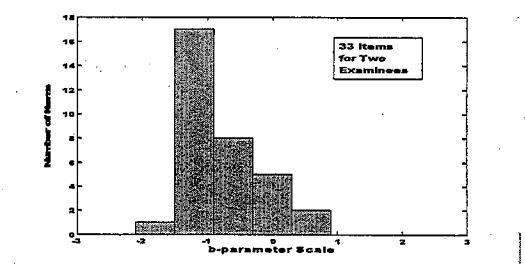
Approximately Optimal Pool for Two Examinees with $\theta = -1, .5$

- Each examinee needs twenty items matched to the current estimate of θ .
- But, if the test is low stakes, and the items are added in as needed for each examinee, the items for Examinee 1 can be used for Examinee 2.
- The approximately optimal pool for the two examinees is the first set plus the unique items in the second set.
- This is the union of the two sets – 33 instead of 40 items needed.

2007 GMAC® Conference on Computerized Adaptive Testing



Approximately Optimal Pool for Two Examinees with $\theta = -1, .5$



2007 GMAC® Conference on Computerized Adaptive Testing



Determining the Approximately Optimal Pool for a CAT

- Specify the CAT procedure.
- Simulate the selection of examinees from the expected population.
- Determine the approximately optimal pool (AOP) for each examinee.
- Find the union of the AOPs.
- If the union of AOPs is formed sequentially after each examinee is sampled, the number of items will asymptote to the AOP pool size.

2007 GMAC® Conference on Computerized Adaptive Testing



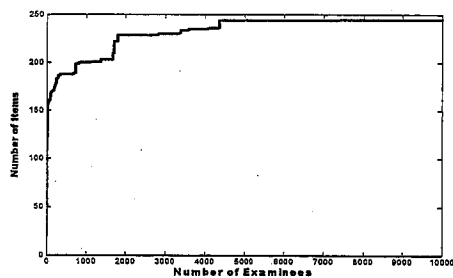
Example – Rasch CAT

- CAT Procedures
 - Max information
 - Maximum likelihood
 - 20 item test length
- Expected examinee population – $N(0,1)$
- Expected size of examinee sample – 10,000

2007 GMAC® Conference on Computerized Adaptive Testing



Number of Items Needed in Pool



2007 GMAC® Conference on Computerized Adaptive Testing



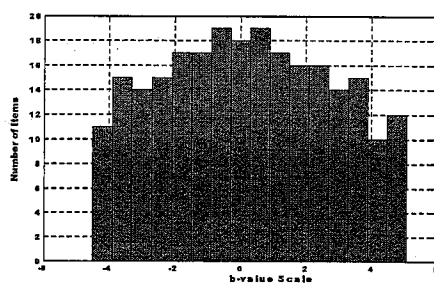
Number of Items Need in the Pool

- The number of items increases quickly with the number of examinees.
- The number is dependent on the test length, bin size and the range of θ .
 - Range of θ -- -3.8 to 4.2
 - Bin size .6
 - Test Length 20
- Pool Size = 245

2007 GMAC® Conference on Computerized Adaptive Testing



Distribution of Item b -parameters



2007 GMAC® Conference on Computerized Adaptive Testing



Distribution of Item b -parameters

- b -parameters range from -4.4 to 4.9.
- The distribution over bins is far from normal.
- The distribution is roughly symmetric around 0.
- Tails are much heavier than normal because a minimum of items are needed for extreme examinees.
- No bin has 20 items.

2007 GMAC® Conference on Computerized Adaptive Testing



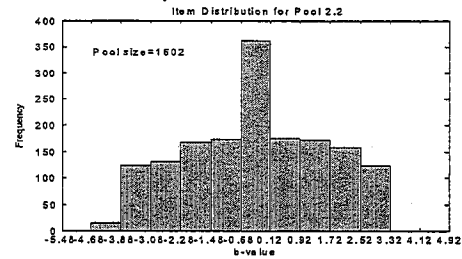
Real Example – Certification/Licensure Exam

- Rasch Model;
- Maximum information item selection;
- Bayesian/maximum likelihood estimation;
- Content balancing;
- 8 content areas
- First 60
- 61 to 250
- Exposure control procedure;
- Randomly selecting one of 15 items that closest to the desired item – the item with the same b-value as the proficiency estimate;
- Test termination rule;
- Variable-length test ranging from 60 to 250 items;
- Stop when the confidence interval around the proficiency estimate no longer contains the cut score of -.28;

2007 GMAC® Conference on Computerized Adaptive Testing



Real Example – Certification/Licensure Exam



2007 GMAC® Conference on Computerized Adaptive Testing



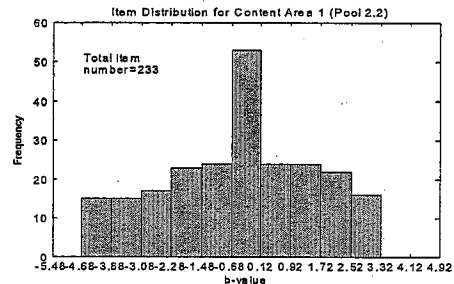
Real Example – Certification/Licensure Exam

- Need at least 15 items per bin for exposure control.
- Need many more items near cut score to provide enough appropriate items for the maximum length test.
- Total item pool is sum of item pools for each content area.

2007 GMAC® Conference on Computerized Adaptive Testing



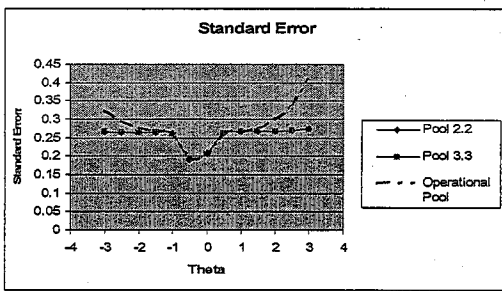
Content Area Pool



2007 GMAC® Conference on Computerized Adaptive Testing



Comparison to the Operational Pool



2007 GMAC® Conference on Computerized Adaptive Testing



Another Example – Two-stage Test

- Fixed length test of length N items.
- First stage has items concentrated at the cut score with k items.
- Two second-stage tests with $N - k$ items in each.
- What is the best combination of test lengths for the two stages?
- How should the item difficulties be distributed in the second stage?

2007 GMAC® Conference on Computerized Adaptive Testing



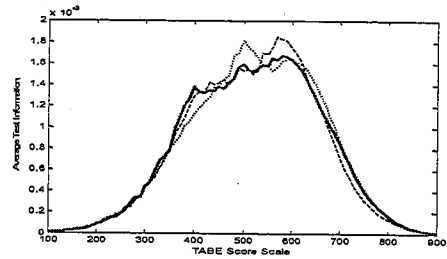
Two-stage Test

- Select θ from Expected Distribution
- Simulate stage one test and classify is high or low
- Select items equal to true θ for the second stage test until target information is reached.
- Continue until second stage of desired length is obtained.
- Replicate and average to smooth out sampling variation.

2007 GMAC® Conference on Computerized Adaptive Testing



Information Functions for Three Test Length Combinations



2007 GMAC® Conference on Computerized Adaptive Testing



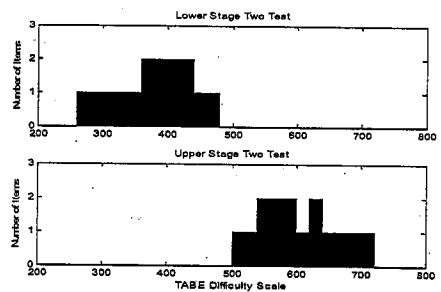
Test Lengths

- Considered four combinations
 - 4 and 16
 - 5 and 15
 - 6 and 14
- Green curve is 5 and 15, a reasonable compromise between amount of information and spread of information.

2007 GMAC® Conference on Computerized Adaptive Testing



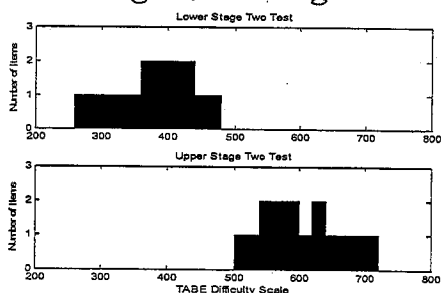
Second Stage Tests – 15 Items



2007 GMAC® Conference on Computerized Adaptive Testing



Second Stage 6/18 Design



2007 GMAC® Conference on Computerized Adaptive Testing



Second Stage Tests

- Tests are not symmetric because 3pl items used.
- Item sets skewed in opposite directions.
- No overlap because rounded average of items over replications less than .5 for overlapping regions.
- ¼ to ¾ split between first and second stages seem best.

2007 GMAC® Conference on Computerized Adaptive Testing



Major Findings

- The simulation approach to item pool design works very well.
- The method can be used whenever it is possible to specify the testing procedure in sufficient detail to develop an accurate simulation.
- The results are dependent on test design and the assumed examinee distribution.

2007 GMAC® Conference on Computerized Adaptive Testing



New Directions

- Use of procedure with the 3pl model is challenging.
 - Optimal items have infinite discrimination and zero guessing.
 - Need to model possible item pools rather than optimal item pools.
 - Need to model the correlations between item parameters.
- See paper by Lixiong Gu later in this conference.
- More operational tests are needed.

2007 GMAC® Conference on Computerized Adaptive Testing



FUMIKO SAMEJIMA, UNIVERSITY OF TENNESSEE (keynote speaker)

2007 GMAC CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

June 7, 2007 at Radisson University Hotel, Minneapolis, MN

[SUMMARY OF ACTUAL PROCEDURE]

- (1) Selection of a set of **300 core items** (Old Test) from the **2,131** dichotomous items that **LSAC** previously administered and estimated their item parameters presuming the three-parameter logistic model (**3PL**). These core items become the **foundation** of the online item calibration, so effort was made to select items that provide **large** amounts of **test information** for a **wide** range of ability.

First, all items whose estimated third parameters (**guessing parameter**) are **0.2** or **greater** were **discarded**, and the remaining **1,452** items were considered. **[TABLE1]

Second, because the first two parameters in **3PL** are **no longer** the discrimination and difficulty indices, the **ICF** of each item is approximated by the **logistic model (2PL)** for the interval of θ **higher** than the **critical value** below which the basic function does not decrease monotonically, or item response information function for the correct answer assumes **negative** values. **[FIG.1a,b], [FIG.2]

Depending on the truncated 2PL thus obtained, 300 items whose difficulty parameters were distributed **evenly** for a wide range of ability and having **high** discrimination power were selected as core items.

- (2) Obtaining the **Maximum Likelihood Estimate (MLE)** of θ (ability) for each of **1,202** hypothetical **examinees** whose ability distributes **uniformly** for the interval of θ , **(-3.0, 3.0)**, following the **truncated 2PL**, (*the model that also avoid multi-modal likelihood functions*) in **Computerized Adaptive Testing (CAT)**, using the set of core items as the tempool. (**Concurrently 25 new, target items were presented to each hypothetical examinee non-adaptively, scored 1 (correct) or 0 (incorrect) and kept separately until Step (8).**) **[FIG.3a,b]

The program was written in such a way that the amount of **item information** below the critical value (Samejima, 1973) is **set equal** to **0** for each core item, so that the item will **never** be presented when the examinee's current MLE of θ is **below the critical value**.

Five stopping rules were used, that is, presentation of new items are stopped when the estimated **standard error of estimation**, i.e., the **reciprocal** of the **square root of the test information function** at the **current MLE** of θ , exceeded **0.25, 0.32, 0.40, 0.50**, respectively, and also regardless of the estimated standard error when the number ****[FIG.4a,b]** of presented core items reached **40 for the sake of comparisons**.

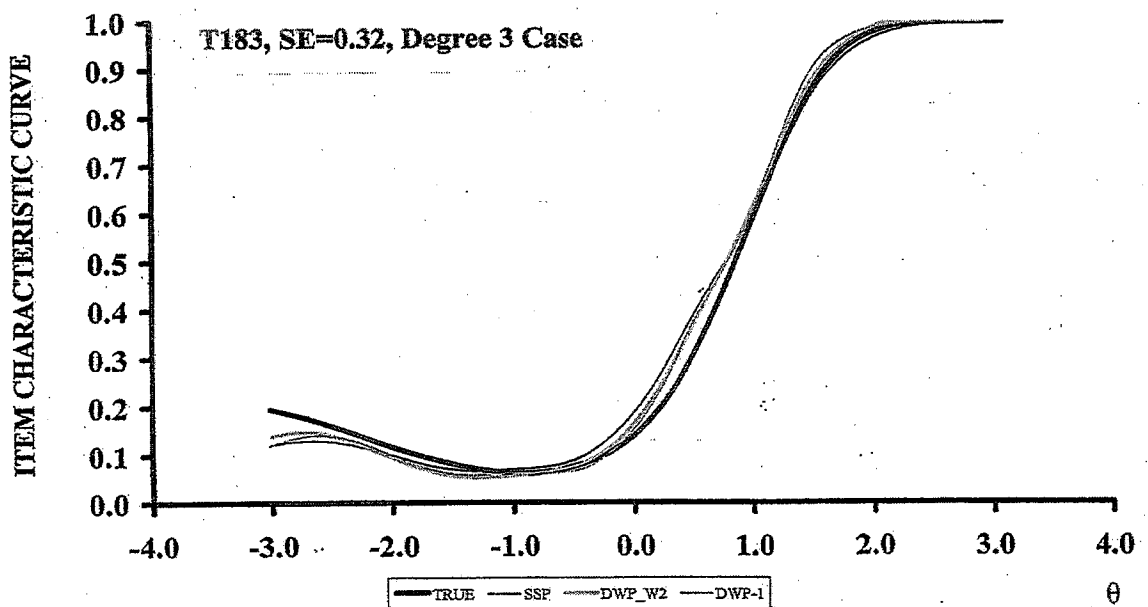
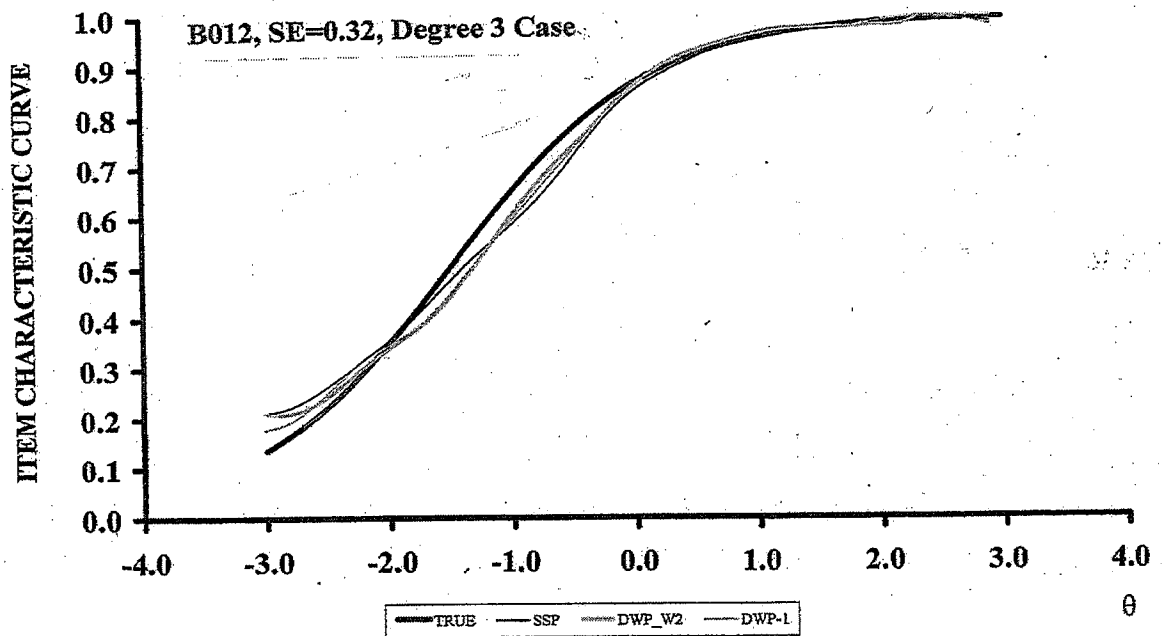
- (3) **Transformation** of θ to τ that has a **constant** test information function.

- (4) Transformation of the **MLE** of θ to the **MLE** of τ for each of the **1,202** examinees, using the transformation formula obtained in (3).
- (5) Approximation to the **density function** of the **MLE** of τ by fitting a **polynomial** of degree 3 or 4 using the **method of moments** whose outcome equals the **least squared solution**. (Degree 3 and 4 Cases.) **[FIG.5a,b]
- (6) Computing the **first through fourth conditional moments** of τ , given its **MLE**, for each of the **1,202** examinees, using the outcomes of (5). **[TABLE 1]
**[TABLE 2]
- (7) Using the outcomes of (6), assignment of one of the Peason's distributions to the **conditional distribution** of τ , given its **MLE**, for each of the **1,202** examinees, using **Peason's criteria**.
- (8) Estimation of the **ICF** for each of **25 target items**, of which **15** follow **3PL** and **10** are **non-monotonic** curves, using the **Simple Sum Procedure (SSP)**. **[FIG.6a-d]
**[FIG.7a-h]
- (9) Computing the **Differential weight function (DWF)** for each target item using the outcome of (8).
- (10) Estimation of the **ICF** for each of the **25 target items** by the **Differential Weight Procedure (DWP)** using the **DWF** obtained by (9).
- (11) Computing the **Criterion ICF** for each of the **25 target items** by the **DWP** introduced in (10), using the **True ICF** as the **DWF** on the outcome of **SSP** in (8).
- (12) The outcomes of the **SSP** in (8) and **DWP** in (10) for each **target item** are compared with its **Criterion ICF** in (11).
- (13) If the **Criterion ICF** turns out to be **very close** to the **True ICF** and the outcomes of **SSP** and **DWP** are **very close** to the **Criterion ICF** for every target item, our conclusion will be that the whole procedures have **successfully** estimated the **True ICF**. **If not**, we will try to **improve** one or more steps of the procedures to **increase** in accuracy of estimation.

[RESULTS & DISCUSSION]

Online item calibration successfully produced estimated ICFs that are close enough to the true ICFs, using truncated 2PL for the core items, detecting nonmonotonicity of ICFs very well. A strength of this nonparametric approach may be the use of the conditional p.d.f. of τ , given its MLE, that prohibits errors in the estimated ability from affecting the OPC estimation of target items, *device that does not exist in many other OPC estimation methods*. The present research outcomes also provided us with several useful suggestions for future research.

Two Examples of the nonparametrically calibrated item characteristic functions of the target items B012 and T183 selected out of 25 target items. (The thick, lines are the true ICFs.) Note, especially, that not only these estimated ICFs are close to the true ICF in each graph, *monotonicity is well detected in the second graph, accuracy that only nonparametric methods can achieve.*



Designing Templates for Innovative Item Types

Handwritten signature

Cynthia G. Parshall, Ph.D.
Measurement Consultant

2007 GMAC[®] Conference on Computerized Adaptive Testing



Introduction to Templates

2007 GMAC[®] Conference on Computerized Adaptive Testing



Why Templates?

1. What are they?
2. What do they do for you?

2007 GMAC[®] Conference on Computerized Adaptive Testing



What Templates Are

- Templates are a structured means of collecting/storing item data.
- Possible aspects to template design:
 1. Database fields
 2. Screen layout elements
 3. More specificity in either of the above (i.e., sub-template)

2007 GMAC[®] Conference on Computerized Adaptive Testing



Sample Template – Database Fields

Item ID #:	Keywords:
Author:	Reviewer:
Instructions:	Reference:
Prompt:	
Graphic file name:	
Correct area(s):	
Incorrect area(s):	

2007 GMAC[®] Conference on Computerized Adaptive Testing



Sample Template – Screen Layout Elements

Task Description:	
Instructions:	Simulated software program:
Navigation Options:	

2007 GMAC[®] Conference on Computerized Adaptive Testing



Sample Template – Sub-Template

Word Processing Task Description:

Instructions:

Simulated word processing program:

Navigation Options:

2007 GMAC® Conference on Computerized Adaptive Testing



What Templates Can Do

- Templates are used in the item writing process. They help provide:
 - *Structure* – they help guide and constrain item writing
 - *Efficiency* – they can save time & money in programming efforts and media development
 - *Security* – they can help replace an “exposed” items with a close *variant*

2007 GMAC® Conference on Computerized Adaptive Testing



与学習 Taxonomy for Innovative Items

2007 GMAC® Conference on Computerized Adaptive Testing



Original Taxonomy

- Item format
- Interactivity
- Response action
- Media inclusion
- Scoring algorithm

(From Parshall, Davey, and Pashley, 2000)

2007 GMAC® Conference on Computerized Adaptive Testing



Overview of Taxonomy for Innovative Assessments

- Assessment structure
- Complexity
- Fidelity
- Interactivity
- Response action
- Media inclusion
- Scoring algorithm

(based on Harmes & Parshall, 2005)

2007 GMAC® Conference on Computerized Adaptive Testing



Assessment Structure

- Item format & beyond
 - from *discrete items*,
 - through *situated tasks*,
 - to *simulated environments*
- Consider:
 - Technology-oriented terms
 - e.g., ‘hotspot’
 - Measurement-oriented terms
 - e.g., ‘selected figural response’

2007 GMAC® Conference on Computerized Adaptive Testing



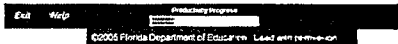
Examples of Discrete Item Types

- Example of various assessment structures:
 - Selected figural responses (hot spot)
 - Constructed figural responses (drag-and-drop)

2007 GMAC® Conference on Computerized Adaptive Testing



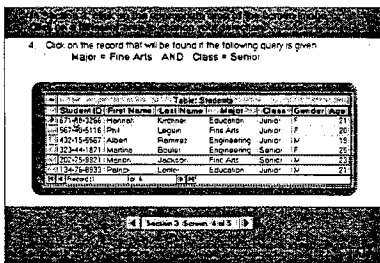
Example: Selected Figural Response



2007 GMAC® Conference on Computerized Adaptive Testing



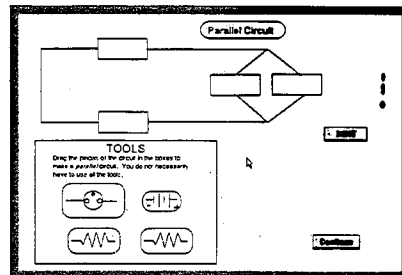
Example: Selected Figural Response



2007 GMAC® Conference on Computerized Adaptive Testing



Example: Constructed Figural Response



2007 GMAC® Conference on Computerized Adaptive Testing



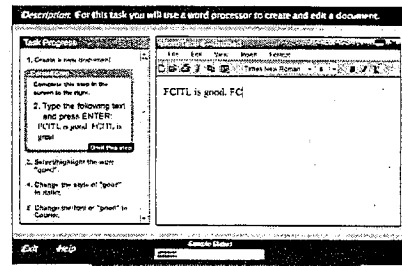
Examples of Situated Tasks

- An integrated, realistic situation or scenario in which examinees are asked to solve a problem or task, typically through a series of actions or steps
- Can be designed to progress as either:
 - Structured
 - Example: Word processing task
 - Unstructured
 - Example: Architectural design problems

2007 GMAC® Conference on Computerized Adaptive Testing



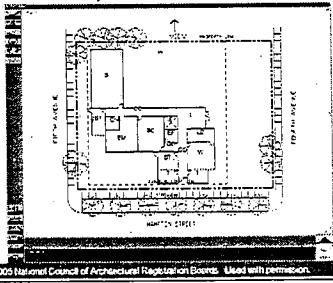
Example: Word Processing Task (Structured)



2007 GMAC® Conference on Computerized Adaptive Testing



Example: Architectural Design Problem (Unstructured)



©2005 National Council of Architectural Registration Boards. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Examples of Simulated Environments

- The simulated environment assessment is presented in a realistic *setting*, and replicates the entire *experience* of interacting in the real environment
- Examples:
 - Flight simulators
 - Information Technology (IT) testing

2007 GMAC® Conference on Computerized Adaptive Testing



Complexity

- The number and variety of screen elements that an examinee *needs to interpret*
 - Conceptual: figuring out the task (e.g., multiple headers, text boxes, portions of graphics that appear clickable, etc.)
- The number of onscreen elements that an examinee *can use*
 - Functional: accessing information and providing a response (e.g., response options, audio/video player, graphical tools, test navigation buttons, etc.)

2007 GMAC® Conference on Computerized Adaptive Testing



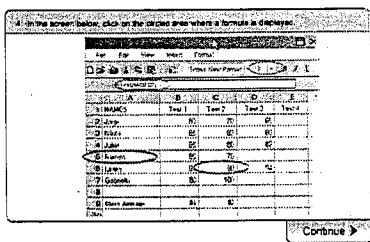
Examples of Complexity

- *Low*
 - Traditional 4-option multiple choice
- *Moderate*
 - Select element from a spreadsheet
- *High*
 - CPA exam forms completion

2007 GMAC® Conference on Computerized Adaptive Testing



Example of Complexity: Moderate



©2005 Florida Department of Education. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Fidelity 保真度

- The degree to which the assessment provides
 - a realistic and accurate reproduction of the actual actions, tasks, and environments that are part of the construct being measured
- The closer the assessment approximates the actual construct being measured, the higher the fidelity

2007 GMAC® Conference on Computerized Adaptive Testing



Examples: Fidelity

Fidelity relates to the nature of the construct.

- **Low**
 - Text-based scenario descriptions with multiple-choice
- **Moderate**
 - Video-enhanced oral communications assessment
- **High**
 - CPA exam forms completion
 - Send and reply email

2007 GMAC® Conference on Computerized Adaptive Testing



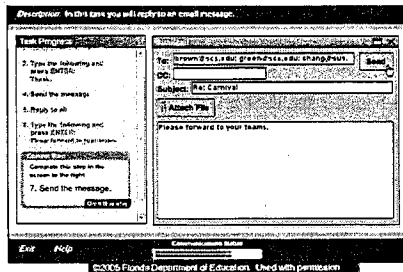
Example of Fidelity: Moderate



2007 GMAC® Conference on Computerized Adaptive Testing



Example of Fidelity: High



2007 GMAC® Conference on Computerized Adaptive Testing



Interactivity

- The degree to which an assessment responds to examinee actions
- Interactivity is increased in an assessment when there are:
 - Multiple cycles of examinee input and assessment reactions

2007 GMAC® Conference on Computerized Adaptive Testing



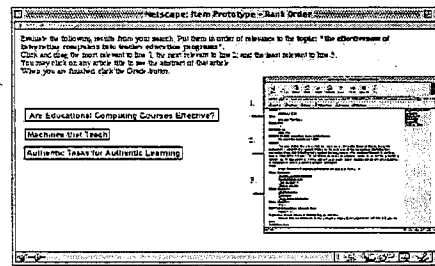
Examples: Interactivity

- **Low**
 - Click an on-screen figure, which displays a single, contextualized result
 - Example: English 'editing' task
- **Moderate**
 - Two or three iterations of examinee/computer interactions
 - Example: Library skills ordered response item
- **High**
 - Potential for a series of examinee/computer iterations
 - Example: Interior design task from NCARB

2007 GMAC® Conference on Computerized Adaptive Testing



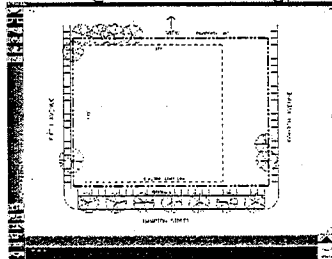
Example of Interactivity: Moderate



2007 GMAC® Conference on Computerized Adaptive Testing



Example of Interactivity: High (1. Initial Background Drawing)

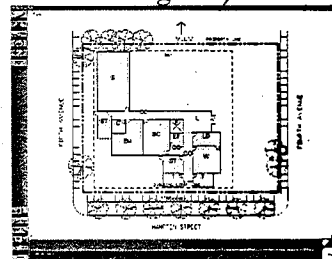


©2005 National Council of Architectural Registration Boards. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Example of Interactivity: High (2. Solution in Progress)

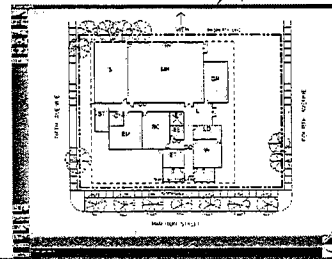


©2005 National Council of Architectural Registration Boards. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Example of Interactivity: High (3. Finished Floor Plan)



©2005 National Council of Architectural Registration Boards. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Response Action

- The actions the examinee is asked to take, and
- The input devices an examinee is asked to use
- Examples:
 - Keyboard
 - Mouse for selection
 - Mouse for other actions
 - Other input devices (e.g., light pen, touch screen, microphone)

2007 GMAC® Conference on Computerized Adaptive Testing



Media Inclusion

- The addition of non-text material
 - in the item stem and/or response options
- Examples of media:
 - graphics
 - animation
 - audio
 - video

2007 GMAC® Conference on Computerized Adaptive Testing



Scoring Algorithm

- A range of scoring approaches
 - Dichotomous
 - Polytomous
 - Complex modeling
- Examples:
 - traditional MC items
 - multiple response
 - integrated, contextualized tasks (e.g. patient case management problems)

2007 GMAC® Conference on Computerized Adaptive Testing



Using the Taxonomy To Design Item Templates

2007 GMAC® Conference on Computerized Adaptive Testing



Designing Item Templates – General Principles

1. Develop a comprehensive list of **goals and requirements** for the exam program
 - e.g., content features, planned **test length**, **available funds**
2. Determine any **innovation item types** are already available
 - i.e., will programming be **necessary?**
3. Consider the **taxonomy levels** in terms of **test development concerns**
 - e.g., **construct**, **psychometrics**, **programming**, **cost**

2007 GMAC® Conference on Computerized Adaptive Testing



Example: Inventory of Teacher Technology Skills

1. **Goals and Requirements**
 - Computer skills will be measured by **performance tasks**
 - Separate skills should be presented in **context**
 - "Simulated" software should be **realistic** but "non-specific" (e.g., Mac Word, PC spreadsheet)
 - "tasks" should be highly **representational**
2. **Current availability?**
 - None; **custom programming** needed

2007 GMAC® Conference on Computerized Adaptive Testing



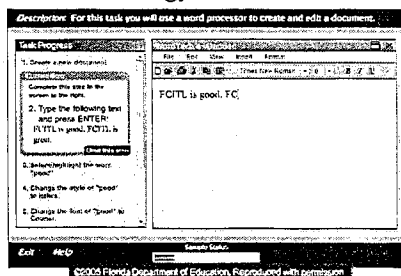
Example: Inventory of Teacher Technology Skills

- Taxonomy level targets, based on TD concerns:
- **Assessment structure** – develop "situated tasks"
 - **Complexity** – moderate complexity
 - **Fidelity** – high fidelity
 - **Interactivity** – 'flash animation' to provide an interactive context
 - **Response action** – standard input devices; standard 'technology' response actions
 - **Media inclusion** – high use of graphics; no video or audio
 - **Scoring algorithm** – dichotomous scoring

2007 GMAC® Conference on Computerized Adaptive Testing



Example: Inventory of Teacher Technology Skills



2007 GMAC® Conference on Computerized Adaptive Testing



Implications for Exam Program Construct

- **Increased innovation** is often sought for better measurement of the construct in all areas of the taxonomy:
 - assessment structure**
 - complexity**
 - fidelity**
 - interactivity**
 - response action**
 - media inclusion**
 - scoring algorithm**

2007 GMAC® Conference on Computerized Adaptive Testing



Implications for Psychometrics

- There is less information about the psychometric functioning of innovative items
 - across all levels of the taxonomy.
- Additional psychometric development or analysis may be needed for:
 - Complexity, interactivity, some media uses, and polytomous scoring

2007 GMAC® Conference on Computerized Adaptive Testing



Implications for Programming Needs

- Additional programming may be needed for many levels of the taxonomy
 - e.g., assessment structure, complexity, interactivity, response action

2007 GMAC® Conference on Computerized Adaptive Testing



Implications for Examinee Computer Skills

- More innovation at some levels of the taxonomy will be dependent upon higher levels of examinees' computer skills
 - e.g., complexity, interactivity, response action, and some uses of media

2007 GMAC® Conference on Computerized Adaptive Testing



Implications for Cost

- Most levels of the taxonomy can potentially increase initial costs and/or ongoing costs
 - e.g., assessment structure, complexity, interactivity, response action, media inclusion, scoring algorithm
- These additional costs are often related to:
 - programming, psychometric development, and media sourcing/development

2007 GMAC® Conference on Computerized Adaptive Testing



Example: Inventory of Teacher Technology Skills

3. Taxonomy level targets, based on TD concerns:
- Construct
 - Psychometrics
 - Programming
 - Examinee computer skills
 - Cost

2007 GMAC® Conference on Computerized Adaptive Testing



Example: Inventory of Teacher Technology Skills

Description: In this task you will reply to an email message.

Task Progress

3. Type the following and press ENTER:
"Thanks"
4. Send the message
5. Reply to all
6. Type the following and press ENTER:
"Please forward to your teams."

Complete Step
Complete this step in the screen to the right.
7. Send the message.

Next Step

Reply

To: brown@cs.edu, green@cs.edu, chang@cs.edu. **Send**

CC:

Subject: Re: Carniva

Attach File

Please forward to your teams.

Exit Help Communications Status

©2005 Florida Department of Education. Used with permission.

2007 GMAC® Conference on Computerized Adaptive Testing



Summary

- Templates can make test development more:
 - Structured
 - Efficient
 - Secure
- The taxonomy may be helpful for guiding the design and development of templates.

2007 GMAC® Conference on Computerized Adaptive Testing



Thank you!

Cynthia G. Parshall, PhD
Measurement Consultant
cparshall@tampabay.rr.com

2007 GMAC® Conference on Computerized Adaptive Testing



Choices in CAT models in the Context of educational testing

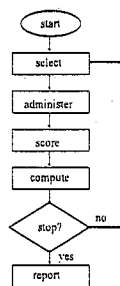
Theo Eggen
Cito, Netherlands
theo.eggen@cito.nl

2007 GMAC® Conference on Computerized Adaptive Testing



CATs at Cito

- Collection of items: IRT calibrated item bank
- Algorithm: rules for starting, selecting scoring, stopping



2007 GMAC® Conference on Computerized Adaptive Testing



Overview presentation

- CATs at Cito
- Item calibration: incomplete designs and estimation methods
- Item selection: efficient and easy

2007 GMAC® Conference on Computerized Adaptive Testing



Cito cats

- **Item response theory based**
 - 1 or 2 parameter logistic model
 - Classification: application SPRT (Wald, 1947; Reckasè, 1983)
 - Estimation: wmls (Warm, 1989)
 - Item selection: max Fisher information at current ability estimate
- **Optimal measurement under practical constraints**
 - Content control (Kingsbury & Zara, 1991)
 - Exposure control (over and under) Sympson & Hetter, 1985; Revuelta & Ponsoda, 1998
 - Difficulty control (Eggen & Verschoor, 2006)

2007 GMAC® Conference on Computerized Adaptive Testing



Calibration: High demands for CATS

- Results of calibration are fixed in the CAT algorithm and discarded are
 - standard error item parameters estimates
 - poor model fit of items
- All computations in algorithm are based on the fixed item parameters
- Possible different effects for individuals

2007 GMAC® Conference on Computerized Adaptive Testing



Calibration designs

- Incomplete designs
 - anchoring
 - stochastic structure
- Sampling (Cito) practice
 - education: no random samples of individuals
 - combination of data from more than source

2007 GMAC® Conference on Computerized Adaptive Testing



Itempool: MATHCAT teacher training

1068 items arithmetic/mathematics
 Mental arithmetic (v/n)
 Basic skills arithmetic
 Fractions, percentages, ratios
 Measurement/geometry
 Information

	No of items	Pretest 2005
Old(2000)	541	92
new	527	527
	1068	619

2007 GMAC® Conference on Computerized Adaptive Testing



Calibration design

2000	541	
<small>N=1112</small>		
2005	92+	527
<small>N=1424</small>		

2007 GMAC® Conference on Computerized Adaptive Testing



Incomplete: Multistage

β	JML	CML	MML
-2.0	-3.07 (0.09)	-2.60 (0.08)	-1.90 (0.09)
-1.0	-2.01 (0.08)	-1.71 (0.07)	-0.96 (0.08)
-0.5	-1.56 (0.08)	-1.33 (0.07)	-0.53 (0.07)
0	-0.06 (0.07)	-0.11 (0.05)	-0.08 (0.05)
0	0.07 (0.07)	0.08 (0.05)	0.02 (0.05)
0	-0.00 (0.07)	-0.05 (0.05)	-0.03 (0.05)
0.5	1.56 (0.08)	1.24 (0.07)	0.42 (0.05)
1.0	2.19 (0.08)	1.75 (0.07)	0.98 (0.08)
2.0	3.24 (0.08)	2.81 (0.09)	2.08 (0.09)
			-0.01 (0.03)
			1.04 (0.05)

2007 GMAC® Conference on Computerized Adaptive Testing



Incomplete: targeted

β	JML	CML	MML	MML (2 marginals)
-2.0	-2.40 (0.09)	-2.09 (0.08)	-1.57 (0.08)	-2.08 (0.08)
-1.0	-1.11 (0.08)	-0.96 (0.07)	-0.41 (0.07)	-0.96 (0.07)
-0.5	-0.53 (0.08)	-0.45 (0.07)	0.12 (0.07)	-0.46 (0.07)
0	-0.05 (0.07)	0.05 (0.05)	0.04 (0.05)	0.05 (0.05)
0	0.00 (0.07)	0.12 (0.05)	0.01 (0.05)	0.01 (0.05)
0	-0.00 (0.07)	-0.03 (0.05)	-0.03 (0.05)	-0.03 (0.05)
0.5	0.72 (0.08)	0.64 (0.07)	0.07 (0.07)	0.64 (0.07)
1.0	1.14 (0.08)	1.00 (0.07)	0.45 (0.07)	1.00 (0.07)
2.0	2.09 (0.08)	1.83 (0.08)	1.32 (0.08)	1.83 (0.08)
			-0.01 (0.04)	-1.03 (0.05) 0.96 (0.05)
			1.19 (0.04)	1.00 (0.05) 0.92 (0.05)

2007 GMAC® Conference on Computerized Adaptive Testing



Estimation method and design structure

	JML	CML	MML
Complete/random	Correct/biased	Correct	Correct
Multistage	Biased	Biased	Correct
Targeted	Biased	Correct	Biased/correct

Little & Rubin, 1987; Eggen, 2004

2007 GMAC® Conference on Computerized Adaptive Testing



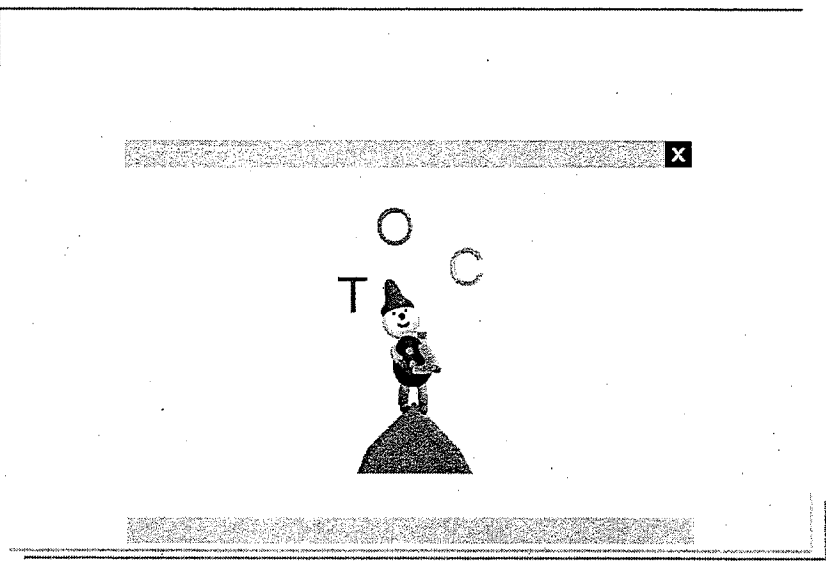
Conclusion item calibration

- Sound calibration important in CAT
- In calibration in incomplete designs in selecting the item estimation method care should be taken of the stochastic nature of the missing data
- Standard IRT calibration software has implemented only one (or two) estimation methods

update the calibration

2007 GMAC® Conference on Computerized Adaptive Testing



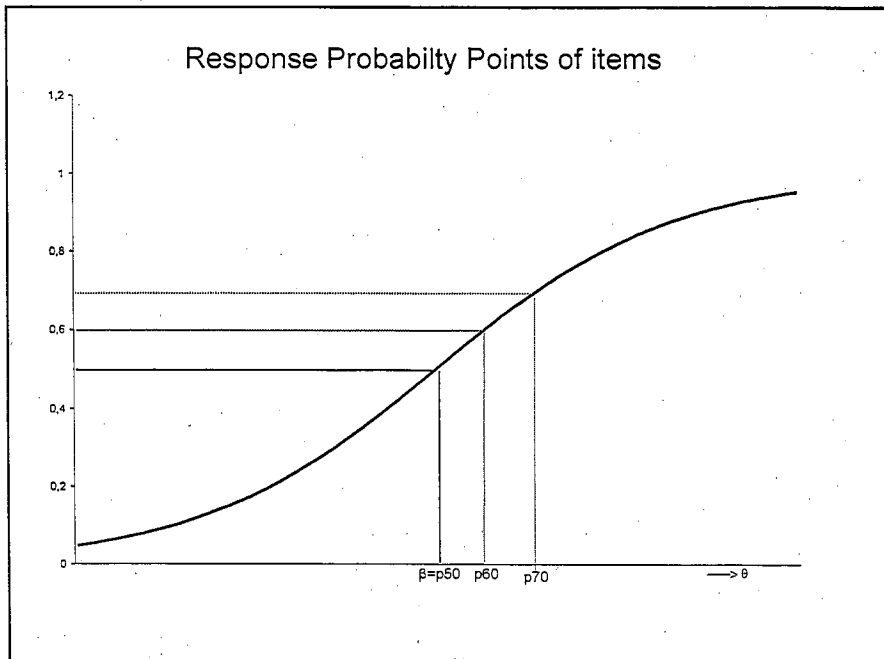


2007 GMAC® Conference on Computerized Adaptive Testing

Item selection in Cito CATS

- max Fisher information at current ability estimate
 - Content control; exposure control
- Expected success probability on each item: about 50%; expected test score: about 50% of maximum
- For young children: Can we find alternative selection methods giving easier items ?
Maintain child's self confidence and retain measurement precision





Evaluation: selection on success probability

- Simulation studies
- 4000 persons from the normal distribution
- Item bank met 300 items 2pl model
- Adaptive algorithm
 - 40 items
 - Selection methods with varying difficulty
 - Random selection and maximum information (mi) selection as benchmarks



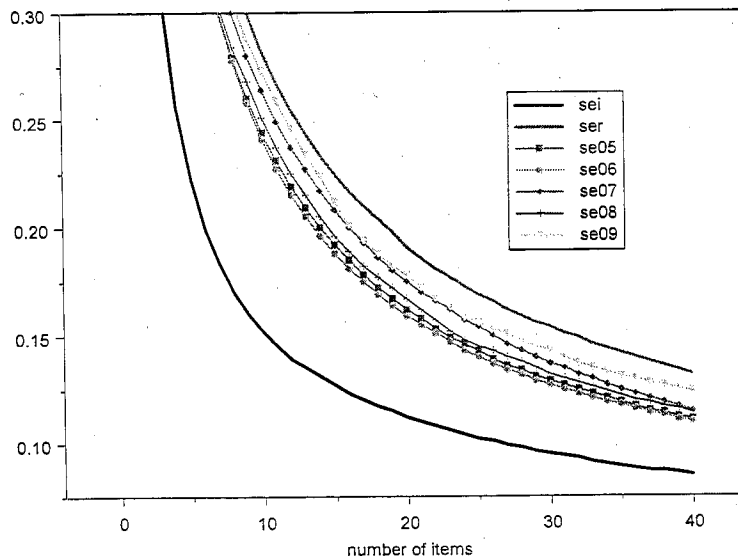
Selection on success probability

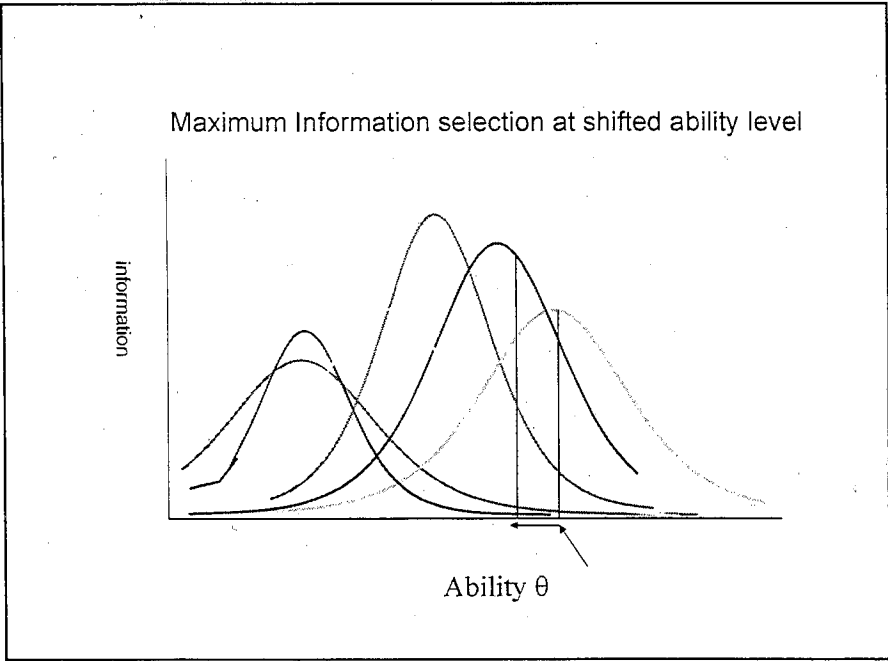
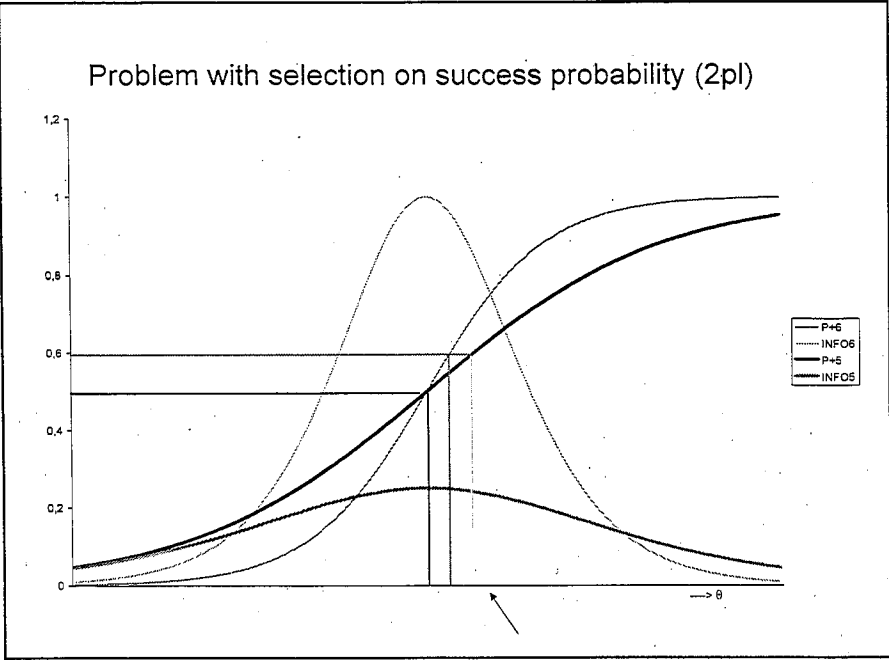
Selection	Mean se (sd)	Mean % correct (sd)
Max info	0.085 (0.013)	49.0 (11.9)
P_50	0.111 (0.008)	49.7 (8.5)
P_60	0.110 (0.009)	58.0 (9.4)
P_70	0.115 (0.015)	64.9 (11.8)
P_80	0.114 (0.018)	70.8 (15.0)
P_90	0.124 (0.033)	74.5 (17.6)
Random	0.132 (0.033)	49.7 (19.5)

2007 GMAC® Conference on Computerized Adaptive Testing

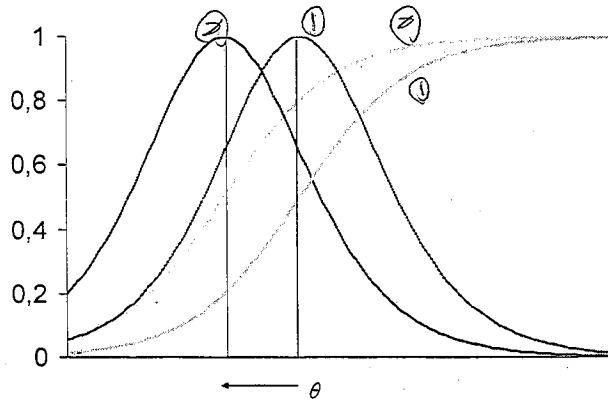


Mean se ability and test length; pp selection





Selection m_i at shifted ability level



Selection m_i at shifted ability level

$$p_i(\theta) = \frac{\exp(\alpha_i(\theta - \beta_i))}{(1 + \exp(\alpha_i(\theta - \beta_i)))} = \frac{\exp(\alpha_i \delta_i)}{(1 + \exp(\alpha_i \delta_i))}$$

$$\alpha_i \delta_i = \ln\left(\frac{p_i(\theta)}{1 - p_i(\theta)}\right)$$

$$p_i(\theta) = p^*$$

$$\delta_i^* = \frac{1}{\alpha_i} \ln\left(\frac{p^*}{1 - p^*}\right)$$

$$\theta_i^* = \hat{\theta} - \delta_i^*$$

$$I_i(\theta_i^*) = \alpha_i^2 p_i(\theta_i^*)(1 - p_i(\theta_i^*)) = \frac{\alpha_i^2 \exp(\alpha_i(\theta_i^* - \beta_i))}{(1 + \exp(\alpha_i(\theta_i^* - \beta_i)))^2}$$



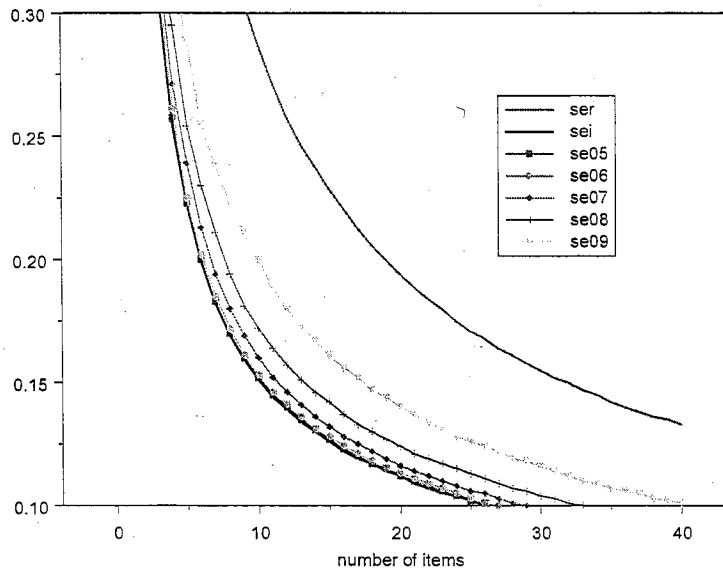
Selection mi at shifted ability level

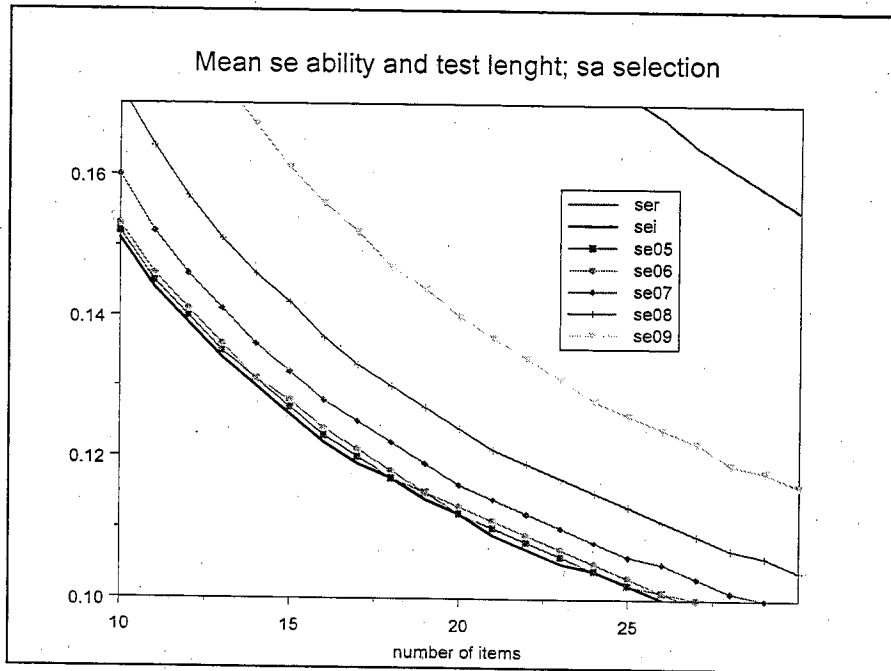
Selection	Mean se (sd)	Mean % correct (sd)
Max info	0.085 (0.011)	49.0 (12.2)
P_50	0.085 (0.013)	49.4(12.3)
P_60	0.085 (0.012)	55.1(12.2)
P_70	0.088 (0.012)	60.9(12.6)
P_80	0.092 (0.015)	65.4(14.1)
P_90	0.101 (0.015)	71.7(15.6)
Random	0.133 (0.031)	50.5(19.6)

2007 GMAC® Conference on Computerized Adaptive Testing



Mean se ability and test length; sa selection





Conclusion item selection

- With a good item bank: high measurement efficiency and easy (difficult) tests in CAT are possible
- Selection on higher success probability does work only satisfactory in the 1pl model
- Selection on max info at a shifted ability level performs satisfactory also in the 2pl



Emerging Topics

Richard M. Luecht
University of North Carolina at
Greensboro

2007 GMAC® Conference on Computerized Adaptive Testing



Three Emerging Themes in CBT

- Uses of CAT for diagnostic testing
 - Educational settings
 - Failing examinees on certification and licensure tests
 - Quality of life and patient self-report measures
- Assessment engineering as new approach to test design, development, and delivery
- Computer-based simulations and performance exercises
 - Measuring "something else"
 - Controlling method variance and other sources of variation
 - Auto-adaptive designs (AAD)

2007 GMAC® Conference on Computerized Adaptive Testing



Psychometric Rethinking About Diagnostic Information

- Useful diagnostic information must be inherently *multidimensional*
- Multidimensional diagnostic metrics should also be validated as *instructionally sensitive*
- The joint goal of test developers and psychometricians should be to embrace the quest for principled multidimensional information (PMI) by using sound *assessment engineering* practices

2007 GMAC® Conference on Computerized Adaptive Testing



Three Assertions

- Inherently unidimensional data cannot be decomposed to produce useful diagnostic scores on multiple "dimensions"
- Assessment engineering practices for item and test design/creation must be employed to produce (and reproduce over time & contexts) principled multidimensional [measurement] information (PMI)
- Idiosyncratic multidimensionality, based upon EFA and related procedures, is not easily scaled for diagnostic scoring purposes

2007 GMAC® Conference on Computerized Adaptive Testing



Example: Three Correlation Matrices

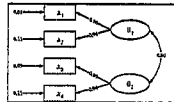
$$R_{\text{un}}^{[1]} = \begin{pmatrix} 1.00 & & & & \\ 0.89 & 1.00 & & & \\ 0.84 & 0.82 & 1.00 & & \\ 0.93 & 0.91 & 0.86 & 1.00 & \\ & & & & & 1.00 \end{pmatrix} \quad R_{\text{id}}^{[2]} = \begin{pmatrix} 1.00 & & & & & \\ 0.90 & 1.00 & & & & \\ 0.83 & 0.76 & 1.00 & & & \\ 0.64 & 0.52 & 0.84 & 1.00 & & \\ & & & & & & 1.00 \end{pmatrix}$$

Unidimensional

Idiosyncratically
Multidimensional

$$R_{\text{pm}}^{[3]} = \begin{pmatrix} 1.00 & & & & & \\ 0.90 & 1.00 & & & & \\ 0.46 & 0.45 & 1.00 & & & \\ 0.45 & 0.44 & 0.90 & 1.00 & & \\ & & & & & & 1.00 \end{pmatrix}$$

Principled Multidimensional
Information



2007 GMAC® Conference on Computerized Adaptive Testing



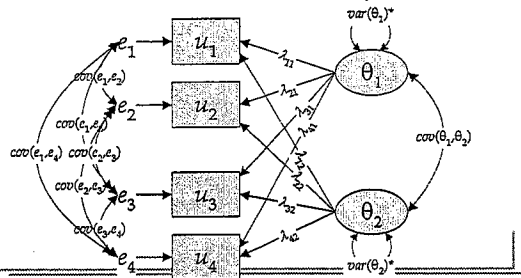
Barriers to Developing Useful Diagnostic Scores

- Unidimensionality of the total test scores as the primary measurement goal during test construction
- Content-driven perspectives about the nature of "dimensionality" (for diagnostic purposes)
- Exploratory factor analytic and related dimensionality detection procedures that "discover" multidimensional information in a particular data set
- Creative item writers

2007 GMAC® Conference on Computerized Adaptive Testing



The Nemesis of Factor Models: Rotational Indeterminacy



2007 GMAC® Conference on Computerized Adaptive Testing



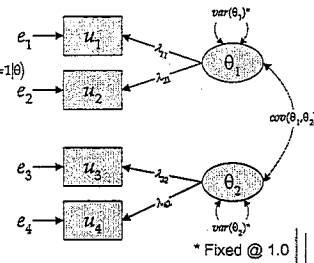
Local Independence (LI) to the Rescue (Under PMI)

Strict LI:

$$L(u_i|\theta) = P(u_i=1, u_j=1, \dots, u_n=1|\theta) \\ = P(u_i=1|\theta)P(u_j=1|\theta)\dots P(u_n=1|\theta)$$

Weak LI:

$$\text{cov}(u_i, u_j|\theta) = 0$$



2007 GMAC® Conference on Computerized Adaptive Testing



If we can generate item banks that exhibit PMI, MIRT CAT is entirely viable...

- Bayes scoring (e.g., Segall, 1996, 2000)
- Item selection maximizes the posterior information matrix or a linear composite of the underlying traits
 - A composite objective function can be easily derived using an appropriate multidimensional item response model, e.g., $P(U|\theta_1, \theta_2, \dots, \theta_m)$
 - $\text{Cov}(\theta_1, \theta_2, \dots, \theta_m)$ can be exploited, even under simple structure

2007 GMAC® Conference on Computerized Adaptive Testing



A Linear CAT Approximation to Maximizing PMI

- van der Linden (2005, Chapters 8 & 9; also see Segall, 1996, 2000; Luecht, 1996; Veldkamp & van der Linden, 2002)
- For $g=1, \dots, n$, maximize

$$\sum_{i=1}^I w^T \text{diag} [I_i(\hat{\theta}^{(g-1)})] x_i$$
- subject to <constraints>
- where w is a weight function that can be based on statistical or logical criteria

2007 GMAC® Conference on Computerized Adaptive Testing



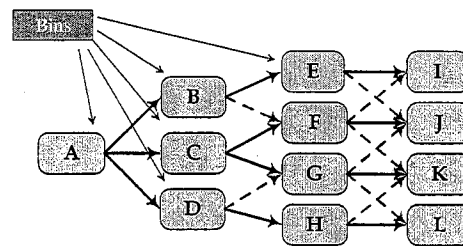
Bundled Multidimensional Computer-Adaptive Multistage Tests

- Luecht's (1997; 2003) derivative of CAST (Luecht & Nungester, JEM, 1998)
- Basic design unit = a bundle (of testlets)
- One design template for multiple bundles
 - Design template shows the position of bins in the bundle
 - Bins have a target TIF and an ATA specification set
- ATA used to create replications of the testlets for every bin
- Uses smaller testlets to facilitate replication and allow for greater adaptation
- PMI built into "diagnostic testlets"

2007 GMAC® Conference on Computerized Adaptive Testing



1-3-4-4 BMAT Design Template



2007 GMAC® Conference on Computerized Adaptive Testing



A 1-3-4-4 Bundle

Score Routing Table

$(\theta_1, \theta_2, \dots, \theta_m)^{\max}$

A₃
C₁
G₃
K₄

2007 GMAC® Conference on Computerized Adaptive Testing

Assessment Engineering and Construct-Based Design

- Constructs should be articulated in terms of hierarchical levels of procedural knowledge and skills, or in terms of levels of cognition applied to well-defined content strands, similar to PLDs
- All salient construct dimensions should be specified, along with the expected relationships
- Ultimately...focus on a specific number of useful, interpretable score scales

2007 GMAC® Conference on Computerized Adaptive Testing

Constructs and Targeted Measurement Information

- Measurement information is largely a function of two statistical characteristics of assessment tasks
 - The difficulty of item (i.e., its "location" with respect to some score scale)
 - The sensitivity of the item to the underlying construct being measured (i.e., discriminating power of the item)
- We can **TARGET** measurement information where it is needed most by controlling the difficulty of the assessment tasks
- Under AE, we must jointly control sensitivity to the construct of interest and "nuisance" dimensionality

2007 GMAC® Conference on Computerized Adaptive Testing

Target Information Functions Tied to Decisions and Score Interpretations

Test Information, $I(\theta)$

θ

2007 GMAC® Conference on Computerized Adaptive Testing

How AE Works with Target TIFs

- Measurement precision is targeted to specific regions of the construct map
- *Task models* are stacked in the greatest numbers where to approximate the density of measurement precision or test information needed
- Once templates and items are constructed and validated, automated test assembly is used to select tasks (items) to maximize measurement precision as needed

2007 GMAC® Conference on Computerized Adaptive Testing

Density of Task Models Proportion to Measurement Precision Needs

Performance

Integrates and interprets discourse-level text

Interprets sentential level text

Encodes, recognizes, and interprets salient lexical patterns

Encodes and defines whole words

Spelling and letter/symbol identification

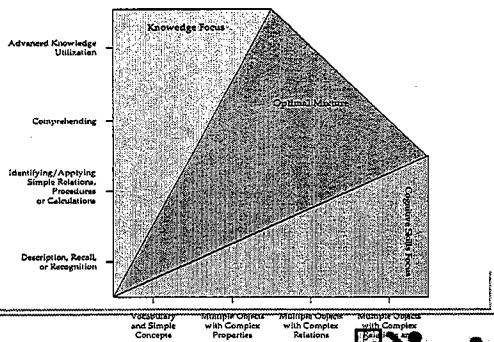
Task Model

Increasing Proficiency

Decreasing Proficiency

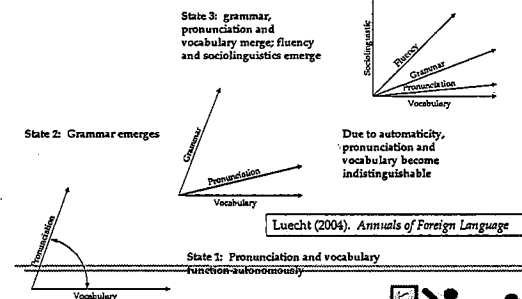
2007 GMAC® Conference on Computerized Adaptive Testing

Targeted Task Designs



2007 GMAC® Conference on Computerized Adaptive Testing

Morphing through Dimensionally More Complex States



2007 GMAC® Conference on Computerized Adaptive Testing

Building Task Models that Control Difficulty and Dimensionality

- A task-model grammar (TMG) captures the salient task challenge, context, and resources of a particular task
- Three components of a TMG
 - Functional clause: what the examinee does (may be composed of multiple nested functional clauses for higher-order cognitive skills)
 - Context: knowledge objects to be actively manipulated, their properties, and relations among the knowledge objects
 - Resources/conditions: auxiliary aids, tools, and facilitators or penalizing conditions that reduce reliance on prior knowledge

2007 GMAC® Conference on Computerized Adaptive Testing

Language-Based Task Design Drivers to Consider Under TMG

Knowledge

- Unique vocabulary/TTR
- Discipline-specific vocabulary
- Grammatical structures
- Semantic relations
- Number of "idea units"
- Key properties of objects
- Nature of relations
- Graphic complexity
- Contextual constraints/setting details
- Formula familiarity
- Auxiliary language

Cognitive Skills

- Auxiliary aids
- Training/direction
- Calculation complexity
- Persistence of relations
- Mental manipulations of images and visual objects
- Derivation or manipulation of formulas
- Functional constraints on applications (e.g., open-ended functionality vs. tight scripting)

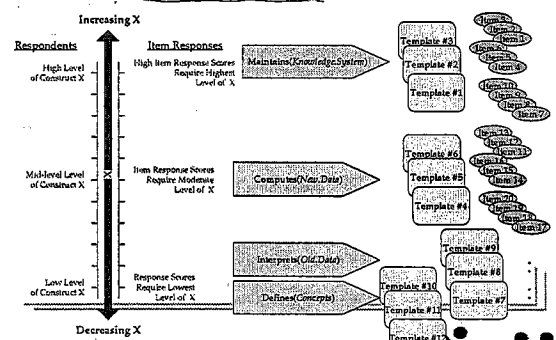
2007 GMAC® Conference on Computerized Adaptive Testing

Templates

- Each task model can be represented by multiple, exchangeable templates
- Templates have three components
 - Rendering model: detailed presentation format data and constrained interactive components for each task (e.g., LaDuca, 1994; Case & Swanson, 1998; Luecht, 2001, 2006)
 - Scoring evaluator: produces item- or measurement-opportunity-level scores from a performance (Luecht, 2001, 2005, 2006)
 - Data model: represents the rendering model, scoring evaluator, associated difficulty drivers (radicals), and incidental surface-level manipulables in database structures that can be used/activated by item writers to generate two or more items

2007 GMAC® Conference on Computerized Adaptive Testing

From Construct Map to Items



2007 GMAC® Conference on Computerized Adaptive Testing

Templates and Item Writing

- All item writing is funneled through one or more templates (i.e., item writers do NOT create their own templates)
- Component palettes can be restricted for each template
- Subtle variations in *templates, component palettes, and content/context* → lots of possible templates, and by extension, even more items, all with *similar* psychometric characteristics

2007 GMAC® Conference on Computerized Adaptive Testing



Other Engineering Steps

- Create pricing sheets to evaluate costs of new templates and component palettes
- Use cost-benefit analysis to evaluate
 - The information-per-unit-of-time for costly components
 - Real costs (\$\$\$\$) per unit of information
- Maximize the number of measurement opportunities and minimize the costs

2007 GMAC® Conference on Computerized Adaptive Testing



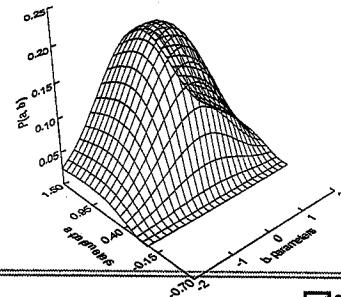
Supporting Psychometrics

- Task models and/or templates can be calibrated instead of individual items, using a hierarchical Bayes framework (Glas & van der Linden, *APM*, 2003)
- For multidimensional data...
 - Multiple, simultaneous unidimensional IRT models for each construct if simple structure is maintained through AE to achieve principled multidimensional information (PMI; Luecht, Gierl, Tan, & Huff, 2006)
 - MIRT models (e.g., Segall, 1996, 2000)
 - Augmented scoring (Wainer, et al., 2001)
 - Constrained latent class models

2007 GMAC® Conference on Computerized Adaptive Testing



Posterior Distribution for a Task Model = $P(a, b / U)$



2007 GMAC® Conference on Computerized Adaptive Testing



Psychometrics (*cont.*)

- Treat the hyperparameters as “super parameters” for the *task model*
- Estimate one set of common means and variance-covariances for the entire family
 - Less pretesting needed, once templates are verified
 - Fewer parameters leads to robust estimation
 - Misfit can be minimized if families are “well formed”
 - Hierarchical framework is extensible as a QC mechanism
 - Minimize posterior variance associated with individual items within templates
 - Minimize posterior variance associated with templates with task models

2007 GMAC® Conference on Computerized Adaptive Testing



Simulations and Computerized Performance Exercises (CPEs)

- High-fidelity simulations of one or more real-life performance environments
- CPEs immerse the examinee in an interactive set of tasks and/or challenges
- Wide range of *measurement opportunities (MO)*
 - Response latencies
 - Strategic problem-solving and decision-making
 - Free-form responses ranging from text to sound
 - Products from applications software
- Measurement opportunities → scores → information functions → adaptive test construction

2007 GMAC® Conference on Computerized Adaptive Testing



Schema for Financial Accounting

- Candidate instructions and expectations, including pacing suggestions and “deliverables”
- Situation (context)
- Develop needed information
 - Regulations
 - Generating evidence and documentation
- Make accounting choice and justify it
- Prepare elements of financial statements

2007 GMAC® Conference on Computerized Adaptive Testing



Instructions to the Examinee

This case involves lease capitalization. You will need to read the background information, located on the Content tab, and perform three tasks:

1. Determine and justify the type of lease (Capitalization)
2. Calculate the Present Value (PV) of the lease
3. Prepare a financial statement (see Statement)

This exercise should take approximately 10 minutes to complete

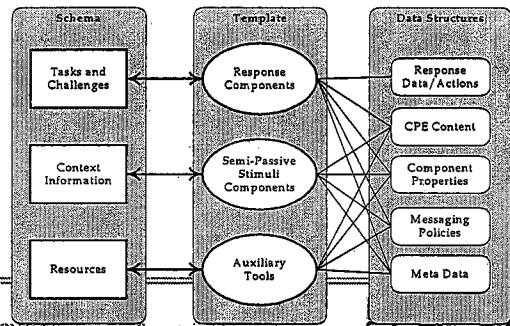
2007 GMAC® Conference on Computerized Adaptive Testing

Financial Worksheet Tab

Account Name	Value
Assets	10,000
Liabilities	5,000
Equity	5,000
Income	1,000
Expenses	1,000
Net Income	0

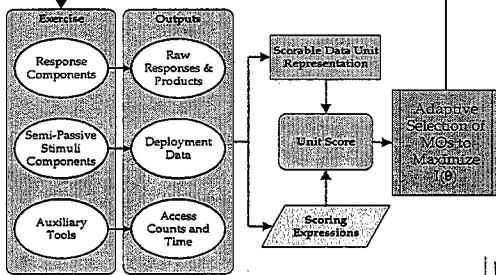
2007 GMAC® Conference on Computerized Adaptive Testing

Data Models for CPEs



2007 GMAC® Conference on Computerized Adaptive Testing

Adaptive Test Assembly and Scoring



2007 GMAC® Conference on Computerized Adaptive Testing



Thank you!

rmluecht@uncg.edu

2007 GMAC® Conference on Computerized Adaptive Testing



2007 GMAC Conference on Computerized Adaptive Testing
Minneapolis, MN, June 7-8, 2007.

*The Shadow-Test Approach:
A Universal Framework for Implementing Adaptive Testing*

*Wim J. van der Linden
University of Twente*

Abstract

In hindsight, the early research on adaptive testing was too narrowly focused on the statistical features of adaptive testing. It mainly addressed such topics as the efficiency of adaptive testing relative to traditional paper-and-pencil testing, approximate ability estimation methods for use in real-time adaptive testing, and the performances of different item-selection criteria. However, the first implementations of large-scale adaptive testing programs revealed a multitude of more practical problems related to, for example, the realization of large sets of content specifications, prevention of item compromise, item-pool design, differential speededness, and the equating of scores on adaptive tests to a released linear form for score reporting. We will present the shadow-test approach as a universal framework for solving such problems and illustrate the approach with several empirical examples.

The Shadow Test Approach: A Universal Framework for Implementing Adaptive Testing

Wim J. van der Linden
University of Twente

Financial support for several studies used in this presentation by the Law School Admission Council (LSAC) is gratefully acknowledged

Outline

- Test specifications and constraints
- Shadow-test approach (STA)
- Applications
- Levels of adaptation

Test Specifications and Constraints

- To realize test specifications, we have to impose *constraints* on the selection of the items in the test
 - Content distributions
 - Quantitative parameters
 - Word counts
 - Time required to complete the test
 - Logical relations between items (“enemies”, item sets, etc.)

Test Specifications and Constraints

- The best way to impose constraints on item selection is using *0-1 integer programming*
 - 0-1 variables to model constraints
 - Powerful IP solvers available to calculate optimal solutions for large problems (e.g., thousands of items and hundreds of constraints)
- van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: Springer

Test Specifications and Constraints

- For adaptive testing, the imposition of set of tests specifications reveals a basic dilemma:
 - Adaptation requires *sequential* item selection
 - Constraint realization requires *simultaneous* selection of all items
- General solutions:
 - Backtracking (impossible in CAT)
 - Projecting future consequences of item selection (*shadow-test approach*)

Shadow-Test Approach

- Modeling adaptive test assembly
- Graphical illustration of STA
- Important principle:

Any type of constraint available to give a fixed test a certain feature can be inserted in the model for the shadow tests to give the adaptive test the same feature

Shadow-Test Approach *Cont'd*

- Shadow tests are assembled on line using an optimized IP solver
 - “Hot start” start of solver
 - Solutions for tests with hundreds of constraints are found in a split second.

Modeling Adaptive Test Assembly

- Items in pool: $i=1, \dots, I$
- Design variables:

$$x_i = \begin{cases} 1 & \text{item } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$
- Items in adaptive test: $g=1, \dots$
- Set of items available for selection of g th item in the test: R_g

0-1 Modeling of Adaptive Test Assembly *Cont'd*

- Basic model for selection of g th item in adaptive test with *free* test length:

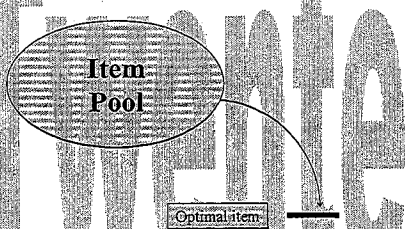
$$\begin{aligned} & \text{maximize } \sum_{i \in R_g} I_i(\hat{\theta}_{g-1}) x_i && \text{(maximum information)} \\ & \text{subject to} \\ & \sum_{i \in R_g} x_i = 1 && \text{(selection of one item)} \\ & x_i \in \{0, 1\}, i = 1, \dots, I && \text{(range of variables)} \end{aligned}$$

0-1 Modeling of Adaptive Test Assembly *Cont'd*

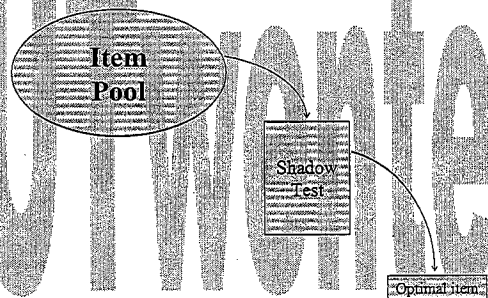
- Basic model for selection of g th item in adaptive test with *fixed* length of n items:

$$\begin{aligned} & \text{maximize } \sum_{i=1}^I I_i(\hat{\theta}_{g-1}) x_i && \text{(maximum information)} \\ & \text{subject to} \\ & \sum_{i=1}^I x_i = n && \text{(test length)} \\ & \sum_{i \in R_g} x_i = g-1 && \text{(previous items)} \\ & x_i \in \{0, 1\}, i = 1, \dots, I && \text{(range of variables)} \end{aligned}$$

Traditional Adaptive Testing



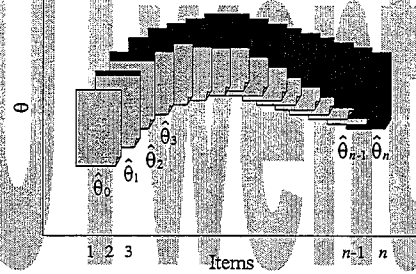
Shadow Test Approach



Shadow-Test Approach *Cont'd*

- Two-stage item selection:
 - Assembly of shadow test to:
 - contain all previous items
 - meet all constraints
 - have maximum information at current $\hat{\theta}$
 - Selection of free item from shadow test with maximum information at current $\hat{\theta}$

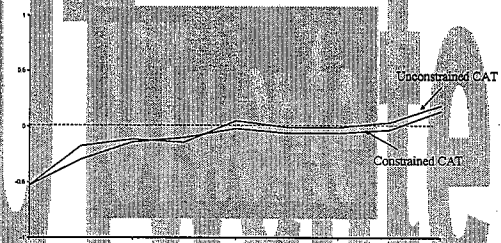
Shadow-Test Approach *Cont'd*



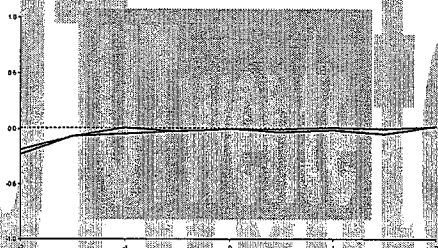
Application 1: Large Sets of Content Constraints

- Adaptive version of the *Law School Admission Test* (LSAT)
- Test length: 50 items
- Pool size: 753 items
- 427 constraints on test length, item content, item type, word counts, gender orientation, item-set structure, etc.

Bias (n=10)

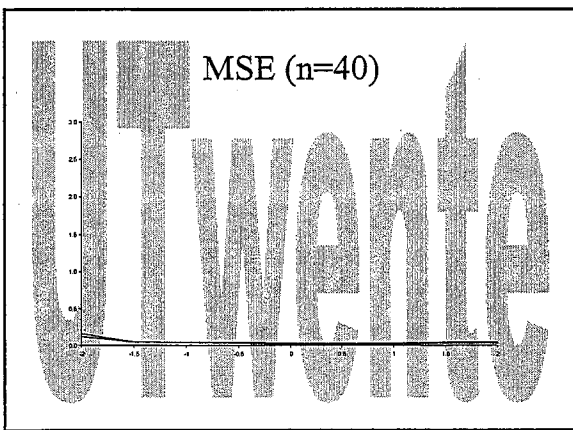
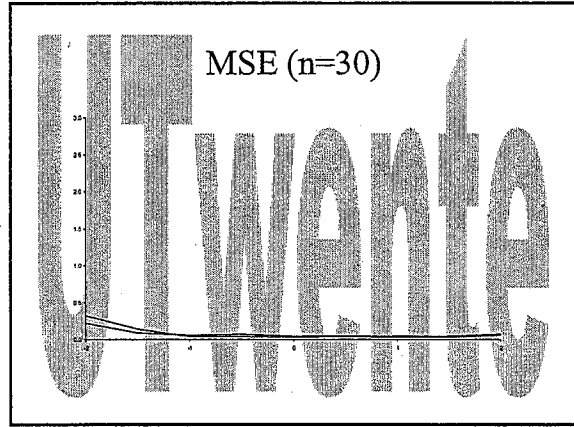
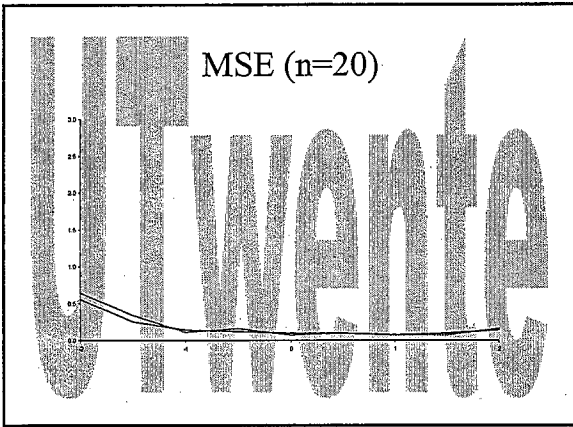
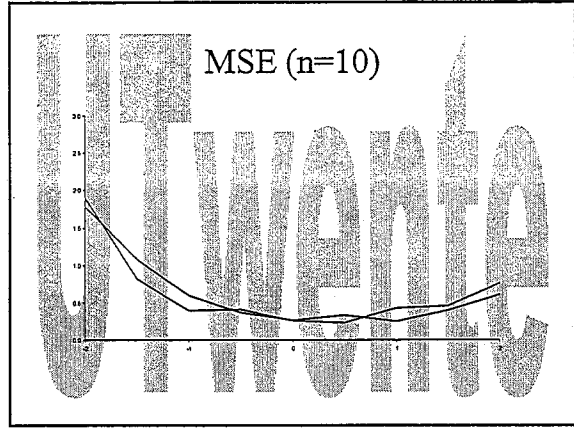
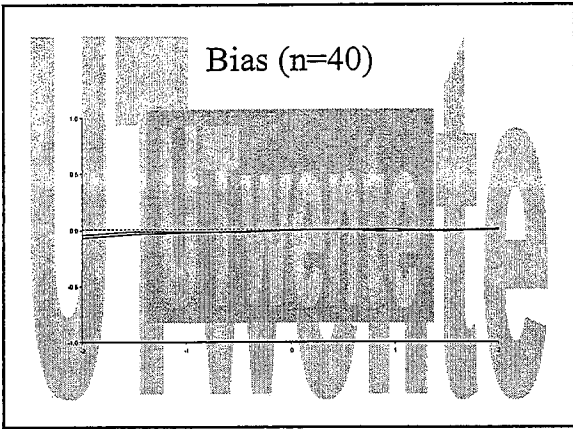


Bias (n=20)



Bias (n=30)



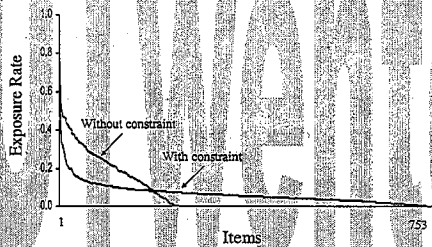


Application 2: Item Exposure Control

- *Alpha stratification*: stratify item pool on discrimination parameter a_i and select items from strata with increasing values for a_i .
- *STA*: add simple constraint to model for shadow test during stratum p :

$$\sum_{i \in \mathcal{Q}_p} x_i = n_p$$

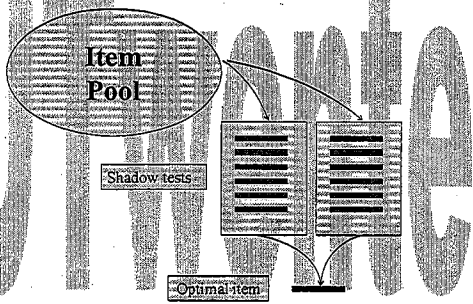
Alpha Stratification



Application 2: Item Exposure Control *Cont'd*

- *Simpson-Hetter method*: run probability experiment after item has been selected
- *STA*: run experiment over free items in shadow test

Multiple-Shadow-Test Approach



Application 2: Item Exposure Control *Cont'd*

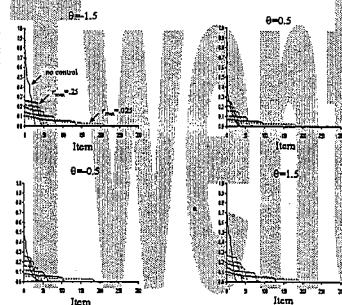
- *Item-eligibility method*: run one probability experiment before the test to determine eligibility of the items for the candidate
- Probabilities of eligibility are self adjusting
- *STA*: add simple ineligibility constraints to model for shadow test

$$x_i = 0$$

Application 2: Item Exposure Control *Cont'd*

- Empirical example for adaptive version of section from the LSAT
 - Conditional exposure control at $\theta = 2(-.5)^2$
 - Maximum exposure rates: .25, .20, .15, .10, .05, and .025

Exposure Rates



Application 3: Eliminating Differential Speededness

- Adaptive testing is liable to the problem of differential speededness
- Calibrate the items with respect to their time intensity in a response-time (RT) model
- Lognormal RT model
 - Speed parameter for the persons
 - Time intensity and discrimination parameters for the items

Application 3: Eliminating Differential Speededness *Cont'd*

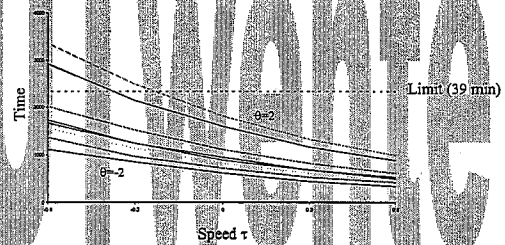
- Update estimate of candidate's speed parameter and use this to predict the RTs on all items in the pool
- *STA*: add simple time constraint to the model for shadow test

$$\sum_{i \in R_s} t_i + \sum_{i \in R_p} t_i^{predicted} \leq t_{tot}$$

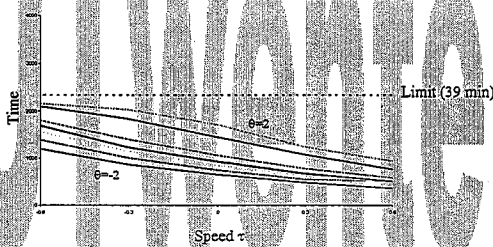
Application 3: Eliminating Differential Speededness *Cont'd*

- Example for CAT version of the ASVAB Arithmetic Reasoning Test
 - 15-item test from pool of 186 items
 - Time limit: 39 min (=2,340 sec.)
 - Simulation study with:
 - Ability: $\theta = -2.0(-5)2.0$
 - Speed: $\tau = -.6(.3).6$

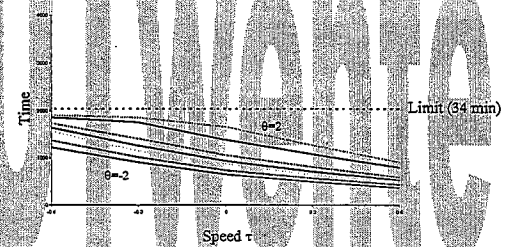
Time Used to Complete Test (Without Constraint)



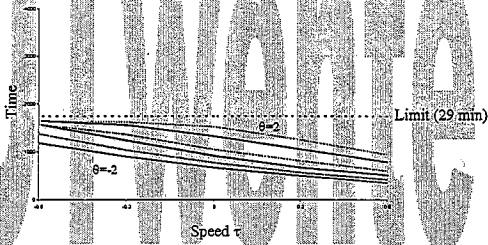
Time Used to Complete Test (With Constraint)



Time Used to Complete Test (With Constraint)



Time Used to Complete Test (With Constraint)



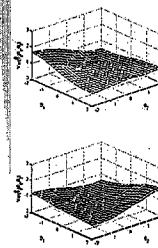
Other Applications

- Alternative item-selection criteria, e.g.,
 - Kullback-Leibler information
 - Bayesian criteria
- Automatic equating of CAT scores to number-correct scores on reference test
- Multidimensional adaptive testing with content constraints

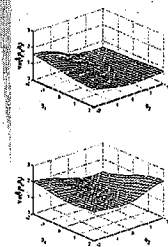
Other Applications *Cont'd*

- Adaptive testing from a pool with cloned items

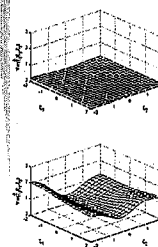
Two-dimensional Mathematics Test (θ_1 and θ_2 both intentional)



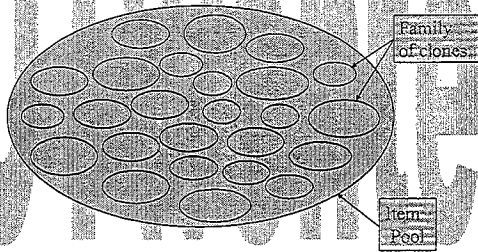
Two-dimensional Mathematics Test (θ_1 intentional)



Two-dimensional Mathematics Test ($\xi_1 = \theta_1 + \theta_2$ intentional)



Item Cloning in Adaptive Testing



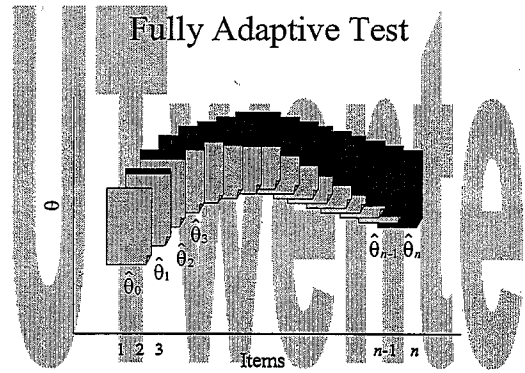
Other Applications *Cont'd*

- Adaptive testing from a pool with cloned items

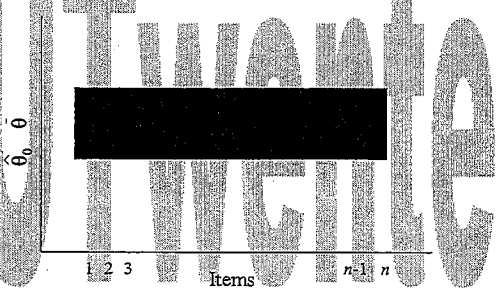
Level of Adaptation

- Shadow-test approach allows us to assemble and evaluate tests with different levels of adaptation:
 - Linear test
 - Adaptive linear-on-the-fly test
 - Multistage test
 - On-the-fly multistage test
 - Fully adaptive test

Fully Adaptive Test



Linear Test



Adaptive Linear-on-the-Fly Test

