Part I Basic Science Item Reengineering Pilot

Report of the August 2004 Administration*

National Board of Examiners in Optometry

This report is the culmination of a multiyear effort in elevating the clinical relevance or relatedness (currently referred to as *clinicality*) of the Part I Basic Science examination. Two task forces were formed to explore item clinicality in the context of a broader effort in restructuring the entire 3-Part examination sequence. The task forces, which submitted their respective reports to the Board in 1998 and 1999, had a broad perspective, as they were comprised of individuals from both within and outside the profession of optometry.

This effort in enhancing item clinicality should be viewed as part of an ongoing evolution to keep up with changes in clinical knowledge and practice, as well as changes in the profession of psychometrics. During the 25-year period in which the Board was located in metropolitan Washington DC, there have been three major restructurings of the examination program. The item reengineering project, in combination with the examination restructuring, the conditions study, and the conditions-based 3-Part Content Outline, will, when implemented, represent the next major program change.

Prior Major Examination Restructurings

In ***1981***, the examination sequence consisting of Parts I, 2A, and 2B shifted from being Section-based to being Part-based, for numerous psychometric reasons. Many psychometric procedures were introduced or changed to be consistent with this initiative, including the formation of examination development committees that had responsibility for test development and test scoring, in conjunction with staff. To promote item clinicality, each committee had to include at least one clinician and one recent graduate.

In ***1987***, the examination sequence was compressed and redefined as Part I Basic Science and Part II Clinical Science. A substantial amount of new content was added; specifically, Human Biology in Part I, and Systemic Conditions in Part II. The rationale for the new content was to reflect the expanded scope of optometric practice, which had begun to include use of drugs for therapeutic purposes. A stand-alone certification test in the Treatment and Management of Ocular Disease (TMOD) had been implemented in 1985, sponsored by the association of state optometry boards (ARBO, which was known as IAB at the time).

The new Part I and Part II content was to assess the knowledge underlying therapeutic drug use. In addition, with a goal of greater clinicality and integration, the examination development committees were structured as two subcommittees: one for a Part I section, and the other for the corresponding Part II section. For example, one committee was responsible for both the Human Biology Section in Part I and the Systemic Conditions Section in Part II. Another committee was responsible for the Optics Section in Part I and the Refractive … Conditions Section in Part II. The intent was for the Part II subcommittee to instill a clinical perspective on the Part I items.

To further the integration of content, two examination councils were formed: one for each examination Part. Each council was structured as a committee of examination subcommittee chairs, with the addition of a second examination subcommittee member for lengthy sections (e.g., Optics). The examination councils replaced the Board (i.e., by Board delegation) as the final post examination committee meeting integrator and arbiter of content. The councils also replaced the examination committees in conducting the usual telephone conference call reviews of the initial scoring data in determining what, if any, items were flawed, and how they should be handled in the final scoring iteration.

In ***1993***, the TMOD content was absorbed within Part II. To avoid a resultant skewing of content, Part II was lengthened by nearly 50% to retain the robust sampling of content from the other content areas within Part II. A comparable lengthening of Part I occurred also, similarly to avoid a skewing of content that would have resulted from the expanded content in Human Biology.

Also significant in 1993 was the launch of the new Part III Patient Care examination. The Clinical Skills examination (CSE), which had been administered as a special test since 1989, and the VRICS (Visual Recognition and Interpretation of Clinical Signs) examination, which had been administered as a special test since 1991, were complemented by the initial administration of Patient Management, which consisted of paper-and-pencil simulations of patient encounters, known as patient management problems (PMPs).

The combination of these three sections formed Part III. Although Part III has since evolved to include more TMOD content within CSE, and PAM (Patient Assessment and Management) has replaced the VRICS and PMP sections, these changes are small in impact in comparison with the impact of adding Part III to the National Board examination sequence in 1993. There was a sense of completeness that characterized the launch of Part III, as this examination included an assessment of all of the skills (including psychomotor, affective, and communication) and formats (including a practical or performance test, and visual) for the broadest feasible assessment of entry-level competence.

Item Reengineering

This history of continual enhancing and updating is an important context for the Part I item reengineering project. The underlying philosophy guiding this project was that the Basic Science examination could and should be enhanced with regard to item clinicality.

The project assumed the moniker of *reengineering* to convey a rethinking and redesign of how items could be written and/or presented for optimal assessment of the clinical essentials of underlying Basic Science knowledge.

An item reengineering task force (subsequently referred to as the *Task Force*) was formed in 2002 to explore alternative assessment concepts and formats. However, preceding formation of the Task Force, the Part I examination subcommittees were given various assignments during their annual test development meetings to better understand item clinicality, and to increment the level of clinicality on the examination. These activities included classifying each Basic Science item against the Clinical Science Content Outline, and identifying the items in each Basic Science subcontent area that were highest and lowest in clinicality, to identify their commonalities. Although these activities appeared to be successful in elevating the clinicality of Part I, there was a pervasive sense that item reengineering and/or examination restructure had greater potential to enhance the test.

The Task Force held a series of telephone conference calls during 2002, which preceded the annual test development committee meetings. These teleconferences were intended to explore the issues and alternatives prior to the Task Force live meeting in September 2002, during which a report to the Board was prepared containing 13 recommendations.

The Task Force recommended for pilot development and administration an item clustering strategy. The pilot proceeded following Board approval at its March 2003 meeting. The cluster format involved Basic Science themes in which items from diverse sections or subsections would be juxtaposed around a clinical context or introduction. This type of clustering differed from the type used in PAM, in that the Basic Science clusters contained few visuals, the items did not reference each other or require a particular sequence (e.g., as in diagnosis items preceding follow-up items in PAM), and the items were neither dependent on each other nor on the clinical introduction for correct answers to be selected. The latter trait can appear to be illogical and wasteful of candidate testing time; however, it allowed very useful performance comparisons to measure the effect of clustering on item difficulty. Details regarding these comparisons are provided later in this report.

Item clustering was recognized by the Task Force to be the most difficult to implement. The low yield (items written) and even lower “hit rate” (items considered acceptable for use within clusters) validated this Task Force perception. Nonetheless, the clustering approach was chosen because it offered the potential for maximal elevation of clinicality. In addition, the resultant content integration was seen as helping the Basic Science test sharpen its focus on entry-level knowledge, as items that reference multiple content areas are regarded by the Basic Science subcommittees as more likely to assess essential knowledge at an entry level of difficulty.

A cluster authoring meeting was held on June 29 - July 1, 2003 at the Board office in Bethesda, MD. In addition to the four original Task Force members, there were five other attendees, four of whom were former National Board examination committee members. The yield from this 3-day session was 27 clusters containing 141 items. The items were edited by staff to conform to National Board style and then sent to the nine attendees for further review and editing, and additional judgments regarding content, appropriateness, clinicality, and difficulty.

Of the 27 clusters written, six were selected by the Part I Examination Council for inclusion on the August 2004 Part I examination. A seventh cluster of comparable quality was selected to serve as a sample cluster on the National Board website, to accompany a narrative explaining the purpose and nature of the pilot to prospective examinees.

The August test was selected because it is the administration that is taken by the full student cohort. An experimental design was established to include each of these clusters in the last of the three administrative sessions. This would enable the first two sessions to establish a baseline performance measure for subsequent performance comparisons.

Analyses

In order to measure any effect of the clustering, all candidates were administered half of the pilot items within their cluster groups. The other half of the pilot items were administered as unclustered, stand-alone items, embedded within the test booklet. This experimental design was intended to measure how the clinical context and thematic focus of clustering impacts candidate performance, if at all.

The table on the next page displays how the item clusters were administered between candidate groups. Candidates are identified as orange or red, corresponding to the colors of the alternative answer sheets. As this table indicates, each candidate group served as the experimental control for the other.

Production of the two versions of the session three test booklet was flawless, and the candidates generally seemed to have an adequate familiarity with the cluster format. Candidate familiarity was anticipated by staff, as considerable effort had been expended to promote awareness of the clustering project. Nonetheless, the complex but successful production was an important accomplishment, as it created the empirical conditions for adequate statistical controls.

Another concern regarding the pilot was that candidates might not have sufficient time for completion of session three, as the introduction to each of the three item clusters required additional text for candidates to process. However, offsetting the additional text was the content integration of the cluster items that resulted from their focus on a common theme. Several candidates commented on this efficiency, noting that the pilot items' common focus facilitated processing the item content. The standard test time analysis data for the examination, which compares candidate test time utilization among the three sessions, did not indicate that additional time was utilized by candidates for session three in comparison with the time utilized for sessions one and two. The comparative time utilization data are displayed in Table 2 and were presented to the Board previously at its November 2004 meeting.

Table 1: Experimental Design

		BASIC SCIENCE EXAMINATION SECTION		
		Human Biology	***Ocular / Visual Biology***	***Optics***
CANDIDATE TYPE	***Orange***	C-1 grouped	C-4 grouped	C-5 grouped
		C-2 embedded	C-3 embedded	C-6 embedded
	Red	C-1 embedded	C-4 embedded	C-5 embedded
		C-2 grouped	C-3 grouped	C-6 grouped

Candidates were encouraged to use the standard test critique form to indicate their reactions to the pilot. A transcribed copy of the candidate critiques that pertained to the reengineering component comprises Appendix A of this report. The critiques were written during session three, and are sequenced in test center order. The comments are numbered "*200,*" as this is the number assigned to candidate critiques of the overall test.

As the comments indicate, candidates were split in their perception of the clusters. Some candidates stated that the introductions were essentially "window dressing" and a waste of time. Other candidates stated that the introductions provided a useful clinical context that was appreciated.

Of particular significance were the differing reactions of candidates to the same item based on whether it was in a cluster, or isolated. The items that were perceived to have high clinicality when administered within a cluster were not seen as having high clinicality when administered in the alternative booklet as stand-alone items, despite being nearly or completely verbatim reproductions of each other. This raises the issue of how well candidates perceive clinicality, regardless of the importance and clinical generalizability of a Basic Science item, if a patient is not provided or referred to for context.

The statistical analyses are based on the 136 items in session three that were scored, and all 1609 candidates who sat for the test. This population was the largest number of candidates to sit for any National Board examination. Of the 1609 candidates, 808 used booklet A, while 801 used booklet B. The two session-three test booklets were structured to be as similar as possible, with the exception of the cluster items. Of the 145 total items comprising session three, 127 were identical (i.e., not reengineered). Of these 127 items, 106 had the same sequence number in both test booklets.

Nine of the 145 items administered in session three were deleted from the final scoring during the Part I Examination Council telephone conference call review, following standard operating procedures. In addition, nine items were deleted in session one and three items were deleted in session two. In total, 21 items were deleted from the final scoring. Of the nine items deleted in session three, two had been reengineered. The final performance comparisons for session three were based on 16 reengineered items and 120 "conventional" items.

Table 2: Test Retention by Time Criteria by Test Center

		Session 1				Session 2				Session 3			
		# Retained		% Retained		# Retained		% Retained		# Retained		% Retained	
Test Center*	Number of Candidates	5 Min	15 Min	5 Min	15 Min	5 Min	15 Min	5 Min	15 Min	5 Min	15 Min	5 Min	15 Min
Berkeley	73	15	28	21%	38%	11	22	15%	30%	14	20	19%	27%
Big Rapids	56	1	2	2%	4%	1	2	2%	4%	0	0	0%	0%
Birmingham	56	3	8	5%	14%	1	3	2%	5%	2	4	4%	7%
Bloomington	85	4	11	5%	13%	2	3	2%	4%	5	11	6%	13%
Boston	110	12	27	11%	25%	11	15	10%	14%	10	16	9%	15%
Chicago	152	18	24	12%	16%	9	22	6%	14%	16	21	11%	14%
Columbus	69	3	9	4%	13%	0	1	0%	1%	2	4	3%	6%
Ft. Lauderdale / Miami	145	20	32	14%	22%	18	27	12%	19%	16	23	11%	16%
Houston	114	16	19	14%	17%	9	15	8%	13%	8	11	7%	10%
Memphis	130	24	24	18%	18%	12	12	9%	9%	15	15	12%	12%
New York	88	19	30	22%	34%	12	22	14%	25%	18	21	20%	24%
Philadelphia	153	17	27	11%	18%	7	14	5%	9%	9	20	6%	13%
Portland / Forest Grove	96	17	28	18%	29%	3	11	3%	11%	7	12	7%	13%
San Juan	60	11	21	18%	35%	13	13	22%	22%	14	17	23%	28%
St. Louis	56	5	7	9%	13%	5	7	9%	13%	7	8	13%	14%
Tahlequah	34	4	7	12%	21%	5	7	15%	21%	2	4	6%	12%

** Fullerton was not analyzed because a prior day change in test center location that required shuttle service may have affected time usage patterns. Waterloo was not analyzed because of a small N.*

Appendix B is a compilation of the comparative performance data for each of the items that comprised session three. These data include the item format (i.e., MCQ – multiple choice question – or cluster item), separated by a slash (/). For reengineered items (i.e., items with high clinicality), the format to the left of the slash refers to test booklet A, while the format to the right of the slash refers to test booklet B.

The usual difficulty and discrimination indices (p- and r-value, respectively) are provided for each item. For reengineered items, the p-value difference is provided. A positive value indicates that the item received a higher percentage of correct responses (i.e., was somewhat easier, empirically) in its booklet A format than in its booklet B format. The absence of p- and r-values indicates that the item was deleted from scoring.

The two columns of p-value disparities for reengineered items indicates whether the performance disparity (i.e., A-B) is a comparison of MCQ in booklet A vs. cluster item in booklet B, or a comparison of cluster item in booklet A vs. MCQ in booklet B. The "*same*" column provides the performance comparison for items that were in the same format for both booklets. The "same" column is applicable to items 1-46, 109-145, and 44 other items positioned elsewhere in the booklets.

The performance comparisons are summarized in Table 3 in decimal form for alignment purposes. In the discussion that follows, the difficulty indices are referenced as percentages.

The session mean scores for test booklets A and B were 63.3% and 64.2%, respectively. This could suggest that booklet B was slightly easier than booklet A, possibly due to the selection of items presented within clusters. However, this difference between booklet A and booklet B candidates was also present for sessions one and two, indicating that the session three disparity resulted from differences in candidate ability rather than differences in item difficulty.

Table 3: Performance Comparisons for Reengineered Items in Clustered and Unclustered Format

	Number	*Mean Item Performance*								
	of Items	*Booklet A*		*Booklet B*		*Mean Item P Difference (A-B)*				
Test Session	*Scored*	*P*	*R*	*P*	*R*	*Same*	*MCQ/Clus*	*Clus/MCQ*	*Mean MPI*	*Mean P - Mean MPI*
Session 3										
Total	136	0.633	0.233	0.642	0.231	-0.008	-0.004	-0.016	0.611	0.026
High clinicality	16	0.701	0.232	0.711	0.218				0.628	0.078
Other	120	0.624	0.233	0.632	0.233				0.609	0.019
Session 1	136	0.670	0.224	0.677	0.223				0.617	0.057
Session 2	142	0.659	0.218	0.667	0.216				0.603	0.060

The session three means for both test booklets were lower than the means for session one (67.0% and 67.7%, respectively) and session two (65.9% and 66.7%, respectively). The greater difficulty of session three is not attributable to the reengineered items, which exhibited relatively high means for the high clinicality items (70.1% and 71.1%, respectively). Rather, the non-reengineered (i.e., "*other*") items were responsible for the poorer session performance, with means of 62.4% and 63.2%, respectively.

The means of the item performance differences (A-B) are sufficiently low to indicate that any effect of item presentation format on performance was negligible. The differences that are observed appear to be more reflective of candidate ability disparities than item characteristic disparities.

The last two columns in Table 3 refer to the pass-fail standards. The National Board determines the pass-fail cutoff score through an item-by-item assessment of difficulty, using a version of the Nedelsky procedure. A detailed explanation of this procedure with an example is posted on the National Board website.

Using this procedure, a minimum performance index (MPI) or standard is established for each item. The pass-fail cutoff score for the overall test, on a raw score basis, is equal to the sum of the MPIs. The percentage equivalent of the cutoff score is equal to the average MPI.

As each item has a performance standard, it is possible to calculate a cutoff score for any group of items, such as a content section or subsection, a timed session, or items that have a special or distinct trait, such as being reengineered. Calculating a cutoff score in this manner is intended for rendering individual pass-fail decisions but rather, to have a benchmark for evaluating the aggregate performance levels.

The next-to-last column in Table 3 displays the cutoff score for each session, as well as for the reengineered and "other" items in session three. These comparisons indicate that the highest cutoff score was for the reengineered items in session three. The cutoff score for these items was 62.8%. This relatively high performance standard (i.e., relative to the other items in session three, and to the items comprising sessions 1-2) indicates that the examination committee members regarded the reengineered items to be slightly easier than the other items, and therefore expected a higher level of performance.

The expectation of better performance was exceeded. The last column on the right summarizes the relationship between the mean and pass-fail cutoff score for each of the item groupings. This column calculates the disparity between the mean item performance level (i.e., mean MPI). The greater the disparity above zero, the higher the level of performance is relative to the corresponding cutoff score. The 0.078 difference for the reengineered items exceeded the disparity for all of the other item groups. This greater difference indicates that candidates not only performed best on the reengineered items, but more importantly, the candidates performed better on the reengineered items *relative* to the corresponding cutoff score. The implication of this relationship is that a test comprised exclusively of items with comparable content and difficulty characteristics would have a higher pass rate. This probable outcome is discussed in greater detail later in this report.

Discussion

One of the most important findings in this study is that the presence of a clinical context and thematic integration exerts little if any effect on candidate performance. To the extent that there is an effect of clinicality and integration on elevating candidate performance, the effect is inherent in the principle tested by the item, rather than contextual enhancement. The significance of this effect for future item reengineering and examination restructure is that it appears unnecessary to struggle with the complexity and low yield of item cluster development to increase item clinicality. Rather, the goal of higher clinicality may be attained more simply by better item focus, content, and selection. Integration may be achieved easier by grouping tiems based on conditions rather than background contexts.

Discussions of examination restructure based on a conditions-oriented Content Outline have considered redesigning Part I to blend current Basic Science and Clinical Science content. If the current 435-item length of Part I were retained, the inclusion of some Clinical Science content would broaden the Part I Content Outline and therefore, the content that could be included in the test, resulting in greater content selectivity. The content that would be shifted from the current Clinical Science examination would be items classified in skills 1-2. These skills cover epidemiology / history / symptoms (skill 1), and clinical signs / techniques (skill 2).

This content is below the level of diagnosis and treatment, and is the material that has typically been most difficult to distinguish between being Basic Science or Clinical Science content. Placing this content in a restructured Part I examination would eliminate the blurred distinction while simultaneously elevating the clinicality of Part I, without the need for item clusters.

A conditions-based Content Outline for this restructured test would foster content integration and clinicality by sequencing items around conditions. However, the clinical context or introduction that characterized the reengineered item clusters would be unnecessary, as the conditions themselves would be the focal point of item groupings. The clinicality of this approach would be quite evident. In comparison with the introductions in the cluster items, conditions would represent an abridged context for item clinicality much like taxonomy and cognitive skill levels represent an abridged proxy for instructional or performance objectives.

A second important finding is the mean score exceeding 70% for the reengineered items. Although the sample size of 16 items is small, the finding is nonetheless significant, as the mean score for the Basic Science examination has never reached 70%. In fact, since its inception in 1987, the highest Basic Science mean score is 67%. In contrast, the mean score for every targeted administration of Clinical Science has exceeded 70%.

A related finding is that candidates performed better on the reengineered items than on each of the other item groupings. In addition, as the mean-minus-cutoff statistic indicated, when projected over the length of a complete test, the pass rate on the reengineered items would be elevated. Converting this statistic to a z-score indicates that the probable pass rate for such a test would be 78%, considerably higher than the actual pass rate on the overall test of 67%.

As a note of caution, in addition to the small sample size cited earlier, the content sampling of the reengineered items was narrow. In fact, the reengineered items did not include any content from the Psychology section. Items from this section typically exhibit the highest section mean score; however, they also exhibit the highest mean MPI (i.e., cutoff score). The combined impact of these item characteristics might reduce the projected pass rate, although the projected pass rate would still exceed the actual pass rate.

Summary and Conclusions

The Task Force that guided the item reengineering project invested substantial effort to resolve the lingering issue of how to raise the clinicality of the Basic Science examination. Several approaches were identified. The Task Force selected an item clustering approach – the format that was most complex in nature and difficult to develop, but which offered the greatest potential for achieving the goal of a more clinical, integrated Part I examination. The reengineering item clusters that were developed were administered on the August 2004 examination in both cluster and stand-alone mode. The effect of mode of presentation on candidate performance was negligible.

This finding can appear to be discouraging. However, it conveys that clinicality in a Basic Science examination, at least with regard to candidate perception, is not necessarily the result of a clinical context. Rather, it results from the importance and centrality (i.e., core knowledge) of the fact or principle to be recalled or applied. While the inclusion of clinical contexts can elevate the character of the test, the contexts are not needed, and are perceived as undesirable by some candidates, if used to explain or justify the clinicality of items. Recognition of this lack of association is significant, as simpler approaches to raising clinicality can and should now be pursued enthusiastically, rather than tentatively. Shifting skills 1-2 from Clinical Science to Basic Science, which will allow items to be grouped around clinical conditions, as well as better item selectivity, should prove to be an effective solution.

Restructuring Part I in this manner has significant implications for restructuring Part II and Part III, such as for the optimal timing of the tests with regard to the academic curriculum and graduation. These implications are discussed in considerable detail in other reports that have been presented to the Board. However, unlike the major restructurings in 1987 and 1993, no new content would be introduced, and the number of examination Parts would neither be reduced nor expanded. While examination restructurings are major undertakings that can be somewhat confusing and disruptive, the rationale, implications, and mechanics of this next restructure should be more readily understood.

Upon implementation, the entire 3-Part examination sequence will be more clinical and more integrated. The three examination Parts will also be more distinguishable from each other in both content and format. With this greater clarity, clinicality, and integration, this new examination structure should be well-received by the profession's constituencies.

* For further discussion of this study, contact Leon Gross, Ph.D., Associate Executive Director, and Director of Psychometrics & Research.

APPENDIX A: Candidate Critiques of the Item Reengineering Project

ITEM # 200 Comment # 1

The grouped question format does nothing to help answer questions. It only gives us more to read.

ITEM # 200 Comment # 2

The theme-clustered questions were not as helpful as I expected them to be. The picture was helpful, but the other information presented seemed as if you were telling an irrelevant story before you asked the questions. I just didn't understand the relevance of the presentation. This type of information (basic sciences) would probably be best suited for just questioning and answering.

ITEM # 200 Comment # 3

The patient scenarios given during the examination were not useful in answering the questions. I feel that they were not necessary to have been included in the exam.

ITEM # 200 Comment # 4

Just wanted to say that I really liked the clinic-related questions and they are actually relevant!

ITEM # 200 Comment # 5

[This critique refers to Version B of Session 3.] The grouped questions were good except for 62-64. If you want these questions to be more clinical, they should relate to the eye and optometry. Questions 62-64 is much more medical based - systemic. I think that this test should be more clinically related because most of this info is not used in general practice.

ITEM # 200 Comment # 6

The clustering questions were good because the info was given in clinical context and helped you to stay in the same mind-set for a couple of questions. This helped to reduce the amount of time spent taking the test.

ITEM # 200 Comment # 7

I liked the cluster questions. It was a little easier to think about them with an actual (pretend) patient to picture.

ITEM # 200 Comment # 8

I like the "clustering"; I believe the test should be made as clinically relevant as possible, and this is a good start. The picture seemed a little out of focus, but I was still able to discern gram (-) diplococci.

ITEM # 200 Comment # 9

Cluster questions were OK, but seemed like they could have been asked anywhere instead of a cluster.

ITEM # 200 Comment # 10

[Candidate took Version B of Session 3.] Introduction for items 62-64, 65-67 is pretty much useless.

ITEM # 200 Comment # 11

I would have preferred not to have the "Introductory Information" questions. I thought it was unfair to those students who had these types of questions, since 1/2 of the students taking the test did not have this format. These questions are more clinical and should not be on the first part of the boards.

ITEM # 200 Comment # 12

More case-related questions should be asked.

ITEM # 200 Comment # 13

The group questions have little to do with the original case present.

ITEM # 200 Comment # 14

I really like the new format of the "sample/trial" questions. They are much more clinically applicable, as opposed to a lot of the irrelevant, minute detailed-oriented questions found in Part I. You can understand an entire system or mechanism and not know 1 minute detail about it and miss that (or a # of) question. It seems somewhat unfair. Whereas the new format tests applicable knowledge optometrists are expected, or should, know.

ITEM # 200 Comment # 15

I liked the clustered items; it helped to orient my thinking!

APPENDIX B: Performance Comparisons for Reengineered Items in Clustered and Unclustered Format

		Item	*Item Performance Data*							
Booklet A	*Booklet B*	*Format*	*Booklet A*		*Booklet B*		*Item Performance Differences (A-B)*			
Item #	*Item #*	*(A/B)*	*P*	*R*	*P*	*R*	*Same*	*MCQ/Clus*	*Clus/MCQ*	*MPI*
1	1	Same	0.55	0.31	0.60	0.34	-0.05			45
2	2	Same	0.64	0.22	0.65	0.28	-0.01			60
3	3	Same	0.90	0.11	0.90	0.12	0.00			60
4	4	Same	0.46	0.13	0.40	0.20	0.06			60
5	5	Same	0.73	0.25	0.77	0.22	-0.04			60
6	6	Same	0.63	0.23	0.65	0.17	-0.02			60
7	7	Same	0.65	0.17	0.67	0.13	-0.02			45
8	8	Same	0.61	0.16	0.58	0.11	0.03			60
9	9	Same	0.33	0.23	0.36	0.19	-0.03			60
10	10	Same	0.74	0.16	0.77	0.21	-0.03			60
11	11	Same	0.55	0.27	0.53	0.26	0.02			60
12	12	Same	0.58	0.24	0.61	0.18	-0.03			45
13	13	Same	0.43	0.18	0.40	0.20	0.03			60
14	14	Same	0.61	0.44	0.62	0.41	-0.01			60
15	15	Same	0.58	0.13	0.59	0.23	-0.01			60
16	16	Same	0.61	0.18	0.65	0.16	-0.04			60
17	17	Same	0.78	0.26	0.79	0.24	-0.01			60
18	18	Same	0.75	0.35	0.74	0.32	0.01			60
19	19	Same	0.69	0.26	0.68	0.25	0.01			60
20	20	Same	0.46	0.29	0.47	0.26	-0.01			60
21	21	Same	0.68	0.11	0.71	0.09	-0.03			60
22	22	Same	0.50	0.11	0.52	0.11	-0.02			60
23	23	Same	0.53	0.18	0.56	0.11	-0.03			45
24	24	Same	0.61	0.30	0.63	0.29	-0.02			60
25	25	Same	0.58	0.32	0.59	0.25	-0.01			60
26	26	Same								
27	27	Same	0.74	0.33	0.75	0.34	-0.01			60
28	28	Same	0.65	0.17	0.65	0.19	0.00			90
29	29	Same	0.61	0.35	0.61	0.41	0.00			90
30	30	Same	0.68	0.24	0.73	0.20	-0.05			60
31	31	Same	0.30	0.27	0.30	0.27	0.00			60
32	32	Same	0.72	0.27	0.72	0.29	0.00			60
33	33	Same	0.52	0.24	0.52	0.27	0.00			60
34	34	Same	0.61	0.31	0.63	0.27	-0.02			45
35	35	Same	0.47	0.15	0.50	0.16	-0.03			60
36	36	Same	0.76	0.08	0.76	0.12	0.00			90
37	37	Same								
38	38	Same	0.65	0.24	0.65	0.23	0.00			60
39	39	Same	0.47	0.29	0.45	0.30	0.02			60
40	40	Same	0.82	0.22	0.82	0.26	0.00			60
41	41	Same	0.52	0.23	0.59	0.24	-0.07			60
42	42	Same	0.71	0.30	0.74	0.30	-0.03			60
43	43	Same	0.57	0.31	0.57	0.31	0.00			60
44	44	Same	0.88	0.27	0.90	0.18	-0.02			90
45	45	Same	0.51	0.34	0.54	0.38	-0.03			60
46	46	Same	0.53	0.12	0.51	0.11	0.02			60
47	64	MCQ/Clus	0.43	0.24	0.47	0.19		-0.04		45
48	48	Same	0.90	0.30	0.92	0.22	-0.02			90
49	49	Same	0.37	0.36	0.39	0.37	-0.02			90
50	50	Same	0.50	0.22	0.52	0.18	-0.02			60

51	51	Same	0.74	0.32	0.71	0.27	0.03			60
52	52	Same	0.46	0.40	0.48	0.38	-0.02			60
53	53	Same	0.79	0.27	0.80	0.29	-0.01			45
54	54	Same	0.34	0.30	0.40	0.30	-0.06			45
55	63	MCQ/Clus								
56	56	Same	0.73	0.09	0.73	0.18	0.00			60
57	57	Same	0.84	0.33	0.86	0.34	-0.02			60
58	58	Same	0.70	0.22	0.72	0.16	-0.02			90
59	59	Same	0.61	0.41	0.65	0.36	-0.04			60
60	67	MCQ/Clus	0.60	0.29	0.62	0.27		-0.02		90
61	61	Same	0.42	0.11	0.38	0.07	0.04			60
62	62	MCQ/Clus	0.87	0.38	0.88	0.38		-0.01		60
63	55	Clus/MCQ	0.54	0.29	0.57	0.25			-0.03	45
64	47	Clus/MCQ	0.55	0.16	0.59	0.07			-0.04	60
65	60	Clus/MCQ	0.31	0.03	0.39	0.13			-0.08	60
66	84	Clus/MCQ	0.92	0.30	0.94	0.24			-0.02	90
67	95	Clus/MCQ	0.85	0.20	0.82	0.23			0.03	90
68	86	Clus/MCQ	0.63	0.33	0.63	0.28			0.00	60
69	102	Clus/MCQ								
70	97	Clus/MCQ	0.92	0.27	0.92	0.17			0.00	60
71	108	Clus/MCQ	0.54	0.19	0.53	0.26			0.01	60
72	71	Same	0.42	0.37	0.47	0.25	-0.05			60
73	72	Same	0.55	0.23	0.56	0.18	-0.01			60
74	73	Same	0.80	0.29	0.83	0.27	-0.03			60
75	70	MCQ/Clus	0.90	0.22	0.88	0.27		0.02		60
76	74	Same	0.56	0.32	0.57	0.25	-0.01			60
77	75	Same	0.66	0.31	0.66	0.36	0.00			60
78	76	Same	0.76	0.32	0.78	0.26	-0.02			60
79	77	Same	0.72	0.27	0.74	0.26	-0.02			60
80	78	Same	0.72	0.33	0.70	0.46	0.02			60
81	79	Same	0.77	0.25	0.76	0.28	0.01			45
82	80	Same	0.58	0.17	0.58	0.17	0.00			60
83	81	Same	0.51	0.25	0.53	0.30	-0.02			45
84	82	Same	0.91	0.25	0.92	0.20	-0.01			90
85	83	Same								
86	85	Same	0.81	0.31	0.82	0.24	-0.01			90
87	66	MCQ/Clus	0.80	0.11	0.76	0.10		0.04		45
88	87	Same	0.78	0.31	0.77	0.32	0.01			60
89	89	Same	0.65	0.12	0.63	0.19	0.02			45
90	88	Same	0.63	0.23	0.65	0.16	-0.02			60
91	65	MCQ/Clus	0.67	0.29	0.66	0.28		0.01		60
92	90	Same	0.44	0.27	0.41	0.26	0.03			60
93	91	Same	0.47	0.21	0.48	0.16	-0.01			60
94	92	Same	0.61	0.12	0.60	0.13	0.01			60
95	93	Same	0.59	0.27	0.59	0.24	0.00			60
96	94	Same	0.87	0.23	0.87	0.27	0.00			60
97	96	Same	0.24	-0.05	0.24	-0.03	0.00			60
98	98	Same	0.34	0.17	0.34	0.20	0.00			60
99	99	Same	0.35	0.17	0.33	0.20	0.02			36
100	100	Same								

101	101	Same	0.38	0.02	0.41	0.09	-0.03		60
102	69	MCQ/Clus	0.87	0.26	0.87	0.30		0.00	90
103	103	Same	0.54	0.03	0.55	0.11	-0.01		60
104	104	Same	0.32	0.21	0.33	0.18	-0.01		60
105	105	Same	0.74	0.34	0.75	0.28	-0.01		60
106	106	Same	0.52	0.30	0.52	0.29	0.00		60
107	107	Same	0.91	0.19	0.93	0.24	-0.02		60
108	68	MCQ/Clus	0.82	0.15	0.85	0.07		-0.03	30
109	109	Same	0.78	0.28	0.79	0.30	-0.01		60
110	110	Same	0.87	0.18	0.91	0.16	-0.04		45
111	111	Same	0.78	0.25	0.78	0.26	0.00		60
112	112	Same	0.53	0.21	0.51	0.23	0.02		45
113	113	Same	0.74	0.22	0.73	0.18	0.01		60
114	114	Same	0.68	0.11	0.71	0.26	-0.03		45
115	115	Same							
116	116	Same	0.77	0.22	0.77	0.27	0.00		60
117	117	Same	0.87	0.30	0.87	0.25	0.00		60
118	118	Same	0.40	0.19	0.43	0.26	-0.03		90
119	119	Same	0.77	0.23	0.76	0.16	0.01		60
120	120	Same	0.43	0.14	0.46	0.17	-0.03		60
121	121	Same	0.69	0.39	0.69	0.39	0.00		45
122	122	Same	0.89	0.13	0.90	0.15	-0.01		60
123	123	Same							
124	124	Same	0.70	0.26	0.67	0.27	0.03		60
125	125	Same	0.72	0.28	0.72	0.26	0.00		60
126	126	Same	0.86	0.26	0.85	0.26	0.01		60
127	127	Same	0.90	0.17	0.86	0.20	0.04		60
128	128	Same	0.87	0.19	0.84	0.19	0.03		60
129	129	Same	0.37	0.09	0.40	0.11	-0.03		45
130	130	Same	0.59	0.25	0.61	0.28	-0.02		60
131	131	Same	0.72	0.13	0.73	0.14	-0.01		90
132	132	Same	0.32	0.38	0.33	0.39	-0.01		60
133	133	Same	0.39	0.34	0.37	0.33	0.02		60
134	134	Same	0.60	0.29	0.63	0.32	-0.03		60
135	135	Same	0.68	0.21	0.67	0.20	0.01		60
136	136	Same	0.56	0.22	0.57	0.26	-0.01		60
137	137	Same	0.54	0.03	0.58	0.18	-0.04		60
138	138	Same	0.56	0.08	0.51	0.07	0.05		60
139	139	Same	0.32	0.36	0.31	0.37	0.01		60
140	140	Same	0.60	0.20	0.65	0.29	-0.05		45
141	141	Same	0.71	0.09	0.71	0.04	0.00		60
142	142	Same	0.71	0.30	0.71	0.32	0.00		45
143	143	Same	0.74	0.22	0.74	0.14	0.00		90
144	144	Same	0.81	0.26	0.85	0.26	-0.04		90
145	145	Same							