AB-1 - Paper

Using Questionnaires to Gain Insight into Retest Effects

John Thain
Defense Language Institute Foreign Language Center (DLIFLC)

The Defense Language Institute Foreign Language Center (DLIFLC), located on the Presidio of Monterey, California, is the proponent for the Defense Language Aptitude Battery (DLAB); it is also the agency responsible for basic language training within the Department of Defense (DoD). DLAB is used to screen candidates for language training who have previously met Armed Services Vocational Aptitude Battery (ASVAB) requirements for language-related occupational specialties. Certain sections of DLAB are presented on audiotape, while others are presented in the test booklet.

Recent studies have suggested that some parts of the DLAB could be reduced in scope to save administration time, while other parts of the DLAB could benefit from a corresponding lengthening of administration time. These studies have led DLIFLC to consider several options comprising relatively small-scale changes in the DLAB; these changes would retain most of the current test content, reduce the number of total test items, and retain current methods of test scoring and score interpretation. Part of data collection in support of this goal involved administration of a modified and reduced form of DLAB (DLAB 1.5) to a population that had previously taken the operational DLAB. DLAB 1.5 included all core items that would probably be included in a future version of DLAB. In order to interpret DLAB 1.5 test data, questionnaires given to DLAB 1.5 examinees were extensively used to gain insight into retest effects. The objectives for DLAB 1.5 administration in order of importance were: (1) to determine if changing the time allotted (but not test content) for completing the printed part of the test would affect test statistics and predictive validity, (2) to determine if differences in audio quality in test administrations at MEPS would have any effect on retest scores for retained audio portions of the original test (in which no test content was changed), and (3) to field test revisions of four test items with faulty statistics in the printed portion of the test.

This paper deals only with a comparison of a part of the original DLAB item response record with the item response record for DLAB 1.5, a reduced and modified version of the original test. Test scores in the data in this paper represent only the number correct in subpart totals and have no relation to the familiar DLAB standard scores used operationally.

DLAB 1.5, the reduced and modified test, retained the following parts from the original DLAB:

Audiotape paced parts. Part II (P2) involves identification of auditorially presented syllable stress patterns. It consists of 18 items requiring seven and one half minutes administration time; In P2, test administration is paced by audiotape. Part III (P3) involves deductive language learning of an artificial language; stimulus material is presented on audiotape. P3 consists of 44 items arranged into three subtests (P3S1, P3S2, and P3S3) and requires 28 minutes administration time; like P2, it is also paced by audiotape. These two parts are identical to and require the same administration time as a corresponding block of 62 items in the original DLAB.

Written part with time limit. Part IV (P4) involves inductive language learning based on pictures accompanied by printed artificial language work samples. It consists of 26 items carried over from the original DLAB plus the four replacement items referred to above. The examinee has 35 minutes to complete P4 in DLAB 1.5, rather than the 25 minutes allotted in the original DLAB for P4.

In total, DLAB 1.5 consisted of 92 items and requires just over 70 minutes administration time, consisting of 88 original items and 4 replacement items. The same number of items in original DLAB require slightly only 60 minutes, because the P4 administration time in DLAB is 10 minutes less the P4 administration time in DLAB 1.5. While increased administration time for P4 is one option for change in future DLAB revision, such an increase is not to be introduced operationally without reducing the administration time of the remaining other portions of the DLAB by a roughly equal or even greater amount.

The idea for trying out an extended administration time for P4 was influenced by a DLI-funded psychometric study of DLAB conducted by O'Mara (1994). O'Mara drew a large sample from the "unrestricted" population (N=32,866) to analyze P4 examinee data collected during 1987-1988. The unrestricted population includes examinees with scores so low that they were not subsequently eligible for any language as well as examinees scoring high enough to be eligible for language training. Thus, in addition to "selected" examinees with complete

language training criterion data, this sample included examinees who had taken the DLAB, but for whom there was no criterion training data because they had not subsequently been assigned to language training. The data from this sample indicated that the series of items toward the end of P4 had progressively higher nonresponse rates as the end of P4 approached. The mean DLAB total test scores of the examinees not responding to each item were also increasing toward the end of P4.

It was decided to administer DLAB 1.5 to a substantial sample of those students arriving at DLIFLC to attend language classes. These students had been selected to attend DLIFLC based on their initial DLAB scores at Military Enlistment Processing Stations (MEPS).

DLIFLC was strategically placed to carry out such a study with a minimum of resources, Regulatory guidance insures that DLIFLC receives completed answer sheets from DLAB administrations at MEPS. DLIFLC was able to easily arrange DLAB 1.5 testing for incoming students upon arrival in Monterey. When all the original DLAB answer sheets from MEPS also arrived, it was easy to scan both sets of answer sheets and create a reduced item response record of original DLAB item response data record corresponding to just those 92 items retained or replaced one-for-one in the DLAB 1.5 test administration data record. Any number of "what if" comparisons could then be facilitated by cutting and pasting together smaller and more manageable databases composed of "virtual item response records" drawn from parts of the two sets of item response records. Subsequent criterion data and self-reported information from the examinees about both DLAB and DLAB 1.5 administrations could easily be added to such databases of interest.

Part of the plan was to solicit the observations of the examinees taking the modified DLAB about the differences they perceived between the original DLAB administration at MEPS and the administration of DLAB 1.5 which they had just undergone. After the administration of DLAB 1.5 in Monterey, they responded to questionnaire items by marking multiple-choice alternatives on their test answer sheets. The numbering of the questionnaire items followed the last item number on the modified DLAB 1.5. Questionnaire response data were scanned along with the DLAB 1.5 test data. This made it possible to correlate the item responses on the DLAB and DLAB 1.5 test administrations with the opinions of the examinees about the two test administrations.

Another part of the original plan, as yet not carried out, was to subsequently gather the language training criterion data for these examinees who had been double-tested with both DLAB and DLAB 1.5. A variety of statistical analyses were to be carried out to suggest which version of DLAB maximized prediction benefits at which costs of test administration time tradeoffs. We hoped that criterion data would enable us to select an option with appropriate cutoff scores varying across foreign languages of varying difficulty. However, the most important criterion data were available only after the completion of the DLI basic course training, which lasted from 25 to 63 weeks dependent on the language. During the interim period, the DLI Research Division conducted an exploratory analysis of the questionnaire data and the data from the two test administrations. Partly due to changes in local research priorities, the criterion data, though now available in DLI school databases, has yet to be analyzed. In the absence of criterion data, the exploratory nature of the questionnaire analysis of retest data is even more pronounced and conclusions at this point are even more tentative than originally intended.

During 1994, the DLAB 1.5 and the questionnaires were administered to 1058 students who had previously taken DLAB. DLAB 1.5 was administered in DLIFLC language laboratories through a central audio console to cohorts of thirty to sixty examinees wearing headphones. The original DLAB administration at MEPS took place under conditions varying by locality.

Table I summarizes information about the first and second administrations. In columns under the label "first test administration," the raw mean score and maximum possible score are provided for the correspondingly reduced portion of the original DLAB item response set (labeled DLAB-R)--in other words, the original DLAB item response set reduced to just those items for which there is a counterpart in DLAB 1.5. Presented under the columns labeled "gain scores" are the gain scores of DLAB 1.5 over DLAB-R and the average p-value gain per item in each item set.

TABLE I

Raw score test gains in DLAB 1.5 Administration over Official DLAB administrations

Mean Raw Score with Maximum Possible Score in Parentheses (N=1058)

	From original administration	From second administration	Gain Scores	
Item set	DLAB-R	DLAB 1.5	Gain	Gain per item
62 items from P2 and P3	44.39(62)	47.90(62)	3.51(62)	.06(62)
Faulty original Part IV items Replacement Part IV Items	2.20 (4) ++	** 2.58 (4)	.38(4)	.09(4)
Part IV items not replaced	16.34(26)	18.42(26)	2.08(26)	.08(26)
TOTAL SCORE	62.92(92)	68.90 (92)	5.98(92)	.07(92)

++ Not present in original DLAB ** Replaced by replacement items in DLAB 1.5

Table I indicates that the per item increase in mean raw score in the second administration for the initial set of 62 items is much less than the increase per item in the common 26 items retained in P4. Above and beyond the bare data in Table I, further triangulation using questionnaire data and other information presented later in this paper seems to indicate that the gain scores in the first 62 items are largely due to conventional practice effects; however, differences in administration conditions between MEPS and DLIFLC related to the quality of audio subtests may also play a somewhat lesser role in these test gains. Questionnaire data presented later in this paper also suggest that a conventional practice effect seems to play as important a role as the increase in administration time in the gain for the last 26 common items; indeed there is surprisingly little strong evidence to suggest that the increase in administration time is the major factor responsible for the gain scores found in this last group of 26 items.

Had our ultimate goals not included introducing an operational change in administration time in P4, the task of merely determining the effect of the replacement of four items from the original test on the norming and score interpretation of a future operationally revised DLAB would have been much simpler. The means and standard deviations of the original DLAB subtests in the sample of 1058 students selected to enter DLIFLC closely tracks the typical means and standard deviations of other even larger samples drawn from the same population. The change resulting from introducing four replacement items into operational would have minimal effect, if any, on the standard score reporting system.

If a gain score for the 26 common items from P4 could be somehow shown to result not from the change in administration time, but from conventional practice effect, such a gain score should not be reflected in new norms or standard scores for a revised DLAB with extended administration time for P4. One source of evidence as to the cause of P4 gains was to compare these gains to gains in the audio portion of the tests in which there was no extension of administration time; another way is was to consult examinee input about the relative significance of practice effects and extension of administration time.

The audio portion of the DLAB 1.5 consisted of four groups of items P2, P3S1, P3S2, and P3S3. Table II indicates the relationship between examinees' self-reported perceptions about practice effect and their actual retest gains on these four audio subtests and the P4 common items. 1022 of the 1058 examinees responded to the question "Do you think having previously taken DLAB helped you taking this modified version?"

TABLE II

MEAN RAW SCORE GAINS FOR AUDIO TESTS AND P4 COMMON ITEMS

Examinees categorized in terms of own beliefs about the efficacy of practice effects

(Mean increase in item p-values for each subtest in parentheses)

How much help?	N	Pct	P2	P3S1	P3S2	P3S3	P4 Common Items	TOTAL
Great deal	209	23.2%	.83(.05)	.59(.05)	.45(.03)	1.79(.11)	2.40(.09)	6.06(.07)
Somewhat	556	52.6%	.60(.03)	.45(.03)	.21(.02)	1.73(.10)	2.14(.08)	5.13(.06)
Maybe a little	198	18.7%	.37(.02)	.41(.03)	-.26(-.02)	1.84(.11)	2.13(.08)	4.49(.05)
Not at all	59	5.6%	-.15(-.01)	.10(.01)	.08(.01)	.54(.03)	1.14(.04)	1.71(.02)
TOTAL	1022	100%	.56(.03)	.45(.03)	.16(.01)	1.69(.10)	2.13(.08)	4.99(.05)

Because of the peculiarities of this data set, if N=1000, an average gain in p-values of .01 across 15 items (or a mean increase of .15 for a 15 item subtest) would probably be significant beyond the .05 level on a two-tailed paired samples T-test, whereas an average p-value gain of .03 per item might be required to be significant at the same level if the sample size were only 100.

In the whole sample (N=1022), the 45 items in P2, P3S1, and P3S2 taken together display moderate (but highly statistically significant) average p-value gains of .025, while the 17 items in P3S3 displayed much greater average p-value gains of .10 per item (t=16.21). Table II indicates that 94.5% of the examinees acknowledged some degree of practice effect and that the relatively few examinees (5.6%) not acknowledging a practice effect had much smaller test gains than the other examinees. It is not surprising from a psycholinguistic point of view that P3S3 and P4 would show larger gain scores in terms of conventional practice effects than P2, P3S1, and P3S2. The retest situation for these former subtests would be more likely to stimulate the examinee to reactivate cognitive and procedural "mental hooks than the latter subtests. While all subtests in DLAB are artificial in the sense that they are based on an artificial language, P3S3 and P4 tend to approach sentence level syntax and semantics rather than consisting of simple morphological or nonsense word content.

TABLE III

Where examinees thought previous experience helped

and where experience actually did help the most

(Mean increase in item p-values in each subtest in parentheses)

	N	P2	P3	P4	TOTAL
P2	143	.49(.03)	1.92(.04)	1.40(.05)	3.82(.04)
P3	260	.79(.04)	2.83(.06)	1.69(.06)	5.31(.05)
P4	211	.73(.04)	2.19(.05)	2.95(.11)	5.87(.06)
None	283	.41(.02)	2.03(.05)	2.20(.08)	4.64(.05)
TOTAL	897	.61(.03)	2.28(.05)	2.10(.08)	4.99(.05)

Examinees who thought that previous experience with the DLAB helped a great deal or helped somewhat were asked a further question: did their prior experience with DLAB help more with some parts of DLAB 1.5 than others? Table III shows that examinees who thought previous experience had helped them more on P3 or P4 did

in fact have much higher retest gains on corresponding test than persons who answered otherwise. However, examinees who said their previous experience had helped them most on P2 had lower retest gains across the board.

Examinees were also queried concerning the administration of the audio portion of the DLAB at MEPS. While headphones were used in all the DLI DLAB 1.5 administration, only 14% of the examinees said that headphones were available for use in the official MEPS administration. In addition, while DLAB 1.5 was administered in cohorts of 30 or 60 examinees, 75% of the examinees reported taking MEPS DLAB alone or with one or two examinees, and only 6% reported MEPS administrations involving 10 or more examinees.

Examinees were asked whether the audio portion was clearer during the official MEPS administration or during the DLAB 1.5 administration at DLIFLC . Table IV below indicates that all three groups responding to this question had overall test gains; those who believed DLI audio was better than MEPS audio had larger gains on audio tests than those who believed that there was no difference in audio between DLI and MEPS. Relatively few examinees thought that MEPS audio had been better than DLI audio; however this small subgroup had a consistent pattern of lower retest gains on audio tests.

Although other questions were asked comparing test administration conditions at the MEPS and DLI, there was little evidence that any factors other than conventional practice effect and overall better audio at DLIFLC had any effect on audio retest scores. The data in Tables II, III, and IV taken together suggest that conventional practice effects were a more important factor in retest gains on audio tests than the improvement of audio quality at DLI over MEPS audio quality.

TABLE IV

How examinees compared MEPS audio and DLIFLC audio and how much examinees gained upon retest (Mean increase in item p-values in each subtest in parantheses)

	N	Pct	P2	P31	P32	P33	All audio tests	P4 common items
DLI audio better	467	44.1%	.66(.04)	.55(.04)	.37 (.03)	1.82(.11)	3.40(.05)	1.82(.11)
MEPS audio better	45	4.3%	.29(.02)	.09(.01)	-.20(-.01)	1.69(.10)	1.87(.03)	1.69(.10)
No difference	535	51.6%	.47(.03)	.40(.03)	.00(.00)	1.55(.09)	2.42(.04)	1.55(.09)
Total	1047	51.6%	.55(.03)	.45(.03)	.16(.01)	1.68(.10)	2.83(.05)	1.68(.10)

The impetus to try out an increased administration time for P4 came from O'Mara's study of the DLAB item response data from the unrestricted (unselected) population. What was true of the unrestricted population did not hold up for the DLAB-R item and DLAB 1.5 response sets, which were derived exclusively from the tests of "selected" examinees with passing cutoff scores. In neither of these samples from the selected population did the mean total test scores of nonresponding examinees increase toward the end of P4. The number of omits was also minimal in both cases.

Questionnaire data did not initially yield a clear picture of any group of examinees that benefited substantially enough from the expanded administration time in P4 to plausibly account for the total retest gains in P4. Examinees were asked to answer the following question with one of five options: Did you finish the last portion of the test before time was called? Table V indicates that all response groups answering this question showed retest gains on the first 10, and second 10, and final 6 common items in P4. Despite the fact that the number of omits was negligible (in both the official DLAB administration and the DLAB 1.5 administration), 16.5% of the DLAB 1.5 examinees responded that they "failed to finish" P4 on DLAB1.5. Furthermore, examinees choosing "failed to finish" as a response did not score lower on the last six items in P4 in DLAB 1.5 than on earlier P4 items. Nor was their mean score lower on the last six items than colleagues who responded differently on this questionnaire item. It is possible some respondents broadly interpreted "failure to finish" to mean "failure to be able to give fullest attention to all of the items regardless of test item sequence on the test under the conditions of

expanded administration time." In support of this interpretation, one notes that the first three response groups in Table IV (arguably the groups working till the last minute on DLAB 1.5) showed relatively higher retest gains on the last six items than the other two response groups. Nevertheless, overall comparison of Table IV data with other tables offers only weak support for assigning retest gains to the increased administration time rather a generalized conventional practice effect.

TABLE V

Self-report of adequacy of administration time for P4 common items in DLAB 1.5 and how much examinees gained upon retest
(Mean increase in item p-values for each subtest in parantheses)

	N	Pct	P4- (1rst 10)	P4 (2nd 10)	P4 (Last 6)	TOTAL P4 Common Items
Didn't finish	175	16.7	.67(.07)	.74(.07)	.51(.08)	1.92(.07)
Finished, little or no time to check work	91	8.7	.79(.08)	.87(.08)	.45(.08)	2.11(.08)
Finished early, just able to check work	174	16.6	1.12(.11)	.71(.07)	.56(.09)	2.39(.09)
Finished early, checked work, still had time left	294	28.0	.93(.09)	77(.08)	.20(.03)	1.90(.07)
Finished early, could have checked, but didn't	314	30.0	1.22(.12).	69(.07)	.30(.05)	2.21(.09)
TOTAL	1048	100	.99(.10)	.74(.07)	.36(.06)	2.09(.08)

Official DLAB test scores at MEPS become part of the examinee's records. DLAB 1.5 administration data did not enter into examinees' official records. With this in mind, the questionnaire asked examinees whether they were motivated to do their best on DLAB 1.5. 75% responded that they were highly motivated and 24% responded unsteady motivation. Only 1% admitted not trying very hard.

So far this paper has examined retest gains after categorizing examinees in terms of response to a single questionnaire item at a time. Our project also experimented with combining several questionnaire responses in order to create new coding variables. The scope of this paper allows us to do little more than briefly touch on such efforts. We attempted to code certain examinees' scores as more likely to reflect retest gains due to administration time and other examinees as reflecting a gain due to conventional practice effects. For example, examinees who did *not* finish the first DLAB administration, but who did finish the DLAB 1.5 administration with time to *recheck* responses, and also claimed that previous experience with DLAB didn't help them on the test would be high on a scale constructed to reflect "gain due to change in P4 administration time rather than conventional practice effect ." Conversely examinees who *finished* the first DLAB administration before time was called, *didn't* check answers even though time was available on a subsequent DLAB 1.5 administration, confessed mediocre motivation to take DLAB 1.5, and claimed that previous DLAB experience helped them a great deal during P4 would be coded as low on the same scale (meaning that this kind of examinee's score could be assigned more to practice effects rather than administration time). Taken individually, all of these experimental recoded scales may seem to have a certain arbitrary ad hoc quality; a skeptic might even question whether test gain from changes in administration time could ever be cleanly separated from conventional practice effects in these circumstances. However, results from various scales seemed to converge on a conclusion. Conventional practice effects seemed to have more to do with bringing about P4 retest gains than did the change in administration time.

References

O'Mara, F., (1994), Improving the measurement of language learning aptitude: a psychometric analysis of the Defense Language Aptitude Battery. PRC Inc. Work performed under contract for Defense Language Institute Foreign Language Center.

INTERNET DOCUMENT INFORMATION FORM

A . Report Title: Using Questionnaires to Gain Insight Into Retest Effects

B. DATE Report Downloaded From the Internet 4/21/99

C. Report's Point of Contact: (Name, Organization, Address, Office Symbol, & Ph #): **Navy Advancement Center**
Attn: Dr. Grover Diel (850) 452-1815
Pensacola, FL

D. Currently Applicable Classification Level: Unclassified

E. Distribution Statement A: Approved for Public Release

F. The foregoing information was compiled and provided by:
DTIC-OCA, Initials: VM_ **Preparation Date**: 4/21/99__

The foregoing information should exactly correspond to the Title, Report Number, and the Date on the accompanying report document. If there are mismatches, or other questions, contact the above OCA Representative for resolution.