UNCLASSIFIED

AD NUMBER

ADB951151

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

**Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; JUN 1952. Other requests shall be referred to Adjutant General's Office (Army), Washington, DC 20310.**

AUTHORITY

AGO ltr 29 Apr 1980

THIS PAGE IS UNCLASSIFIED

THIS REPORT HAS BEEN DELIMITED
AND CLEARED FOR PUBLIC RELEASE
UNDER DOD DIRECTIVE 5200.20 AND
NO RESTRICTIONS ARE IMPOSED UPON
ITS USE AND DISCLOSURE.

DISTRIBUTION STATEMENT A

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED.

Best
Available
Copy

AD B951151

RESEARCH NOTES

Research Note 52-38 ✓ UNANNOUNCED (11) June 1952

(6) PROBLEMS IN THE STANDARDIZATION AND USE OF PSYCHOLOGICAL (screening) TESTS FOR THE ARMED FORCES,

(12) 11 (10) J. E. Uhlaner

Presented to the Panel on Personnel, Committee on Human Resources, Research and Development Board, at its seventeenth meeting, 6-7 March 1952.

(14) AGO-PRS-RESEARCH NOTE-52-38

DDC
RECEIVED
NOV 9 1979
E

DDC FILE COPY

PERSONNEL RESEARCH SECTION
PR AND P BR, AGO 003 650 slt

79 11 07 139

PREFACE

Endorsement for the attached paper, Problems in the Standardization and Use of Psychological Screening Tests for the Armed Forces, was given by the Panel on Personnel Committee on Human Resources, Research and Development Board at its seventeenth meeting, held 6-7 March 1952 in the Pentagon. At that meeting, the Army representative, E. A. Rundquist, explained that the paper had been prepared "because of the present confusion and need for clarification concerning the Congressional intent and the actual operation and results of the psychological screening tests..." After review and further discussion of the document, which J. E. Uhlaner had presented earlier at the Symposium, the Air Force Member recommended that it be endorsed by the Panel as an official, factual statement of the problems concerning standardization and use of these tests. The Panel approved the document as recommended.

Accession For
NTIS GRA&I
DDC TAB
Unannounced
Justification
By
Distribution/
Availability Codes
Dist | Avail and/or special
2

PROBLEMS IN THE STANDARDIZATION AND USE OF PSYCHOLOGICAL SCREENING TESTS FOR THE ARMED FORCES

J. E. Uhlaner
Personnel Research Section
The Adjutant General's Office
Department of the Army

The paper "Specifications of Psychological Screening Tests for the Armed Forces" has discussed the major human variables which can be profitably assessed by mental screening tests. Further, it has been pointed out that these variables are not only determined by research results but that prevailing operational requirements and conditions must also be considered in order to develop optimally useful personnel measurement tools. Assuming, then, that we have general agreement on the variables which are to be assessed, we must now direct our attention to the standardization of the instruments which are to provide the scores and measures for various purposes that are anticipated by the personnel policy makers, the personnel administrators, as well as the research people.

Basically, the question of standardization of any of these psychological instruments involves the establishment of norms for some purpose. (See Chart 1) It is my belief that one of the primary questions to be settled before we determine the kind of standardization we desire, methodologically speaking, is what interpretation will we need to make from the norms that are developed. It has been our experience that a considerable amount of confusion has resulted from the interpretation, or perhaps misinterpretation, of a set of particular standardized scores of these psychological screening instruments when the intended purpose of a specific set of norms had not been sufficiently delineated. Typical of this kind of confusion with respect to standardized scores are allegations that the screening test is rejecting a much higher percentage of inductees than is thought to be proper.

For the purpose of our discussion, I should therefore like to suggest that some of the anticipated interpretations that personnel policy makers and personnel administrators may need to make from these norms are the following:

1. Determining equivalence of legal standards of mental ability, as set by the Congress.

2. Understanding the significance of any score on these variables as it relates to a broad military population base, with possible projections under mobilization conditions.

3. Providing a basis for estimating the rejection rates which different cutoff scores are likely to yield.

There may be other objectives but it seems to use that these three mentioned are sufficiently fundamental to deserve our special attention.

Many of the methodological questions concerning problems of adequately standardizing the psychological instruments are more readily resolved when related to one or another of the above objectives. However, it soon becomes apparent that no one method is ideally suited to meet each and all of the objectives enumerated above.

Before discussing the procedures and methods that may meet out present requirements, it may perhaps be helpful to review briefly the methods used and problems that were encountered in the standardization of the Armed Forces Qualification Test. As you know, one of the consequences of the Unification Bill was the establishment of the Personnel Policy Board as a part of the organization of the Office of the Secretary of Defense. This Bill established a policy providing for the construction and use of a single Armed Forces mental test for the dual purpose of screening enlistees and inductees, and for qualitative allocation to the services, if required. The development of this first Armed Forces screening instrument was initiated in 1948 and necessary research and development conducted during that year and in 1949, culminating in the operational use of the Armed Forces Qualification Test on 1 January 1950. Our major concern with respect to standardization for that instrument involved primarily objectives one and two.

For the first objective, where equivalence of legal standards of mental ability, as set by the Congress, are required for improved alternate forms of established psychological screening instruments, the methodological question is relatively straightforward. Assuming the relationship between the new test and the established test to be relatively high, and it certainly is, a tie-back standardization on a sample population possessing the necessary range of the abilities in question seems sufficiently appropriate for securing the norms in order to determine equivalences of scores.

More specifically, the legal requirement contained in the Selective Service Act of 1948 provided that in the event of the screening of inductees, the cutoff score for acceptability would be 70 on a General Classification Test. Although the law was not specific, the context in which it was written suggested that a standard score of 70 on a test like the Army General Classification Test was the standard intended. More recently this Act was amended to reduce the standard to a standard score of 65. The Army General Classification Test was used extensively by the Army and Air Force during World War II. A quite similar test, the Navy General Classification Test, was used by the Navy. Score equivalences between the Army General Classification Test and the Navy General Classification Test were secured in 1947. This comprehensive body of data on the Army General Classification Test contributed to the establishment of standard score of 70, and later the 65, as the legislated standards of mental capacity to be met by new forms, specifically, the improved Armed Forces Qualification Test. For the purpose of establishing that equivalent standard, a tie-back standardization was all that was required, thus avoiding the necessity of painstakingly setting up a special sample population representative of one or another universe population.

However, it was recognized that we could not limit our goal to securing equivalence of some very specific score, such as the 65 or 70 established by Congress. Rather, we had to have a measure with greatest accuracy in that portion of the range of ability where accurate measurement is most important, namely, in the score ranges where cutoff scores are likely to be set by personnel

policy and operating people for voluntary enlistment or for induction. This requirement presented quite a challenge. Experience had shown that the Services desired to set the standard anywhere from a score which would reject somewhere between 10 and 20 percent to almost half of the total male population of military age. There was some concern whether the same item types could yield adequate results at standard score points 65, 70, 80, 90, or even 100. Recent experience does show that 65 is dangerously close to the level at which verbal type tests cease to discriminate efficiently differences in mental ability. That is something we have to keep in mind in future test developments. While we are on the topic of discrimination of mental ability, it should be pointed out that the same problems discussed in the standardization of the Armed Forces Qualification Test apply to tests used to evaluate mentally marginal personnel.

Further, the mere determination of equivalence (as for example, a raw score of 27 on the Armed Forces Classification Test being equal to the standard 65 and a raw score of 31 as equal to a standard of 70) does not provide information with respect to the second and third objectives that have been mentioned above. In other words, such a simple equivalence of scores would not throw light on the meaning of that score in relation to a broad military population which might be used under mobilization nor provide good estimates of rejection rates under varying conditions.

So now we come to the more pressing problems, which are related to the second and third objectives. The main distinction between the standardization problems where we are essentially interested in equivalents, and in the other objectives, that is meaningful interpretation with respect to a broad military population base and estimates of rejection rates, involves the reference population on which norms are to be established. If we keep this distinction clearly in mind I think we can avoid a considerable amount of confusion in this area.

As examples of the kinds of problems related to the reference population which had to be considered in the standardization of the Armed Forces Qualfication Test in order to provide greater meaning for the scores, we had such questions as: "Must such a reference population provide a sample of the population currently in the Armed Forces," or, "Must this reference population be a sample of the total potential military population under emergency mobilization conditions?" In this case, in order to provide information for the second objective, the latter type of population was decided. It was reasoned that successful military operations require planning for mobilization.

Another problem posed in connection with the reference population was the kind of sample which would be representative of the total potential military population. Was it necessary to do a complete sampling of the entire civilian population, or could use be made of the previous military population for which data were already available? After very considerable discussion and careful consideration of the points of view of the mathematical statisticians and the experts on sampling in other Government agencies, as well as in the military, it was decided to use the group of more than eleven million men in the Armed Forces as of December 1944 as the reference population in order to provide a meaningful base for scores obtained on the improved alternate form of the mental screening test. It should be pointed out that this reference population

includes only men who were found acceptable for military service and therefore excludes an unknown percentage of the male population of selective service age. However, World War II screening test standards, especially for the period preceding December 1944, were so low that estimates by a number of experienced technicians conversant with all the data agreed that probably not more than about 4% of the male population subject to selective service were not represented in this reference population. It is this reference population that was used by the research people to develop another scale primarily intended for a better understanding of the raw score on the Armed Forces Qualification Test, and for that matter it also yielded a better understanding of the scores on the forerunners of the Armed Forces Qualification Test. Careful consideration was given of the advantages of standard scores in general and the various specific standard score scales used by each of the Services in particular. It was concluded that the scores on the Armed Forces Qualification Test should be expressed in percentiles, and a policy decision to this effect was made by the Personnel Policy Board. The research technicians recommended this policy in the belief that personnel policy makers, operating people, and the Congress would more readily follow the meaning of any score in relation to this broad military population base if expressed in percentiles than in any of the standard score systems then in use. It is of course obvious that all of the percentile scores can be translated into any particular standard score system used by any Service.

Perhaps our greatest need for clarification is a standardization scale which would enable one to provide estimates of rejection rates with different cutoff scores on any of the screening tests for various conditions. It is readily admitted that the percentile scale based on the total military population will not provide a very effective estimate for that purpose since in many cases where rejection rates seem to radically vary from what is anticipated, it turns out that the particular group on which the rejection rate is computed is not at all representative of the reference population on which the test was standardized. In such cases the distribution of scores will vary widely from that characteristic of the total reference population. It should be emphasized that the percentile scale based on the standardization technique in which the broad military population serves as the reference base is not engineered to yield precise estimates of rejection rates. It is my proposal that we entertain the development of an additional standardization scale which will attempt to reflect more adequately the distribution of the specific populations that we may be concerned with during any one period. For example, with respect to AFQT-3 and -4 it may be desirable to secure, after its introduction, additional standard scales based on selective service registrants forwarded for examination over a period of time. Such scales, which could be called user scales, could then be used to forecast rejection rates for such populations. If these populations would remain sufficiently constant during specified extended periods, such as peacetime periods or partial mobilization periods, better estimates could then be made of rejection rates with different cutoff scores for such periods. Here again the method used for collecting these data and the kind of sample that is selected is of utmost importance in furnishing such estimates. However, it should be stressed that the stability of equivalent scores is likely to be far greater than the stability of rejection rates under any condition.

Related to this general question of rejection rate, it is sometimes pointed out that a test is not adequate because it fails too many people. We have already said that this may result from the level at which the cutoff score was set or from the nature of the group to which the test is applied, or both. The current standardization percentile scale does not provide a precise basis for estimating rejection rates. If such estimates are required, an additional standardization scale based on the particular input examinee population to which the test is to be applied would have to be provided. Such an additional scale for this purpose was proposed earlier.

Another source of much operating difficulty is the fact that the motivations of test administrators and those taking tests are subject to marked variations depending on the administrative consequences of the test scores. These variations can be readily illustrated by this chart. (Chart 2). You will notice that there are marked deviations of the operational test results from those obtained under conditions usually prevailing when tests are standardized. This problem has received considerable attention by the operating and the research people. Some of the steps that have been taken, including the placing of personnel psychologists in the various Armed Forces Examining Stations, seem to have resulted in a marked improvement on that point. This chart (Chart 3) based on data gathered from one of the large examing stations is an example of the kind of operational improvement that is possible.

Another important question in connection with the standardization population involves its composition with respect to enlistees and inductees. Will norms based on an enlistee population be applicable to inductees and vice versa? Or will norms based on a specified composition of inductees and enlistees be applicable to an examinee population in which the ratio of enlistees to inductees changes? It is possible that there may be greater motivation on the part of enlistees than inductees to do better on the screening test - although this difference may be less in times of war when public desire to serve is aroused. This difference in performance will not affect equivalences of scores appreciably, since it is reasonable to expect fairly consistent performance on the reference test and the new test within the same testing session.

However, in the standardization methods which are aimed to yield estimates of rejection rates, differences in motivation for enlistees and inductees complicate the problem. The proportion of enlistees varies from time to time, which would make the total enlistee-inductee rejection rate vary accordingly. In order to minimize the effect of this periodic variation, standardization would have to be accomplished separately on an enlistee applicant population, and a pre-inductee population. Data describing the test score distributions of each of these populations would have to be gathered over a long enough time period to minimize periodic fluctations. Different cutoff scores, then, may have to be set for enlistees and inductees.

The extent of this difference in enlistees and inductees is not known at present. Therefore, the necessity for considering the two groups separately in standardization cannot be determined. It is true, however, that differential treatment will considerably complicate the standardization process. Further

research on possible differences in performance as affected by motivation should be conducted before any attempt is made to consider these groups separately in standardization.

In summary, it is my recommendation that when we concern ourselves with the standardization of psychological screening instruments of the Armed Forces that we keep the objectives clearly in mind and consider the possibility of various types of standardization to meet these objectives. This is especially significant when one considers the difference in cost for the various possible methods that could be used. For the next forms of the Armed Forces Qualification Test, it is my belief that equivalences and percentiles can again be established as they were before, but that in addition we should consider securing a scale which will serve as a better basis for estimating rejection rates at various cutoff points for the various populations under the conditions anticipated.

Chart 1. RAW, PERCENTILE AND STANDARD SCORES FOR THE ARMED FORCES QUALIFICATION TEST

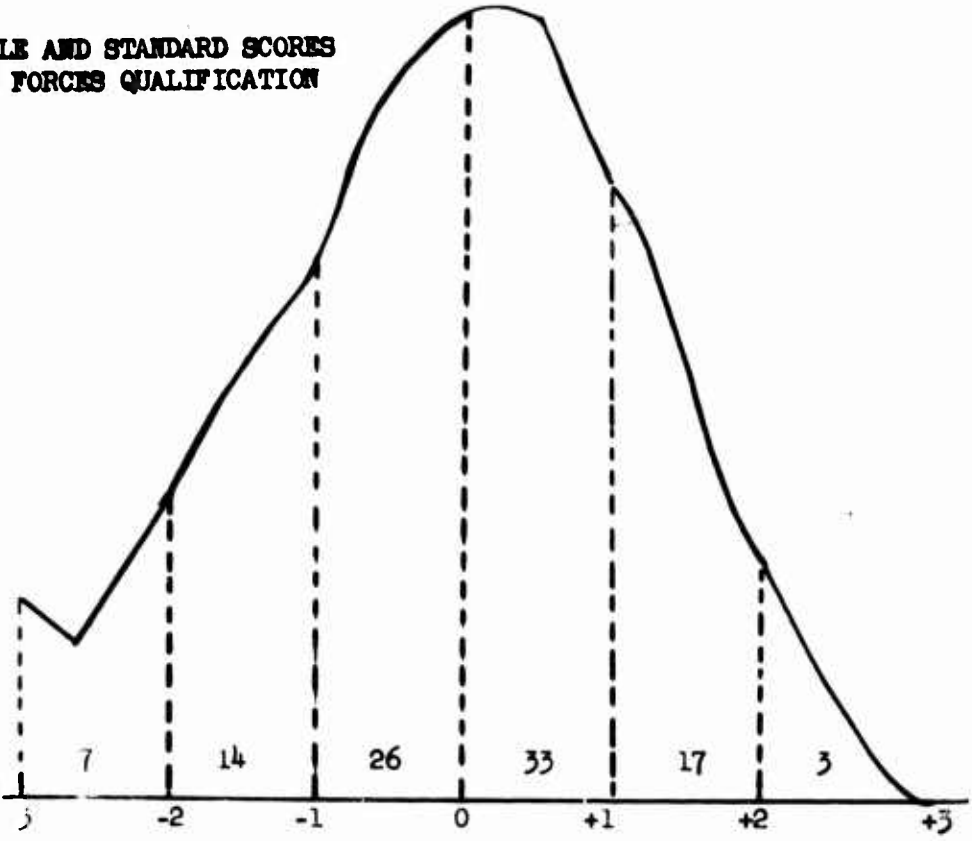

Chart 2. AFQT SCORES OF SAME 1,000 ENLISTED MEN FROM OPERATIONAL AND STANDARD ADMINISTRATIONS

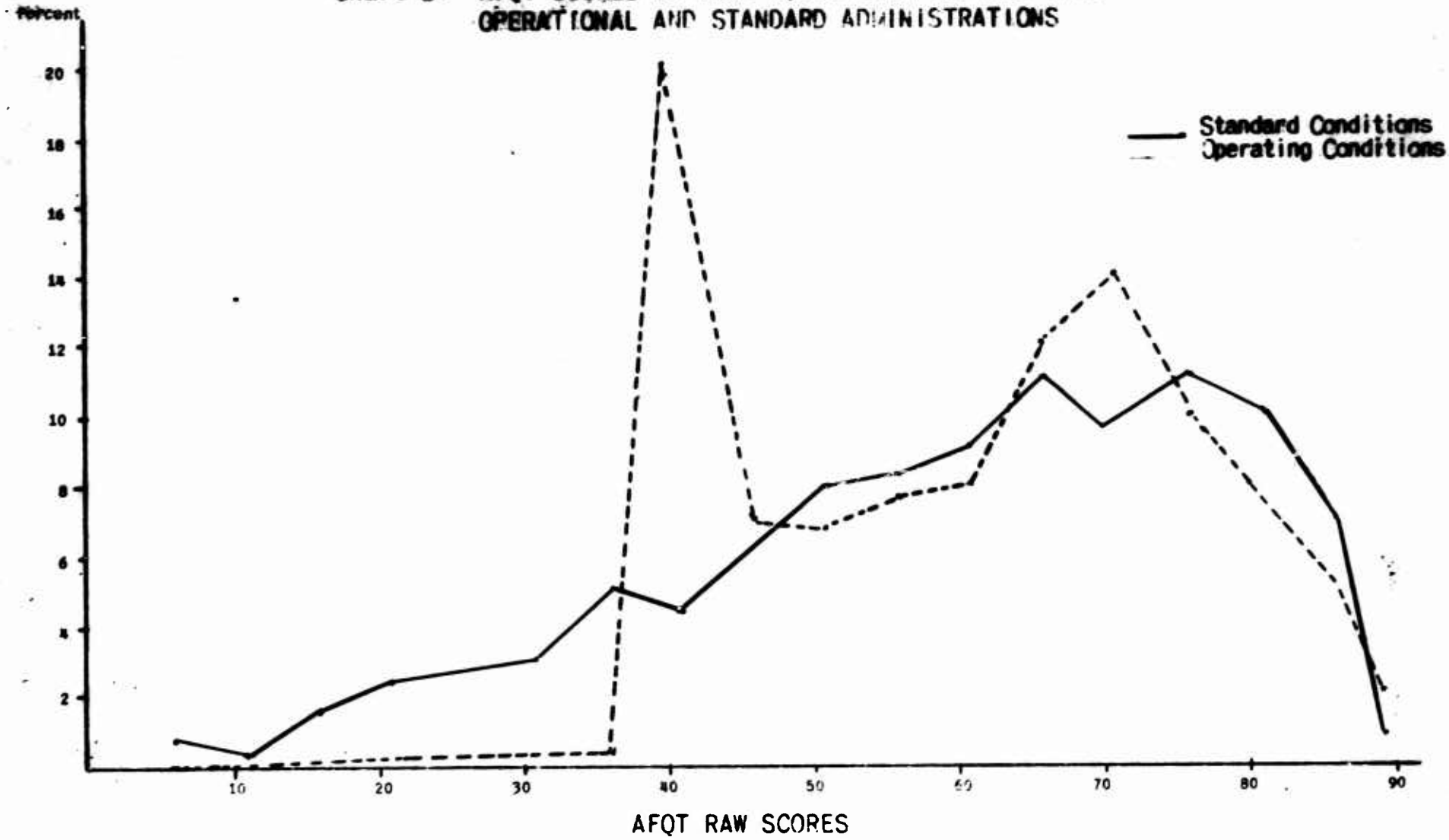

Chart 3. AFQT SCORES OF ENLISTEES AT ONE ARMED FORCES EXAMINING STATION