Defence Science Journal, Vol. 4, No.3, July 1954, pp. 99-106

STANDARDISATION OF GROUP TESTING OFFICERS' ASSESSMENT.*

ABSTRACT

This is the first investigation that has been made to find out the degree of standardisation of Group Testing Officers' assessment. The object of this project was to find out the exact degree of reliability and comparability of markings that has been achieved in the G.T.O. field and express them quantitatively. An experiment was accordingly designed and carried out at two places N and S. Three Group Testing Officers took part in the experiment at each place. At N, they observed simultaneously five batches consisting of 49 candidates and assessed them independently and separately by awarding them actual marks out of a fixed total. Similarly, 36 candidates were observed and assessed at S. The main finding is that there exists a sufficiently high order of reliability between the assessments of different Group Testing Officers, the co-efficient of reliability being 0·68, 0·68, and 0·75 at N and 0·80, 0·85 and 0·93 at S. It has been found during the analysis that the correlation co-efficient which is usually computed to measure the amount of reliability, fails, however, to measure the amount of identity between two sets of markings. For this purpose, a new co-efficient, termed as the co-efficient of comparability, has been found out. The computed values of this co-efficient generally show a similar trend as the correlation co-efficients with a few exceptions. Difference in average marks awarded was calculated for each pair of Group Testing Officers. None of them, barring one, was statistically significant.

Introduction

A Services Selection Board consists of three members, the President or the Deputy. President who acts as the Chairman of the Board, the Group Testing Officer and the Psychological Officer. During their three days stay in the Selection Boards, the candidates are given certain tests. These comprise, among others, out-door individual and group tasks given by the Group Testing Officer, who usually has a good scope to assess a candidate's abilities in the social field and his individual and physical capabilities as applied to the dynamic factors, specific factors of leadership and general aspects of personality. Though a Group Testing Officer's assessment is essentially subjective all attempts are being made to bring it as near to objectivity as possible.

Object.

This is the first investigation that has been made to find out degree of standardisation, that exists in G.T.O. assessment. Efforts have been and are

*This paper was prepared in the Psychological Research Wing of the Defence Science Organisation.

still being made to achieve as much objectivity as possible in this particular field of assessment by the introduction of Rating Scales, central training of the G.T.O.s, following a standardized technique using standard tests etc. The object of this project was to find out the exact degree of reliability and comparability of markings that has been achieved in the G.T.O. field and express them quantitatively.

Method

An experiment was accordingly designed and carried out at two places N and S. Three Group Testing Officers took part in the experiment at each place. At N, they observed simultaneously five batches consisting of 49 candidates and assessed them independently and separately by awarding them actual marks out of a fixed total. Similarly 36 candidates were observed and separately assessed at S. These actual marks formed the basic data for the analysis.

Henceforth in this paper, the three G.T.O.'s who took part in this experiment at N will be denoted as G.T.O. N1, N2 and N3 respectively and those at S as S1, S2 and S3.

Statistical Analysis

The data pertaining to N and S have been analysed separately. The actual marks in percentage awarded to the candidates by the three G.T.O.'s at N are shown in appendix 1, and the appendix 2 gives similar details about the experiment conducted at S.

Some idea about the comparability of the sets of marks as awarded by the different Group Testing Officers can be obtained from their means and standard deviations given in Tables 1 and 2.

Table 1. Centre N.

	Mean	Standard deviation
(1)	(2)	(3)
GTO N1 ..	26·10	9·87
GTO N2 ..	27·96	11·48
GTO N3 ..	27·04	8·87

Table 2. Centre S.

	Mean	Standard deviation
(1)	(2)	(3)
GTO S1 ..	30·00	11·31
GTO S2 ..	29·89	12·83
GTO S3 ..	29·92	11·44

The extents to which different G.T.O.'s scattered their marks differed very little in case of centre N and still less in case of the other centre. None of the differences in overall standards of different G.T.O.'s as measured by the differences in their means were statistically significant except that between G.T.O's N1 and N2, which came out as 1·86% being significant at 5% level. These differences are shown in column (5) of Tables 3 and 4 and they in a way indicate the amounts of relative overall strictness or leniency between pairs of G.T.O.'s.

Reliability

Comparisons of averages and standard deviations of marks however, do not give a complete picture of the amount of similarity in marking by the different Group Testing Officers. To determine the reliability of the G.T.O. assessment, product moment correlations between the assessments of pairs of assessors were computed. They are given in column (2) of Tables 3 and 4 be low.

TABLE 3
Centre N

(1)	r (2)	Lower 95% limit of r (3)	C (4)	Difference of means (5)
GTO N2—GTO N1	0·84†	0·75	0·81	1·86*
GTO N2—GTO N3	0·79†	0·68	0·75	0·92
GTO N1—GTO N3	0·79†	0·68	0·78	—0·94

TABLE 4
Centre S

(1)	r (2)	Lower 95% limit of r (3)	C (4)	Difference of means (5)
GTO S3—GTO S1	0·96†	0·93	0·95	—0·08
GTO S3—GTO S2	0·88†	0·80	0·88	0·03
GTO S1—GTO S2	0·91†	0·85	0·91	0·11

All the six correlation co-efficients are significant and quite large. Keeping in view the subjectivity inherent in such methods as G.T.O. asessment, the reliabilities as measured by these correlation co-efficients seem to be quite encouraging and satisfactory. Since the correlation co-efficients in Tables 3 and 4 are based on only 49 and 36 candidates respectively, they are liable to considerable amounts of sampling fluctuations. Lower 95% confidence limits of these correlation co-efficients have therefore been computed based on their sampling distributions and shown in column (3) of the above Tables. The odds are five in hundred against this limit to have a value greater than the correlation co-efficient in an infinitely large sample. Thus it has been found that even if we take into account the effects of small sizes of samples experimented with, the reliability figures varied from 0·68 to 0·75 at N and 0·80 to 0·93 at S, showing thereby that the reliabilities achieved in this particular ype of assessment are quite satisfactory.

*Denotes significant at 5% level.
†Denotes significant at 1% level.

Coefficient of Comparability vs. Coefficient of Correlation

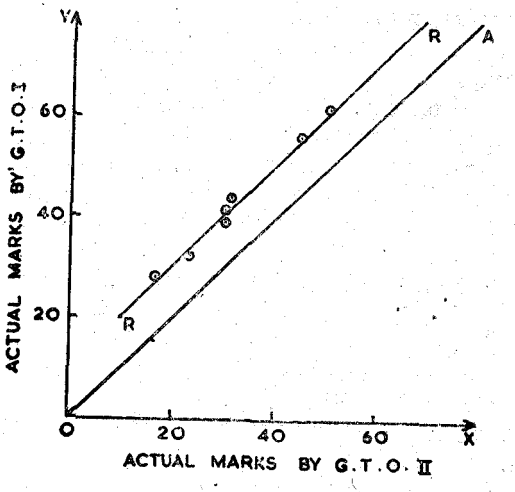
The usual method of measuring the degree of reliability of the marking by two assessors is to calculate the product moment correlation co-efficient between the marks awarded by the two. The value of this co-efficient, which can vary from —1·0 to +1·0, gives an idea of the amount of standardisation achieved. A negative correlation co-efficient will mean that if a candidate receives above average mark by one examiner, he is more likely to get below average mark by the other. An insignificant co-efficient of order zero will mean that the two markings are quite irregular, there being no sinigficant relationship between the two. A correlation co-efficient of the order of +0·5 or higher in a large sample will mean sufficient standardisation between the markings whereas a correlation co-efficient of unity will be achieved only when there is a perfect linear relationship between the two.

The correlation co-efficient, however, has a limitation. Instead of measuring the amount of identity between two sets of markings, it measures, in a way, the degree of similarity, in the relative placement of the candidates by the two assessors. The following figures and the diagram below will make it clear.

The actual marks in percentages awarded by two imaginary G.T.Os. have been taken as an illustration.

G.T.O.	Actual marks in % of candidate No.							Average marks
	1	2	3	4	5	6	7	
G.T.O. I	60	28	54	32	41	38	42	42·14
G.T.O. II	47	18	43	22	30	30	28	31·14

The diagram below shows the scatter of these points when plotted in a graph.



In case the sets of marks were identical for the two Group Testing Officers, all the seven points representing the seven candidates would have comexactly on the line OA, which makes an angle of 45° with both the axes of reference. Thus it is seen that these marks are far from being identical but as these points lie very closely about a straight line which has been denoted by RR in the graph, the correlation co-efficient is also very large and in fact it is equal to +0·99 in this case. This line RR is one of the regression lines in statistical parlance, from which the sum of squares of the deviations of these points measured along y axis is the least, Denoting by σ_e^2 the variation about the line RR, and σ^2 the total variation of Y, the correlation co-efficient r may be written as

$$r = \left(1 - \frac{\sigma_e^2}{\sigma^2}\right)^{1/2}$$

When σ_e^2 is negligible in comparison to its total variance as in the case illustrated here, r will be large, whatever the position and orientation of the line RR may be.

Thus it is found that high value of r fails to guarantee the identity between two sets of makings and it merely points out that there is a strong tendency of linear relationship between the two. Even a correlation of +1·0 does not mean that the marks given by the two assessors are identical, but it only means that they show a perfect linear relationship. There still remains a wide scope for differences in overall standards as in the case illustrated, where there is an average difference of 11% between the marks awarded by the two assessors. Also, the difference in one mark to one assessor may not have the same significance to another assessor.

It is thus obvious that the correlation co-efficient r fails to measure satisfactorily the degree of unanimity in assessment, if by that, is meant the identity of marks awarded by the two assessors. But if the aim is only to order the candidates according to their officer potential on the basis of the G.T.O. tests alone and to select a certain proportion from the top for imparting training, the correlation co-efficient will satisfy all the criteria of an efficient and consistent measure of the degree of standardisation achieved in the G.T.O. assessment.

The selection procedures, now followed in the Selection Boards are, however, entirely different. Here the candidates are selected if they get a total mark equal to or greater than a given critical score, otherwise they are rejected. The mark awarded by G.T.O. is only one of the constituents of this total mark. Thus it becomes imperative that the standardisation in G.T.O. assessment should aim not only at the similarity in ordering of the candidates within a group, but also at the identity of the marks awarded by the different G.T.O's. Correlation co-efficient, as has already been shown, fails to measure this identity successfully.

A new co-efficient has, therefore, been found out to measure this degree of identity. It has been termed as co-efficient of comparability and has been denoted by C. Technically speaking, it is similar to the correlation co-efficient 'r' except for the fact that the variance σ_e^2 in this case is not taken about

the regression line RR but about the 45° line OA. Calling it σ_d^2, the equation now will be

$$C = \left(1 - \frac{\sigma_d^2}{\sigma^2}\right)^{\frac{1}{2}}$$

$$= 1 - \tfrac{1}{2}\left(\frac{\sigma_d^2}{\sigma^2}\right), \text{ expanding and neglecting higher powers.}$$

$$\text{Again } \sigma_d^2 = \Sigma_i \frac{(x - y_i)^2}{n} = \frac{\Sigma_i d_i^2}{n}$$

Where n=total number of candidates,

and, $d_i = x_i - y_i$ = difference between the marks for the ith candidate.

Writing the pooled estimate $\sigma_x \sigma_y$ for σ^2 where σ_x and σ_y are the standard deviations of the respective sets of marks awarded by the two assessors, the equation finally becomes

$$C = 1 - \frac{\Sigma_i d_i^2}{2n\sigma_x\sigma_y}$$

Thus it is clear that the comparability co-efficient will be unity when and only when each candidate has received same mark by the two assessors, that is, only when all d_i's *i.e.* differences will be equal to zero.

This new co-efficient can also be expressed as

$$C = r - \frac{(m_x - m_y)^2}{2\sigma_x\sigma_y} - \frac{(\sigma_x - \sigma_y)^2}{2\sigma_x\sigma_y}$$

Where r is the corresponding correlation co-efficient and m_x, m_y are the averages of the marks awarded by the two assessors respectively.

From the above equation it is obvious that this co-efficient C will all ways be less than the correlation co-efficient r and they will be equal only when the means and standard deviations of the two series of marks are respectively equal, and the above conditions are necessary for the regression line RR to coincide with the 45° line OA but not sufficient. The difference between the two co-efficients otherwise will depend upon the amount of lateral shift in the position of the centroid of the observed points and the amount of angular rotation about the centroid, that will be necessary to make the line RR coincide with OA. The maximum positive value that this new co-efficient can take is also +1, but unlike the correlation co-efficient, it has no limit in the negative side.

Comparability Coefficients

The co-efficients of comparability have been separately calculated for each pair and given in column (4) of Tables 3 and 4. Even these co-efficients have sufficiently large values varying between 0·75 to 0·95, proving the existence of a sufficient degree of standardisation in the G.T.O. assessment. It has not been possible to take into consideration what will be the amount of the effect of sampling fluctuations on the values of the comparability co-efficients as its sampling distribution is not yet known.

APPENDIX 1

Centre N

Actual marks Awarded in %

Candidate No.	1	2	3	4	5	6	7	8	9	10	11	12
G.T.O. N_1	15	30	20	32	22	23	28	10	29	65	22	22
G.T.O. N_2	18	40	22	40	22	25	40	20	25	75	33	30
G.T.O. N_3	18	24	27	42	18	24	40	13	27	49	31	22

Candidate No.	13	14	15	16	17	18	19	20	21	22	23	24
G.T.O. N_1	30	23	25	25	24	20	28	29	25	36	28	29
G.T.O. N_2	30	20	32	28	27	22	32	20	24	42	18	30
G.T.O. N_3	27	18	22	29	20	27	33	22	20	36	24	33

Candidate No.	25	26	27	28	29	30	31	32	33	34	35	36
G.T O. N_1	22	20	30	33	25	19	18	10	22	20	25	33
G.T.O. N_2	20	18	35	42	26	18	20	10	15	15	15	40
G.T.O. N_3	16	18	40	40	22	16	18	20	27	29	20	42

Candidate No.	37	38	39	40	41	42	43	44	45	46	47	48	49
G.T.O. N_1	24	30	24	20	60	45	32	20	20	33	15	17	27
G.T.O. N_2	25	38	30	28	50	30	40	20	30	40	15	10	25
G.T.O. N_3	20	33	27	24	53	35	27	27	27	31	20	18	24

APPENDIX 2

Centre S

Actual Marks Awarded in %

Candidate No.	1	2	3	4	5	6	7	8	9
G.T.O. S_1	60	22	42	25	27	30	32	45	32
G.T.O. S_2	70	22	40	24	25	24	35	45	28
G.T.O. S_3	60	20	50	20	30	32	30	49	42

Candidate No.	10	11	12	13	14	15	16	17	18
G T.O. S_1	20	27	22	33	68	22	20	37	36
G.T.O. S_2	20	42	22	30	70	20	21	35	35
G.T.O. S_3	22	29	27	29	67	20	21	25	30

Candidate No.	19	20	21	22	23	24	25	26	27
G.T.O. S_1	23	17	25	35	55	27	20	30	25
G.T.O. S_2	20	15	28	33	58	28	22	28	23
G.T.O. S_3	20	25	20	28	50	30	22	25	30

Candidate No.	28	29	30	31	32	33	34	35	36
G.T.O. S_1	23	28	25	20	21	30	25	24	27
G.T.O. S_2	22	23	25	20	20	30	27	24	23
G.T O S_3	25	20	25	20	27	25	30	24	27