An investigation into perverting influence in catogory assessment scales, with particular reference to precision in the objective assessment of attitude. Also a comparison between category scaling and an application of paired comparisons.

Swindell, David John

An investigation into perverting influences in category assessment scales, with particular reference to precision in the objective assessment of attitude. Also a comparison between category scaling and an application of paired comparisons.

David John Swindell B.Sc. (Leeds)

M.Sc Thesis 1972

D.J.Swindell Thesis 1972 Abstract

An investigation into perverting influences in category assessment scales, with particular reference to precision in the objective assessment of attitude. Also a comparison between category scaling and an application of paired comparisons.

- - - - - -

The writer's approach to precision in attitude scaling is derived from an exposition of propoganda, defined in terms of attitude manipulation.

The attitude-action discrepancy is discussed in relation to cultural stereotypes and stereotypical behaviour.

Category scaling procedures are critically described and discussed, together with paired comparisons, Gutman scalogram analysis, and questionaire methodology.

Experiments are carried out to amplify and clarify the perverting influences on category scaling as discussed. Also experiments are carried out to provide data on the relative precision in item selection, and ease of application of category scaling methodologies and an application of paired comparisons.

Multiple partial ranking, an adaptation of paired comparisons, is discussed in relation to attitude scaling methodologies.

CHAPTER I

Introduction

One specialised and very practical aim in psychology is the use of words and images in the manipulation of public opinion. It cannot be denied that public feeling and corporate conscience do suffer radical changes in the long term. This has its roots in history, and is difficult to predict in direction or extent beyond a relatively short period. However, the business world has evolved techniques for influencing the man in the street, which increasingly undermine our belief in our ability to think freely for ourselves. Psychology can claim but marginal credit for these discoveries, though psychologists have undoubtedly contributed to their refinement and increasing sophistication.

The production of materials intended to influence people in both political behaviour and spending patterns is best considered a creative art. This cannot be considered remarkable since any scientific process aimed at synthesising affective statements depends on a knowledge of semantic relationships which are highly susceptible to time and social/ linguistic evolution; and this knowledge does not yet exist.

A dictionary definition of a word is quite inadequate for the form of 'semantic engineering' implied above. A combination of dictionary definitions and word counts could be of considerable value, though word counts fail on two important points; firstly they represent written rather than spoken words, and secondly they do not list the frequencies of the various meanings attributed to similarly spelt words. What is more a spoken word count would also have to differentiate between the social

class/occupation/education specific meanings of similarly pronounced words, let alone dialect variations and fashionable jargon. It is hardly surprising that no spoken word count has yet been published.

It is left to basically non-scientific specialists to produce material intended to influence people's opinions and behaviour, and most of this cannot be considered more than merely short term manipulation. Thus attempts at political behaviour manipulation may have the sole aim of getting a particular person elected, and sales promotions are recognised as being of short term value only, being abandonned usually after a certain period beyond which, experience has shown, they produce diminishing marginal returns against effort.

I have avoided the use of the word 'attitude' in the above. Definitions of attitude are various, and are frequently put forward with a particular purpose in view. Allports definition in terms of 'a mental . . . state, exerting . . . a dynamic influence . . .' is perhaps the least exceptionable so far formulated yet it remains undeniably an inferred intervening variable and is inextricably tied to the nature of the instrument used in measuring it. Eysenck related opinion complexes with attitude, they being its verbal manifestations, while Jung had taken a more classical definition, demonstrating introvert and extravert attitudes and relating them to personality types. Murphy went still further in considering attitude an integral part of personality.

In any attempt to examine attitude, then, several fields of study must be encountered and catered for, more especially since research tools bear remarkable similarities in all these fields, and indeed in many cases may be quite interchangeable.

Notwithstanding the variety of depth, or of the dependancy or independancy attributed to 'attitude', it is commonly used as a blanket term for a subject's predispositions to act or express himself in various ways. However, as intimated above, definitions of attitude may be as

various as the tools used to measure them. These tools are almost invariably language based survey methods, and this work will examine one particular genus of survey methods commonly called category scaling procedures.

The investigation of mental states exerting dynamic influences, or of predispositions to act in certain ways is perhaps one of the most intriguingly difficult in social psychology. In pure psychophysics the results tend to justify the assumption that whatever variable is controlled is in fact the primary cause of any results obtained. In investigating attitude only physical manifestations of attitude-potent stimuli can be varied, and this is the crux of the problem. Whatever aspect of a test situation it is chosen to vary, either it is assumed that at least the true attitude quality of the change is known, or variations are made quite randomly with a view to isolating common factors against common results. In the first instance it may be argued that the variables involved in this kind of experiment are too complex for such naive assumptions, and in the second only physical differences in the total stimulus may be analysed and the 'true' underlying factors may remain concealed, to be brought to light only in the framework of some established theory.

To examine this further let us consider interpersonal attitude-potent situations as analogous to phrases of spoken language. As the precise meaning of a phrase or its units may be disputed, so a situation may not, and perhaps cannot, mean the same to any two observers. The linguistic analogy may, with caution, be taken yet further in that linguistic sophistication can convey and detect more variations and precise nuances.

We may now consider again the mechanisms of propoganda. I will begin with the controversial assertion that logical argument in the conventional sense is virtually worthless as a potent opinion/attitude influence. There is however a perverse logic in the effectiveness of repetition as a

promotion method. A product or person is seldom sold to the public on their physical integrity, or on unbiassed factual evidence of their potency; rather the public is told that such and such is the truth, and there is ample evidence for the effectiveness of such assertions. What is the superiority of one dogmatised truth over another? The answer shows where the propoganda artist succeeds over the scientist. There can surely be nothing inherent in tigers' tails to affect the octane rating of petrol, or a blue boy blowing bubbles the power of soap, yet these have proved very effective in the past in selling their related products. The answer lies in the genius of an individual or a group in realising how, within the context of a particular social/ historical instant or epoch, words or images (actual or conceptual) may be juxtaposed to represent a dogmatised assertion powerful enough to achieve for itself the identity of a perversely logical truth. The art in this is the concrete, though seldom conscious realisation of some underlying factor in the psychological makeup of a majority of a population to be propogandised, and the relation to it of physical and mental images which maximise the play on this underlying factor. In productive terms it is useless to play to a psychological factor which will not evoke some positive action, so we must return to a factor which implies a propensity to act, and thus to some manifestation of affective attitude.

Considering the complex and often intangible influencers of opinion we may question not only the definition of attitude, but also the quality measured in survey methods. An ideal experiment would consist of the surreptitious engineering of a situation in such a way that subjects' responses may be precisely monitored without the obvious presence of an experimenter or indeed an experiment. To some extent this is obtained at election times or during intensive sales promotions. We want to know the precise effect of a campaign designed to influence subjects' actions, and from these we indirectly infer their attitudes, and election returns or

sales figures can be considered this result if prudently analysed. However, there are few occasions of direct interest to the social scientist where such an approach may be used, though some will be described below. An attractive and far easier method is to present hypothetical situations to a subject and to ask him to describe how he would respond. The fallacy of this approach has been amply demonstrated, and the lessons from its frequent failure slowly and often expensively learned (see Chapter 2). Failure of this method may be attributed to failure of the hypothetical situation posed to conform fully with the actual situation in life. Perhaps the hypothetical situation fails to represent all the concrete facts of a life situation, or perhaps a subject is too easily able to alienate the situation from himself, and to respond in terms of only a limited number of attitude sets representing a felt ideal or desirable approach. The precise mechanism of this failure is outside the scope of this paper, the only pertinent fact being the failure per se.

Let us look at the nature of the test items described above. Words and figural or figurative images are presented to a subject. The images are the principal controlled variable in a complex psychosocial experiment, other variables being the situation in which and the people to whom they are presented. By holding the population constant the effect of varying the mode or situation of presentation may be studied. The resulting findings may lead to questionable conclusions about the nature of the presentation situations, since it may not be valid to assume a particular situation is the same for all members of any otherwise apparently homogeneous group. To amplify this, humans are notoriously fickle, and although a particular group may respond with a great deal of consonance in certain situations, which may in fact define the group, there is no reason to suppose they will agree in all their responses. Child rearing practices, for example, in any social class are variable, and even within

a single family it may not be assumed that two children will suffer the same formative experiences. The same is true throughout life, and it would be a poorer world if it were otherwise, albeit a more amenable world for the scientist. We have then, a group of individuals who recognise themselves as dedicated to common aims, but who have come to behave in this way by virtue of varied pressures and experiences. We may study gross differences between groups, but the individual differences within groups seems to have been largely ignored, even though such an approach might yield a far more useful definition of the group. This approach is derived from Maslow's exortation to study normal rather than abnormal populations. Extreme subgroups of a population may be easily defined superficially, and results obtained will invariably demonstrate differences between such groups, but their relevance to 'normal' people is questionable. We may conclude that we cannot assume any two otherwise similarly opinioned people to have the same attitudes. They respond to similar situations in similar ways, but we may not assume that the events are in fact perceived similarly.

A situation or an event may be perceived differently by otherwise apparently similarly reared and opinioned individuals by virtue of two closely related factors. Firstly a person's education, explicit and implicit, in both formal and family senses, must preclude or emphasise many words and individual meanings of a word. And secondly the reactions of peers to certain words or modes of expression impress a subtle, though nonetheless potent, meaning. Thus we return to the semantic content of a word or unit of meaning or expression.

I would conclude from this argument that methodologies based on the presentation of hypothetical situations only begin to measure one minor parameter of the complex social stimulus - social response paradigm.

The term 'semantic engineering' was mentioned above, and although its meaning might appear immediately comprehensible it is worthwhile

expanding on it here.

Industrial work study and job evaluation techniques have been evolved for synthesising the time aspect of time and motion studies. In most craft skills there are many operations which are common to different tasks; these might be tightening a nut, applying callipers, or simply moving from one part of a work bench to another. Over the years accurate mean times for such operations have been established which eliminate much tedious and imprecise stopwatch work. However, variations in the precise time for a specific operation are accepted according to the context of the operation. Thus ambient temperature, time of day, length of task or simply the previous movement may all influence the time taken.

A system analogous to that above is what is meant by semantic engineering. Statements may be synthesised from basic units, and the precise meaning of these statements would be known under any defined contextual situation. The context consists of many potent aspects of the situation in which a subject finds himself, as for instance the subject's own psychological makeup and the presence of certain influential people or things in the environment. Thus mere dictionary definitions or word counts are useless in themselves, the only approach of any value being Osgood et al's work on semantic differentiation, though difficulties of their approach will be discussed later.

Let us speculate on how such a system might be operated. Every word or standard unit of meaning or expression would be defined in terms of multiple vectors, each representing a form of 'action press' dependant on certain aspects of situation context. This would specify the most probable actions to be elicitted by the presentation of the unit of expression in any definable context.

However, the presence of any such unit of expression would also constitute a modification of the experimental context for the subject,

and the presence of two or more such units forming a coherent statement would require the computation of a final vector for a situation whose context is defined by the subject, incidental aspects of the environment, and the semantic vectors of the constituent parts of the statement itself, all mutually dependant and modifying one another. The mathematics of such a statement in any defined situation would theoretically enable precise prediction of a subject's response. However, the system presents the dilemma of chickens and eggs, or horses and carts. Semantic engineering as here described requires techniques for measuring attitude at a level of precision which we have not yet attained; and the construction of such techniques would require a knowledge of attitude which perhaps they alone could be instrumental in compiling.

As yet it is only productive to think of semantic engineering in terms of attitude measurement rather than propoganda.

Attitude scale construction consists of the selection of certain items of a number constructed which will exhibit a clear cutoff point in relation to a defined attitude. Thus as we approach subjects with increasing degrees of a certain attitudinal attribute, we would hope to find a very short range of potency of this attribute within which responses to a particular item will exhibit a sharp qualitative or quantitative change. In fact the most common method of item selection is to take responses or subject groups recognised as representing polarised opposites in terms of the attitude to be studied. Even accepting a wide spread of opinion within such groups, an item shown to differentiate accurately between them cannot be assumed to conform to our model, since there may be a very large range within which no responses have occurred. The result of diffuse changeovers rather than sharp cut-offs is that banks of such items will fail to measure subjects in the mid range with the accuracy they do at the extremes.

The semantic engineering approach would aid this procedure in two

ways. An experimenter would no longer have to construct a large number of statements to be sorted, vetted and reduced by a great band of judge/subjects, since the nature of the attribute expressed in all the statements would be controlled within tolerable limits. Also the precision of the cutoff would be implicit in each statement's semantic vector geometry. Ideally therefore, opinion/attitude scale construction, as well as the production of propoganda materials, would be a matter of synthesis by computer in hours or only minutes, rather than months or even years of hard physical effort; and the tedious processes of calibration and validation would be a thing of the past.

This is the background against which the present thesis was formulated. So many factors are involved that it would be the lifetime work of a team to resolve all the problems. In this thesis only one topic is approached, and even then the problems would appear to multiply rather than diminish. However, even though the background would appear ambitious, the results are cogently applicable to more mundane present-day problems in category scaling methodology.

CHAPTER 2

Studies relating to attitude as an abstract quality

In 1934 Richard T. LaPierre published the classic paper "Attitudes vs. Actions". Between 1930 and 1932 LaPierre travelled extensively about the United States in the company of a young Chinese couple. Out of 251 establishments visited, services were refused on racial grounds in only one, while a questionaire presented six months later in the same and other establishments showed positive expressed racial discrimination in 90% of both the establishments visited and a control group of others not visited. In such circumstances LaPierre quite rightly questioned the validity of questionaire data, emphasising the unreality of the hypothetical situation as presented in the questionaire.

A study by Corey (1937) reached similar conclusions to LaPierre's after examining possible reasons for the lack of validity studies in questionaire methodology. He showed how results on attitude questionaires tended to be accepted as valid by the experimenter when they conformed to a class stereotype, and particularly to that of the experimenter.

Despite this evidence of the attitude/action discrepancy at such an early date Linn (1965) demonstrates the continuing tendency to ignore it. Tarter (1966) accounts for the discrepancy by suggesting the subject reacts to what he believes the situation to be, and hence Tarter finds the discrepancy entirely acceptable.

A parellel approach to the problem is given in studies which take an oblique approach to attitude. Instead of attempting to identify individual attitudes or to hold constant and quantify a particular

attitudinal quality, items are presented in such a fashion that the subject may be influenced indirectly in his assessment by an indefinable level of 'prestige' attributed under that guise. Thus Lambert, Hodgson, Gardner & Fillenbaum (1960) and Anisfeld, Bogo & Lambert (1962) showed reactions to carefully controlled items of spoken language. Results of these studies demonstrated the potency of community stereotypes, some culturally less favoured groups expressing some stereotypes of their own group more strongly than did the dominant group.

Francès (1963) showed the effect of attributing a work to a known prestigious author; this was to raise the general opinion of the quality of the work and to reduce the judged superiority of works judged better. As level of education decreased, so the effect of the attribution of the work to a famous author decreased. A similar and more specific finding was shown in Greenberg (1966), when preference to newspaper vs. television newscoverage was studied. Sex and education were shown to be quite independant and potent influences on the believeability attributed to the different media, while age was shown to be a bad predictor. Anast (1966) showed clear relationships between mass media preferences and Jungian personality type.

Finally, referring again to Lambert, Hodgson, Gardner & Fillenbaum (1960) and Anisfeld, Bogo & Lambert (1962) above, Gardner, Wommacott & Taylor (1968) showed the stability of nationally held stereotypes across subcultural groups, suggesting this stereotype is independant of individual attitudes towards the group evaluated.

From these studies it may be seen that many intangibles are involved in the attitude/action process. Conventional methods in which groups identified with certain attitudes are examined, may fail to account for say educational or personality differences, and thus some factors apparently independant of the major attitudinal characteristics might appear blurred where they in fact exert definite underlying influences. Also some

reactions, notably to national stereotypes, show a great degree of independance from national sub-group identity, and thus no differences may be found where radical differences do in fact occur.

On the whole the above studies used techniques other than the category scales to be described below. The conclusions may therefore be taken as independant from other perturbing influences yet to be demonstrated. Where used, quantitative methods were not assumed to represent absolute quantities, rather they were mere indicators of differences.

We have then a body of evidence to suggest that there are influences on expressed attitude which might either completely mask a particular attitudinal attribute, thus leading the investigator to conclude no relationship, or else they may emerge unpredictably during investigation of an attribute perhaps closely though not obviously related. In both cases attempted measurement of a defined attribute may not necessarily lead to the correct prediction of a certain course of action.

Measurement of defined attitudinal attributes has become reasonably precise. Before predictions may be made we must have before us some instrument with an acceptably low degree of imprecision. Whether valid or not we may then relate success of prediction with the precise measure. Apart from the action/attitude discrepancy described above, the nature of LaPierre's 'hypothetical situation' as presented in questionaires merits further study. From the introductory remarks, the writer assumes that response formats represent integral parts of the questionaire items, and an effect analogous to that which might be predicted from this assumption has been demonstrated in Blankenship (1940), in which response formats were shown to affect the range of validly drawable conclusions quite apart from the question wording itself. This might be due solely to the response option as presented, however this paper will attempt to isolate such perturbing influences of a purely psychophysical

nature.

Precision in attitude measurement is not enough. One random group may well show responses similar to another's, but perturbing influences inherent in the scale construction might originate at a purely psychophysical level, and thus similar scalings would be similarly perturbed.

Before extending this argument let us examine the attitude instruments most commonly proposed; and particularly the category scaling procedures.

CHAPTER 3

Attitude measurement

There are many techniques used in the study of attitudes. For many reasons inventories and questionaires are preferable to interview and analytic techniques. Their obvious advantages are in the realm of quantification while a by-product is a sometimes spurious air of scientific validity.

Most attitude assessment techniques can be considered specialised psychophysical methods. The category methods to be described below are essentially custom built for the purpose. However, a central topic of this essay is the method of paired comparisons which came to the field of attitudes from a long career as a more general psychophysical technique from what may be called the 'physicopsychologists'. It was first and is still used as a tool in the study of completely tangible physical entities such as colour and sound tones. Its application to social psychology as well as that of the category scaling methods is described in this chapter.

The method of paired comparisons involves the comparison of every item of an item population 'm' with every other item. In practical terms every subject must make a binary decision on each of ${}_mC_2$ pairs of items. It is usually desirable that each pair be administered but once to each subject, to minimise the work load on the subject and to avoid complications in the procedures used in the analysis of the results.

The data thus gathered are used to calculate the relative scale values of the items on an hypothetical psychological continuum. The

method of obtaining these relative scale values is based on Thurstone's Law of Comparative Judgement explained below:-

When confronted with a series of items to be judged, one must consider an attribute common to all the items and which they possess to different degrees. If one item appears to possess more of this attribute than another then it is said to have a higher discriminal process. The discriminal process is the process within us by which we react differently to different items, objects or specimens.

The most common discriminal process experienced to a particular object by a particular subject is called the object's modal discriminal process for that subject.

The separation between the discriminal process at any instant and the modal discriminal process is called the discriminal deviation.

The standard deviation of the discriminal deviations evidenced by a subject to an object is called the discriminal dispersion of that object for that subject.

The discriminal difference is the difference between the evoked discriminal processes for two objects in the same judgement.

The law may now be concisely stated thus:-

$$\bar{S}_i - \bar{S}_j = z_{ij}\sqrt{\sigma^2_i + \sigma^2_j - 2r_{ij}\sigma_i\sigma_j} \qquad \left(z_{ij}\sqrt{S^2_i + S^2_j - 2r_{ij}S_iS_j}\right)$$

where $\bar{S}_i$ & $\bar{S}_j$ are the modal discriminal processes of the two objects I & J in the same judgement, σ_i & σ_j are the discriminal dispersions of the two objects, r_{ij} is the correlation between the discriminal deviations of the two discriminal processes in the same judgement, and z_{ij} is the normal deviate of the proportion of judgements (I > J).

Consider the equation above applied to m items examined in a paired comparison situation. For each pair there will be an equation like that above in terms of $\bar{S}$, σ, z, and r. There will be $(m^2 - m)/2$ expressions. These expressions must be used to deduce each value of $\bar{S}$ from $\bar{S}_I$ to $\bar{S}_m$, each value of σ from σ_I to σ_m, and every value of r_{ij}. There are thus

$2m + (m^2 - m)/2$ unknowns, but only $(m^2 - m)/2$ expressions. In this form the law of comparative Judgements is therefore insoluble.

Thurstone proceded to make certain assumptions about the relative values of the unknown quantities in order to make the equation soluble. The more assumptions that may be made the easier the solution:-

(I) r is practically constant throughout the stimulus series for the single observer.

(II) when a group of observers perceives an object, the quantity of attribute that they ascribe to it is normally distributed on the continuum of perceived attribute.

(III) $r = 0$

(IV) $\sigma_i = (\sigma_j + d)$ where d is so small that d^2 may be ignored.

(V) all discriminal dispersions are equal.

These assumptions simplify the equation to:-

$$\overline{S}_i - \overline{S}_j = I.4I42\ z_{ij}$$

And this is called Thurstone's 'case V' of the law of comparative judgement since it was the result of the fifth simplifying assumption.

The purpose of the law of comparative judgements is to deduce the scale values of the items examined on a defined psychological continuum such that these values are at least linearly related to the 'true' modal discriminal processes.

In a series of papers, Mosteller (1951 a, b & c) presents perhaps the most concise examination of Thurstone's law of comparative judgement.

Most significantly he shows that where the assumption of equal discriminal dispersions holds for all but one of the items scaled, the result is for all the items to be properly distributed except for the aberrant one. He also derives a formula for the amount of error due to this aberrant item where all the assumptions of case V are accepted.

Greenberg (1965) proposes a modification of the law of comparative judgements to accommodate judgements of 'equal' or 'no difference' rather

than forcing choices. In effect the modification approximates the ad hoc procedure of distributing the 'equal' judgements equally between the items in the pair. However, on purely practical grounds the technique appears useful.

With quite different considerations in mind, and independant of the restricting assumptions of case V, Kendall (1948) presents a method for examining a set of paired comparison data to determine whether they represent a significantly non-random set of choices. Thus m items administered to n subjects by paired comparison may be perceived by all the subjects as representing more or less the same thing, and if so their choices will show some degree of concurrance. A random set of choices would not be meaningfully scalable by the law of comparative judgements, though results might be obtained, and should show as random by Kendall's method. Mosteller (ibid) presents a similar technique based on the law of comparative judgements case V. Kendall's method examines each subject's set of paired comparison responses for 'circular triads'; that is for judgements of the mature $A > B$, $B > C$, $C > A$, which appear inconsistent with the notion of an undimensional continuum of items. Judgements should be of the nature $A > B$, $B > C$, $C < A$. The use of triads as a unit of quantification is challenged by Slater (1961) who maintains that since each triad consists of three pairs and any pair appears in m-2 triads, triads are not strictly independant of one another. The result is to weight some inconsistent responses more than others. Slater's solution is in terms of the actual number of these inconsistent responses insofar as they may be identified. This is a useful tool making as it does no assumption about the unidimensionality of the items judged. However, it has certain practical disadvantages in computation rendering it more cumbersome than Kendall's with no real advantage in power efficiency.

In constructing a scale by the method of paired comparisons a number of statements is collected each of which expresses some sort of comment

upon the object the subjects' attitude towards which we hope to measure. These statements are presented in pairs to each of a group of judges, and each judge is asked to say which of each pair represents a more favourable comment on the object of the statements. Application of the law of comparative judgements establishes the relative scale values of the statements on a continuum of perceived favourableness-unfavourableness of comment.

To apply the final scale, a subject is invited to agree or disagree with the sentiments expressed in the scaled statements and his score is taken as the median of the scale values of the statements he endorsed.

The statements used in this form of scale are usually specialised and selected or constructed, and few in number. There are relatively few statements because of the disproportionate increase in ${}_mC_2$ (the number of pairs to be presented to each subject) with increasing m. In fact few workers have attempted to use more than thirty items while a more usual number is less than ten. Thus the method has rarely if ever been used as a method of item selection but has tended to use all the items examined in its final form as an attitude scale.

In an attempt to cope with the disproportionate increase in ${}_mC_2$ with increasing m, methods have been proposed whereby instead of responding to each pair separately, small samples of items from m are put in rank order by the subjects. Durbin (1951) proposed the use of conventional balanced incomplete blocks and Youden square designs which offers a partial solution to the problem of ranking a large number of items. Schucker (1959) proposed the administration of groups of only three items at a time and indicated methods similar to those of Durbin for generating the 'triads' to be administered. The important difference between these techniques was that Schucker envisaged the triads as consisting of three individual pairs to be analysed as simple paired comparisons, while Durbin used the method to gain as estimate of the population rank orders

of the items.

Slater (1965) proposed a more general method than that of Schucker, in which, subject to certain logical limitations, up to 28 items may be so construed that all the ${}_{m}C_{2}$ pairs are presented as constituent pairs within groups of 3, 4 or 5 items to be ranked. Consider an item population m; all the constituent ${}_{m}C_{2}$ pairs must be presented to a subject within a number of groups all size x, i.e., ${}_{x}C_{2}$ pairs at a time. Thus only $\frac{{}_{m}C_{2}}{{}_{x}C_{2}}$ presentations of groups of x items need be made for all the pairs to be administered. The logical limitations are:-

$$\frac{m-I}{x-I} \text{ must be an integer}$$

$$m > x(x-I)$$

$$\frac{{}_{m}C_{2}}{{}_{x}C_{2}} = \frac{m(m-I)}{x(x-I)} \text{ must be an integer}$$

Slater gives no indication in his paper as to any convenient techniques for generating his presentation groups, and indeed in a private communication admitted no knowledge of any such techniques, prefering 'the mental exercise rather like a game of patience' of generating them by hand. However, Dr. D. Fairley of the Department of Mathematics at Durham University indicated an iterative method whereby the necessary presentation groups (which may be greater than Slater's maximum of x=5) can be easily generated providing m is a prime, and which with some adaptation lends itself to certain other values of m.

Sjöberg (1965) examined four different methods of scoring paired comparisons data. He used case V, case IV with an estimate of the discriminal dispersions for the items derived from the data, successive intervals analysis, and his own correlational scaling. He showed that case V produced a far worse least squares fit than the other methods, while successive intervals scaling was marginally better than the rest.

A criticism of the use of paired comparisons in advertising research

by Blankenship (1966) emphasises the purely relative nature of scale values derived from paired comparisons, in that no absolute value can be attributed to a paired comparison scale value since it is dependant on the total item sample examined. He shows how the use of the 'winner' in paired comparisons can lead to confounded expectations where all the options examined are absolute 'losers'.

Subsequent to his work on paired comparisons, Thurstone developed a procedure for scaling statements which was far less cumbersome and provided an objective means of item selection (Thurstone & Chave 1929). This came to be called the method of Equal Appearing Intervals. Since it is still widely used and since in many respects it is representative of other scaling procedures, it will receive lengthy attention in this paper.

To produce an equal appearing interval scale firstly gather together as many statements as possible descriptive of the object under examination. They should represent opinions about the object ranging from most to least favourable through neutral opinions. These statements are then rated by a number of judges (as many as 900 and as few as 16 judges have been used) on a 9 or 11 point scale from least to most favourable with the centre category defined as 'neutral', The responses are cumulated across categories over subjects to produce an ogival curve of responses to each item. The median of the responses for an item is taken as that item's scale value.

Thurstone then selected twenty or more items more or less equally spaced along the scale. Where there was a choice the item with the smallest interquartile range was chosen since Thurstone considered this a measure of ambiguity.

When used as a test these selected items are presented to the subject who is invited to agree or disagree with each item in turn. His scale value is then taken as the median of the scale values of those

items he endorsed.

As implied in the title this method assumes that as a judge sorts items into the various categories he perceives the category widths to be the same throughout the continuum. This might be considered a questionable assumption. In order to account for this Thurstone developed the Method of Successive Intervals in which items are rated as in the method of equal appearing intervals, responses are cumulated across categories, and a procedure is carried out which normalises all the resulting ogival curves by finding the normal deviates corresponding/each successive cumulative proportion, thus shifting category boundaries. In essence the extreme categories are widened. The scale values of the items are the median responses on the new 'unequal' interval continuum and selection, administration and scoring are otherwise as in the method of equal appearing intervals.

The literature on Thurstone scaling is extensive, and it would be neither practical nor useful to summarise it here. The following represent the most recent information on the contemporary use and theory of the methods.

Upshaw's 1962 article was an important step in the study of the relationship between judges' attitudes' effects on scale values derived from their judgements. Upshaw contended that a judge's ratings of a set of items depended on the range of these items and their relationship to his attitudinal position at the time of judging. He shows that where an item population only covers one extreme of a judge's attitudinal 'span', that is where no items representing the judge's position are presented, then scale values derived from such data approximate Upshaw's t-condition in which the total span of probable statements are presented; however where the extreme of the item population is excluded there is a displacement of item ratings towards the 'missing' end of the continuum. Someapparent inconsistencies in this study were shown by Manis (1964)

to be comprehensible on an assimilation & contrast model when it is considered that Upshaw's subject population was mainly distributed at the pro end of the continuum.

Manis' interpretation was partially endorsed in Upshaw's re-examination of part of his 1962 data (Upshaw 1965) in which he investigated the affect on any final Thurstone scale of varying judgemental perspectives. He shows that variations may be understood in terms of 'judgemental language', and implies the superiority of this interpretation over Manis' assimilation-contrast model. Upshaw states that attitude scale values can not be said to be invalidated by variations in judgemental reference scale parameters.

Robinson (1965) indicates the effects of 'level of information' on judgements made in Thurstone scale construction. The logical nature of the statements judged is considered, and the number and nature of the judges anchors mentioned.

Bruvold (1969) reports findings concerning the relation between equal appearing intervals and successive intervals scaling which conflict with much previous work and theory. He shows a linear relationship between scale values where previous expectations and published evidence had indicated a curvilinear relationship.

Shortly after Thurstone's work Likert developed

The Method of Summated Ratings

A number of statements are dichotomised into favourable and unfavourable categories. Judges are then asked to respond to the statements on a five category continuum in which the categories are defined as:

strongly agree
agree
uncertain
disagree
strongly disagree

These categories are then weighted so that unfavourable statements are weighted 0 for strongly agree through to 4 for strongly disagree while favourable statements are weighted 4 for strongly agree, etc. In this way subjects with the most favourable responses will achieve the highest possible score.

Likert found that scores based on weights assigned by the naively simple integral method correlated 0.99 with scores based on weights assigned by the far more complex normal deviate system of weighting. The normal deviate method requires a normal transformation of the cumulative response distribution. These values are made all positive by subtracting the largest negative value from all the rest and are then rounded to the nearest whole number. This procedure is carried out for every item in order to determine the weights to be assigned to each specific response to every item separately. For most purposes therefore the simple integral method may be employed with no appreciable loss of information and a considerable saving in effort.

In order to select the items to be used in the final test the scores of defined highest and lowest scoring groups are examined (e.g., highest and lowest scoring quartiles). These groups' responses to each item are examined by some form of item analysis and those items shown to distinguish significantly between the high and low scoring groups are selected.

It has been said that a score on a Likert scale has no bearing on a subject's true attitude outside some reference group since the meaning of the 'uncertain' category is not assumed by Likert to be of zero valence as in Thurstone's work. However, as a means of distinguishing between two groups or of measuring change it is considered invaluable; though this might indeed be said for any similar category technique.

Edwards & Kenney (1946) examine the evidence available on the influence of judges' attitudes on Thurstone scaling, relative test-retest

reliabilities of the two methods, relative ease of construction, and the usefulness of a judging group in scale construction. With reservations they conclude that the Likert method would seem less tedious and more reliable than Thurstone's, judges' attitudes have no effect on final scale values, and a judging group is unnecessary.

Barclay & Weaver (1962) carried out a quantitative analysis of some of Edwards and Kenney's conclusions and showed them to be substantially correct.

Guttman's scalogram analysis

In the methods so far described there is no attempt to verify the unidimensionality of items originally examined or those finally chosen. The consequence of this is to allow items into the final test, whose scale values exist on an entirely different dimension from that of the majority of the items. Thus a certain quantitative response to item A on dimension X might mean something quite different from the same quantitative response to item B on dimension Y. If a set of items can be shown to exist on different dimensions then in Guttman's terms they are not scaleable. Guttman's method gives a figure descriptive of the degree to which a subject's responses are exactly reproducable from his scale score. A coefficient of reproducability of above 0.85 or 0.9 means that the test is good on Guttman's criterion.

Consider a number of statements A-Z. In a unidimensional scale a subject scoring positively on item A will score at least as well and no worse on items B-Z while a positive score on item K will mean worse scores on items L-Z, etc. The degree to which a test conforms to this model is Guttman's coefficient of reproducability. A certain score on any item should convey some information about the subject's scores on the other items.

The importance of Guttman's techniques is the recognition of the concept of non-unidimensionality in tests. However it is not strictly

speaking a method for scaling test items, nor does it provide means for selecting items to be included in a test. Edwards & Kilpatrick (1948) attempted to account for these failures by combining Thurstone, Likert and Guttman techniques in

The scale discrimination technique

Scale values are calculated as in the method of equal appearing intervals. The 50% of items with the largest interquartile ranges are eliminated from the test. Phi-coefficients are calculated for each statement from the scores of defined high and low scoring groups above and below a defined neutral score. Statements are then selected from each interval or fractional scale interval of the Thurstone scale values on the basis of the highest phi-coefficients. Edwards and Kilpatrick divided their six category Thurstone scale scores into half scale intervals and from each of the seven half categories containing scores they selected the four items with the highest phi-coefficients. These 28 items were then ranked in order of their Thurstone scale values and were divided into parallel versions of 14 statements each by taking alternate items. These versious were then applied to a fresh group of subjects who were instructed to express their agreement or disagreement to them on a six category defined scale and Guttman coefficients of reproducability calculated. Both of these were above 0.85 in their study.

No comparative work appears to have been done on the scale discrimination technique.

Semantic Differential

Although this is not strictly an attitude assessment technique, it is based on methods similar to those described above and in certain circumstances lends itself to ready adaptation to the problems of attitude assessment. Osgood Suci & Tannenbaum (1952) set out to establish some quantitive measure of meaning. Their intention was to produce a form of 'controlled association and scaling' procedure. A

subject is given a concept to evaluate on a number of scales. The scales consist of pairs of bipolar adjectives, i.e., opposite in meaning, placed at either extreme of 5 or 7 category response formats. The subject is required to indicate the intensity and direction (if any) of the concept's association with each scale. Thus if the concept is closely related to the meaning of one of the adjective pair, the subject marks that extreme category. If the scale is meaningless or irrelevant to the concept, or if the concept can be said to possess none of the attributes specified by the scale then the subject marks the centre category, etc. This is in contrast to free association in which a subject responds with all the related words that spring to mind when the stimulus is presented. In free association there is no apparently valid measure of meaning since without some form of semantic differentiation there are no precise cues as to the qualitative and quantitative connotative meanings of the responses. Semantic differentiation provides specific responses to which the subject may be presumed to respond positively if the scale constitutes a natural response to the stimulus concept and vice versa, and the intensity of the response may be assumed to represent some measure of the probability of this response being ellicited in free association.

The use of bipolar scales stems in Osgood et al's work from the studies of Karwoski, Odbert and others from 1934-1944 on synesthesia and social stereotypy.

Osgood's first study in Measurement of Meaning (1957) was designed to examine the dimensionality of the semantic space, i.e., the nature of the coordinates against which an object may be plotted in defining its meaning and differentiating it from other objects. To do this a large number of bipolar adjective pairs was constructed in an attempt to represent all possible shades of meaning. Factor analysis on the results for 100 subjects evaluating 20 concepts on 50 scales isolated four factors

rotated into simple structure maintaining orthogonality. The last factor extracted accounted for only 1.5% of the total variance, hence no more factors were extracted after this and the last factor was ignored. The three dominant factors were labeled 'evaluative', 'potency' and 'activity' from the most highly loaded adjective pairs on each of the factors. To some extent Osgood confesses that the sampling methods used for concepts and scales did not produce unbiased items. However, similar results were found using a specific factor analysis method (D-factorisation) designed to eliminate the dependance on concepts to be evaluated, which in themselves might be biased as a sample. Further studies duplicated these results when more representative samples of scales were used (Thesaurus analysis in above), specialised concepts were evaluated (using sonar signals), and specialised material was responded to by naive subjects. The result of these studies was the delimitation of the three primary factors used in defining semantic space, evaluative, potency, and activity. Only after this work was it possible to construct a semantic differential ~~(sem diff)~~ scale.

The object of the semantic differential is to determine the meaning of a concept in terms of its valence on each of the orthogonal factors defining the semantic space, and hence its 'position' in that semantic space, for a selected subject or group. Consider a group of subjects who consistently consider a concept good, strong and active; these must represent a population considerably different from one which considers the same concept bad, weak and inactive. Of equal importance is the relative configurations of constellations of concepts. A group which consistently considers 'wife' and 'mother' as being very close in meaning must represent something different from one which consistently places them in diametrically opposite quadrants of the semantic space.

To construct a semantic differential scale to examine a defined set of groups of subjects, firstly gather together a set of concepts

to be evaluated taking care that they are such that there will be a fair degree of disagreement between the groups to be examined as to their meaning, while they remain familiar and unambiguous to the subjects individually. This process may involve selective sampling, but Osgood stresses the value of discernment and good judgement. The choice of scales against which the concepts are to be judged is far less haphazard, since Osgood et al's work indicates those scales which are highly loaded. It is only necessary to select a number of highly specifically loaded scales to represent each dimension of the defined semantic space. It would be unsatisfactory to select only the most highly loaded scale for each factor since they are all poluted to some extent. By selecting a sample of scales to represent each it is possible to sum for each factor and thus gain a reasonably unpoluted estimate.

Considerations in selecting scales are that they must be relevant and meaningful to the concepts to be examined. This may involve selecting a dimensional factor other than the three dominant ones, for which factor loadings for an unbiased population are rather small, but which nevertheless gain meaning in a specific situation.

To administer the sementic differential, each of the concepts is paired in some form of presentation with each of the scales. A subject is asked to indicate by placing a mark in one of the five or seven categories of each scale what the concept 'means' to him. If the concept is closely related to the adjective at one end of the scale, then the subject must place a mark in that extreme category. If the concept is neutral or unrelated to the defining adjectives they must mark the middle category, etc.

The responses are integrally weighted, e.g., from 0 to 6 or from -3 to +3 for a seven point scale, and the resulting data summed over scales for each factor. The mean weighted response for concept C on factor F_I represents one coordinate of C in the semantic space defined

by factors F_I ... F_J ... F_N. Analysis may be carried out by vector geometry to establish the significance of the various points in the semantic space allotted the concepts, and if no more than three factors are used it is a simple matter to construct a 'three dimensional' graph of the relative positions of the concepts.

An early published study similar to semantic differentiation by Jones and Thurstone (1955) examined a number of words and phrases on a single 9 category response continuum of 'greatest like' through 'neither like nor dislike' to 'greatest dislike'. They demonstrated the generally normal distribution of the majority of the items using Thurstone's successive intervals method. However, while most of the distributions were normal some exhibited severe skew within the centre range of the continuum (skew at the extremes is predictable), and others exhibited bimodality. This has a bearing on Thurstone's 'ambiguity' criterion in equal appearing intervals.

While Jones and Thurstone's paper was limited in only examining one dimension of meaning, its lack of recognition in the subsequent works of Osgood, etc., does not appear merited.

In analysing concept clusters within the semantic space, Osgood and Luria (1954) relied on visual inspection to a great extent, and expressed regret that no adequate mathematical procedure was then forthcoming. Hofman (1967) provides a technique for determining the significance of the difference between the positions of concepts and concept clusters within the semantic space. This analysis is only applicable to the analysis of raw linear distances between points in the N-dimensional semantic space, and is not applicable to the analysis of scalar quantities; it is thus cumbersome and wasteful of information, rather like carrying out a number of t-tests on data arranged for analysis of variance.

The application of semantic differentiation to attitude scaling is

treated in Brinton (1961). Brinton selects those assessment scales from an application of the semantic differential on which significant differences between defined high and low scoring groups for the items assessed are found. He showed a high Guttman coefficient of reproducability for a scale thus derived. Brinton suggests that a generalised attitude scale could be constructed by selecting only highly evaluative adjective pairs in this way.

Another application by Barclay (1964) simply sums subjects' responses over all the scales irrespective of their factor loadings. This rejects the very concept of semantic differentiation and can hardly be considered a true application of the method.

Hudson (1967) approximates far more closely Osgood's original work by examining the relative use of evaluative terms by arts and science biased groups of boys. His final data were in the form of graphical representations of the significantly associated clusters for the two groups. Hudson suggests the comparison of individual results with clusters such as these as a basis of some form of index of aptitude for the occupational biases of the groups typified by those clusters. It is but a small step from an occupational index to one of attitude.

Questionnaire methodology cannot strictly be said to fall within the range of category scaling procedures, however where closed ended response formats are presented similar influences may be assumed to act. Also the lack of standard methodology often leads to the unintentional but misleading misuse of techniques of construction and analysis.

Adapting A. N. Oppenheim's (1966) summary of the processes involved in Questionnaire methodology, the following phases must commonly be gone through in the construction and application of a questionnaire:

1. Establish the aims of the study and where applicable state the hypotheses to be tested.

2. Review literature, enter into discussions with knowledgeable

and interested parties. Where applicable, state further hypotheses and reject redundant ones.

3. State hypotheses from above in operationally testable terms, considering step 8 below.

4. Select, adapt or design techniques to examine the above operational hypotheses. Specify sample to be studied.

5. Pilot study.

6. Revision of design in light of the pilot study, (returning to step 4 if necessary).

7. When instrument is satisfactorily refined, carry out the necessary field work and data collection.

8. Data processing and analysis. This should be so designed from the outset that all the hypotheses are implicitly tested in the analysis.

9. Write up study, drawing conclusions directly from the data and comparing with other studies.

This account is obviously quite flexible, and represents to some extent experimental methodology in general, apart from the design of parametric studies, in which no hypotheses are stated (though they may be implied), and an effort is made simply to describe without forcing the data to limited conclusions.

An essential part of questionnaire methodology is the form of the statement or question put to a subject. Considerations in this are:-

a) Logical form of the question must be clear and comprehensible.

b) Wording of questionnaire must not imply consistent bias to one point of view.

c) Form of presentation of question must be constant for all subjects.

d) Permitted responses must bear a complete relationship to question asked.

Questionnaire methodology borders on the realm of artistic

creativity, and any further discussion will be left to the following commentary.

CHAPTER 4

Criticisms of category scaling procedures

The most direct evidence against category scaling procedures concerns the phenomenon variously referred to as response set, bias or style. This is a tendency to produce stereotyped responses. Where subjects are not required to respond within defined categories, but must indicate a position within a homogeneous response continuum, response set may be governed by Gestalt. Thus certain parts of the response continuum may be more easily identifiable even though their identities are specified by extrinsic qualities of the response situation. Taves (1941) showed that dots arranged in an identifiable pattern (a circle) were consistently judged less numerous than the same number arranged randomly. By varying the relationship between figure and ground, Bevan, Maier & Helson (1963) found quite divergent estimates of a constant number of beans in varying sized jars. A similar effect was found by Bevan & Turner (1964) varying the size of the 'frame' around random arrangements of dots.

Granberg & Aboud (1969) confirmed the conclusions of Mokre (1927), demonstrating a linear relationship between judgements of visual numerousness and visual density. However, they failed to control order effects in their study, and in doing so demonstrated the effects of order on such responses.

The most important form of response set, for the purposes of this paper, concerns response behaviour to a categorised response continuum. However, the writer could find few direct references to such work. Mathews (1929), examined responses to a Likert-type format, where five

responses were each defined by verbal terms, (dislike very much, dislike, indifferent, etc.), and found a significant discrepancy in responses when the order of presentation of the response categories was reversed ('like' on the left or on the right) and the shift was greatest where subjects had least pronounced views. This suggested a stereotyped response towards a certain end of the response continuum relative to the subject (possibly 'handedness').

Philip (1947) examined an 11 category Thurstone-type scale, and found variations in scatter between subjects, and 'foci' where individual subjects tend to mass responses.

Many more studies have treated dichotomous responses such as yes-no, agree-disagree, etc., or simply series of items of which only those which a subject considers 'correct' or 'agrees with' etc., have to be checked (Bennet, Seashore & Wesman 1947, Humm & Wadsworth 1943, Lorge 1937, Rubin 1940, Vernon 1949). According to Cronbach (1950) these generally demonstrate problem solving methodology sets, which may influence any derived scores. Cronbach concludes that forced-choice or paired comparison methods should be used wherever possible, and where it is not possible an attempt should be made to induce the same set in all subjects and a final response set score derived to identify any possibly invalid results.

Rorer (1965) reviewed the literature on the topic and concluded that response set was at best a minimal artefact. Rundquist (1966) challenged Rorer's conclusions and indicated an item-response model including response set which may aid in determining an index of bias. This study shows that responses are affected by the form of a statement; thus a negatively stated statement is likely to elicit a different response than the same statement put positively. He also showed differing response characteristics to items with different contents, demonstrating the difficulties of comparing scales with different

contents.

Das & Dutta (1969) examined Soueif's Personal Friend Check List, in which set is equated with response rigidity, and showed a quadratic relationship between age and rigidity with a minimum at about 24 years. They also showed a positive correlation with religiosity and hypnotic suggestion, and a negative correlation with intelligence.

Finally, we consider the effects of question order. Standardised tests have a constant order of presentation of questions. It is accepted that this order was the one with which the test was standardised, and thus any effects of order are integral components of any final score and are thus irrelevant. Granberg and Aboud (1969 above), found a significant order effect in their perceptual study which while it did not appear to affect their final results, nevertheless exhibited wide variations.

The popular belief in attitude survey methodology is that order is quite irrelevant, but the writer has been unable to find published evidence to justify this. Nor, unfortunately, has he found published evidence to the contrary. However, from pure psychophysics there is much evidence of both spontaneous alternation and repetition. Zwaan (1964) has adapted psychophysical methodology to the question of categorical responses. In experiments where there were choices of 2, 4 or 6 responses, he found significant alternation in the form of a less than random repetition of a previous response. These were not ordered categories, but were 'absolute' choices such as card suites and numbers on dice. However Zwaan points to the importance of this finding to question order in 'psychodiagnostics', and the paper appears particularly pertinent to problems of questionaire design.

Wagenaar (1968) comes to the opposite conclusion with a 2-choice verbal reaction task. These diverse findings serve to demonstrate the general confusion in this field, where e.g., choice reaction times show

definite recency effects but unpaced motor tasks ~~and~~ show alternation, and true randomness ~~are~~ is seldom found. It can only be assumed that within a test situation the nature of the task remains constant throughout and whatever response bias is acting acts constantly.

LaPierre's paper of 1934, though in some senses methodologically naive, raised a series of problems which subsequent test constructors and theoreticians have preferred to ignore, or at least forget. LaPierre showed that verbal statement of policy towards accepting orientals as patrons had no relation to actual behaviour in the situations as engineered. In fact whereas all but one out of 250 establishments did not refuse service, 118 out of 128 of those establishments stated six months later in questionaires that they would not serve orientals (i.e., 92%), and a similar percentage of establishments not visited responded in the same way. This may be considered an artefact of social response set in the same way as the psychophysical response sets demonstrated above.

Corey's (1937) masterful summary of the work done till that time, stressed the implicit assumptions of validity in scales with no effort at establishing behavioural measures. His study on attitudes to cheating showed a consistant near zero correlation between attitude and behaviour.

Tarter (1966) cites further studies to suggest that attitudes are situational rather than absolute, insofar as manifest behaviour may be observed (Kutner, Williams & Yarrow 1952, Minard 1952, Lohman & Reitzes 1954).

Tarter (abid) established testable Parsonian hypotheses to account for this phenomenon, but his experimental results were inclusive.

The implications of these findings are extensive. Contemporary definitions of attitude centre on the predisposition to act, and are thus linked with the prediction of behaviour (see Chapter 1). If prediction from current attitude measures fails, then either the

mechanism of prediction or the attitude measures must be at fault. The evidence presented above would appear to lay the blame on inadequate measures.

QUESTIONAIRES

Blankenship (1940a) listed examples of the use of apparently harmless words influencing responses in questionaires. He shows the influence of ascribing a point of view to a nationally known figure, the effect of emotionally toned adjectives, nouns and verbs as apparently harmless as 'involve', and the effect of social class-specific words. Where emotionally overtoned words are used, there is invariably a substantial change in the number of endorsements as compared with a similar statement put 'neutrally'. Where class-specific words are used, some portions of an intended subject population will simply fail to understand the question.

The above writer also carried out an extensive examination of the results of various wordings of questions using objective and subjective questions worded positively and negatively, and also positive objective questions containing a check list of responses (Blankenship 1940b).

All the actual questions used were current topics. Blankenship concluded that of the types of question he used, the most valid was the positively objectively stated question with a check list answer. However, he admits that this only began to approach the problem and more work was necessary before an accurate protocol for constructing questionaire statements could be evolved.

Further studies by Cantril (1940), Rugg (1941), Rugg & Cantril (1942) and Hyman (1944-1945) merely highlighted the problems of question wording without providing adequate solutions. One answer suggested in several articles of the period was a vocabulary of words, etc. it is unadvisable to use in defined situations. The scale of this work would have been formidable and it does not appear to have

achieved fruition. However, the recommendations of Rugg & Cantril (1942) appear quite constructive. They conclude that the stability of subjects' answers in the various question wordings is a function of the stability of his frame of reference and normative system. They recommend that questionaires should take into account the variations in individuals' normative systems by examining a variety of questions on the same issue. They also state their repudiation of Blankenship's (1940b) admittedly limited conclusions and suggest the use of a free answer question in some part of a ballot in order to sample the total population opinion on an issue, and recommend the use of split ballot techniques as a continual assessment of the above.

It is very probable that the growing interest in semantics leading to semantic differentiation in the following decade superceded the rather cumbersome questionaire methodologies suggested in the early '40s. It is certain that very little could have, or indeed has, been added to this work since. However, a slightly different area of study popular in the mid '50s and undergoing a resurgence of interest since 1962 sheds light on questionaire methodology through a wealth of sociological studies of the process of interviewing sumarised by Manning (1967). Interviews are structured on the social norms and values of the interviewers, in that social-linguistic categories are tacitly assumed to represent the actual situation under consideration. Thus opinion is commonly divided between 'for', 'against' and 'don't know', and people falling within these groups are assumed to be homogeneous in their opinions. But social class subsumes various systems of shared meaning, styles of affect and mood, which differentially attribute significance to pregnant pauses, raised eyebrows, etc. By failing to account for such differential situational patterning interviewers fail to represent multiple perspectives on reality, and the full range of responses possible even from the individual.

The interview may be described as a 'two-game situation', in which the 'information game' rewards the interviewee by having his ideas accepted and recorded, and the 'ingratiation game' by gaining the interviewer's approval. The 'good' interviewer is the one who can maximise the rewards to the interviewee of the information game.

Mentioned also by Manning is the method of participant observation which is useful in situations where a group might feel its integrity threatened, and react by producing evasive responses and gambits. Participant observation is conceived as a situation where participant observers gather information and learn the language both verbal and behavioural of a group. It cannot lend itself to large groups and in some respects may still be subject to the above criticisms. Even the observers are subject to the mores and perceptual sets of a particular social and educational background, and during their participation they may also cathect some of the values determining both explicitly and implicitly a 'threatening situation', and hence find themselves unable (or unwilling) to express or even identify certain conflicts.

A psychological assessment of the same problem by Cattel & Digman (1964) identifies seven 'perturbing influences' which may confound or polute values derived from survey & interview methodologies:-

I) Instrument factors such as number & nature of permitted responses, and nature of stimuli, e.g., all biased to an extreme or mentioning a single emotionally potent group.

2) Individual factors such as response set.

3) Stimulus modulation to the subject by differential presentation due to interviewers, test situation, etc.

4) Stimulus modulation to the observer. Role effects on S due to relationship with O. Situational effects on O's scoring of S's responses.

5) General personality differences outside direct influence on perception of stimuli & responses.

6) Effects of inadequate definition or variable meaning of item content. Also inadequate specification of the test situation and inadequate representation of the density of content variables to be examined (mainly a mathematical effect).

7) Effects of misperception due to stereotyped perceptual sets, e.g., social stereotypes, private images, cliches, etc.

SEMANTIC DIFFERENTIATION & SEMANTICS

Considerable work has been done both directly on and concerning semantic differentiation. Jones and Thurstone's paper (1955) has already been referred to. Its importance was both in the early date of the work and in the bimodality it evidenced in the meaning of certain items. It will be noted that this very characteristic would probably have eliminated those items from an equal appearing intervals scheme, but the conceptual set of the paper predisposed the acceptance of bimodality as manifesting a legitimate response parameter.

Mordkoff (1963) challenged the basic assumption in semantic differential that nominally opposite adjectives are functionally opposite, and demonstrated the falsity of this assumption with several commonly used adjective pairs. This replicated the findings of Ross & Levy (1960) and Terwilliger (1962) and unlike those papers it used a method approximating to semantic differential.

Using pairs of oppositely defined scales rather than single 'bipolar' scales, Bentler (1969) demonstrated this effect yet more forcefully by showing a near zero correlation between opposite defined scales where a high negative correlation was predicted. However, when the effects of 'acquiescence set' were partialled out, he found high correlations in the predicted direction. This was put forward as evidence for the assumption of bipolarity. Bentler went on to discuss the validity of partialling out response bias, and suggested that combinations of unipolar scales might well be a more powerful

instrument than the conventional semantic differential.

Ivan Sipos of the Slovak Academy of Sciences has contributed much, albeit quietly, to the study of semantic content. His work has little direct bearing on semantic differentiation itself, but is particularly relevant to problems of wording in surveys generally.

Sipos & Kolada (1966) present a method for determining the 'entropy' or precision of statements or words. The method is subject to the general criticisms of categorical scaling as presented above, but conceptually it reflects Jones & Thurstone's thinking in their paper of 1955. Sipos (1966) applied this method to definite expressions and cliches, and demonstrated sex differences in the expressed meaning of several statements. The criticisms of the categorical method used in the above two papers are partly mitigated in Sipos (1967) in which he applies the method of successive intervals to data gathered on semantic entropy.

Sipos & Adamica (1967) applied Cohen & Hansel's (1956) concept of subjective uncertainty to items selected from Eysenck's MMQ test, and exhibited widely varying within-group measures of meaning with adverbs and adverbial expressions of time. This would appear to endorse the above summarised work on perturbations due to individual differences, and class-specific operational definitions.

The work on semantic differentiation complements the rather crude studies on question wording, by providing precise profiles of meaning for question material which had previously been presumed to approximate the values as understood by the test constructor. The danger of ignoring class-specific meanings has been more than amply demonstrated.

Conclusion and rationale behind the present series of experiments.

Thurstone's LCJ would appear to have been rather too severe in its assumptions. If Mosteller's (1951 a, b & c) arguments were extended to series in which all items were assumed to have different discriminal dispersions, then the complete lack of all but the approximate ordinal

scaling of items would necessarily ensue, and there is little to suggest that this is not the case in the overwhelming majority of cases. Of course Thurstone's approach has been of great value where it has been used, I merely suggest the assumptions have been unfortunately, albeit necessarily, strict.

A rather more interesting approach appears to be offered by Durbin (1951), Schucker (1959) and Slater (1965) in which sets of items are ranked rather than pairs discriminated. There are obvious apparent advantages in time saved alone, but the validity of the Law of Comparative Judgement might be questioned where three or more items are presented. Disregarding for the moment the Law of Comparative Judgement, the present study will attempt to investigate some aspects of ranking over paired comparisons.

An attempt will also be made here to examine Upshaw's (1962) work using tangible stimuli in a conventional category scaling situation. Thus, while disregarding manifestations of social perceptual set, which can only be studied against the background of an established theory of attitude, an attempt will be made to discover certain purely psychophysical response set effects which might be assumed to underly all category scaling procedures. The data will also serve to show relationships between the various methods of item scaling.

Finally the data will also be seen to be applicable to semantic differentiation, and the use of paired comparisons as an alternative in semantic differentiation will be discussed.

Chapter 5

Experiments Carried Out

EXPERIMENT A

Experiment to examine systematic directional biases in category sorting, and to investigate item variance over categories as an artefact of end effect.

Apparatus

The experiment was designed to investigate underlying response sets and as such the stimuli used were designed to be manifestly unidimensional and easily quantifiable. The experimental items consisted of 72 black discs drawn on 4" square cards. The discs varied in size from 4 to 7cm diameter. Every card was identified to the experimenter only by a coding system. The size of each disc could not be precisely controlled due to the small differences involved and the inaccuracies of the drawing instrument, which had an error of approximately 0.3mm (95% confidence limits). However, it was intended that the mean difference between the radii of adjacently size coded cards should be 0.21mm and that this should vary between 0.042mm and 0.38mm. Instrument error meant that this range would be somewhat larger and would probably result in some 12 item pairs becoming reversed in actual size order. This represents a rank correlation between actual and intended disc size in the order of 0.9975.

These items were to be sorted by judges into eleven categories represented by eleven boxes in a single unit some 4' long. The boxes were so designed that there was no indication of the number of cards

already inserted into any box, and subjects could not change a decision once made.

Exact copies of the largest and smallest discs were placed in front of the extreme category boxes as reference points. This differed from the Thurstone technique in not defining the central category as 'neutral', thus removing the criticism that the response continuum represents two continua extending outwards from the central category (McNemar 1946 and Mordkoff 1963). The Likert method and the semantic differential define every point on the response continuum and the present method also diverges from this. The implicit assumption in defining every response category is that these definitions represent equal sized, equally spaced divisions and that they cover every possible degree of magnitude of response. The writer does not assume this.

Method

Forty-two subjects (judges) were used; all were student volunteers at Durham University. Half sorted the cards with the larger comparison card on their left and half with it on their right.

For each subject the cards were thoroughly shuffled and placed in a single pile in a presentation box which hid them from view but enabled the subject to remove the top card quite comfortably. The subject was then told the nature of the task thus:-

"In this box (indicate) there are some cards like these ones (indicate reference cards in front of each extreme category box). They all have black circles on them. The circles are all different sizes but the biggest (or smallest) is the same as this one (indicate appropriate reference card) and the smallest (or biggest) is the same as this one (indicate other reference card). All the other circles are in between these two! Now I want you to pick the cards out of the box one by one. You just have to feel behind and pick off the top one. (Demonstrate with blank cards till competent.) Then I want you to

compare the size of the circle with the ones you can see here (indicate reference cards). If it looks the same size as this one (indicate either of the reference cards) you put it in this box (indicate appropriate extreme box). If it looks half way between them you put it in here (indicate approximately the centre boxes) and so on. Do you understand? (pause for questions).

"All the cards are different sizes but some of the differences are very small. But don't worry if after you've done it you think you put a card in the wrong box, because you most probably didn't. People are usually a lot more accurate than they think! Once you've put a card in forget it, and only judge the card in your hand with the ones you see here (indicate). Remember, we're not probing into your mind or anything like that so just take it easy and take as long as you like. Do you understand?"

Any queries were answered with paraphrases of the above instructions.

Once they indicated they understood the task the subjects were allowed to handle the experimental cards. While performing the task they were all given some random positive reinforcement in the form of occasional favourable comments on their 'accuracy'.

If a subject showed gross errors in his first few responses he was reinstructed and questioned. Any cards already sorted were replaced randomly in the pack and the subject allowed to continue. In only two cases was this thought necessary. If a subject persisted in gross sorting abberations he was allowed to finish the task as above and his results were surrepticiously discarded. Only four sets of results were rejected.

Results and discussion

The raw results may be seen in Table 1 (see appendix).

Hypothesising a position effect, one of two effects might be observed:-

1) if the subjects as a whole tend to sort towards either the biggest or the smallest comparison card then the responses in both the groups will tend to show the same distribution over categories from biggest to smallest.

2) if the subjects tend to sort towards a preferred hand then the two groups will tend to show the same distribution of responses over categories from left to right irrespective of size.

To detect any possible bias towards a preferred hand the distribution of responses over categories from biggest to smallest for each subject in both groups was examined by partitioning chi square (Graph 1). This produced:-

Source	Chi Square	df	P
Group sorting with largest card on the left	337.85	200	<0.0001(z=6)
Group sorting with largest card on the right	270.28	200	0.0006 (z=3.3)
Total	611.81	410	<0.0001 (z=6)
Residual	3.69	10	0.96

It was therefore concluded that there was no difference between the two groups in their sorting of the items over categories from biggest to smallest, i.e., there was no handedness bias.

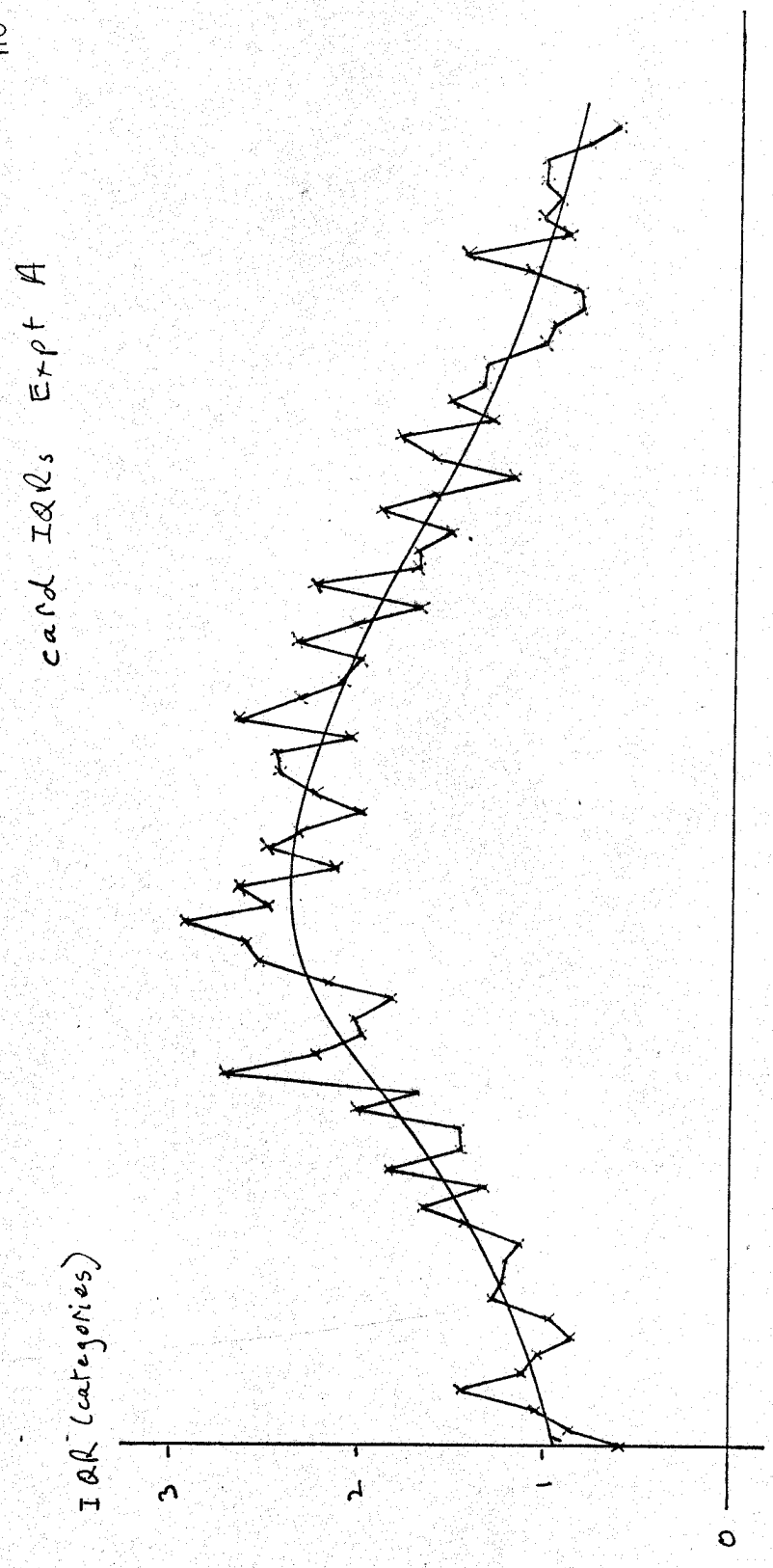
To detect any possible size bias the distribution of responses over categories from left to right for each subject in both groups was examined:-

Graph 1

Distribution of items over
categories for subjects sorting
with the largest card on the right
compared with that for those sorting
with the largest card on the left
in experiment A.

Also the overall mean distribution.

Graph 2

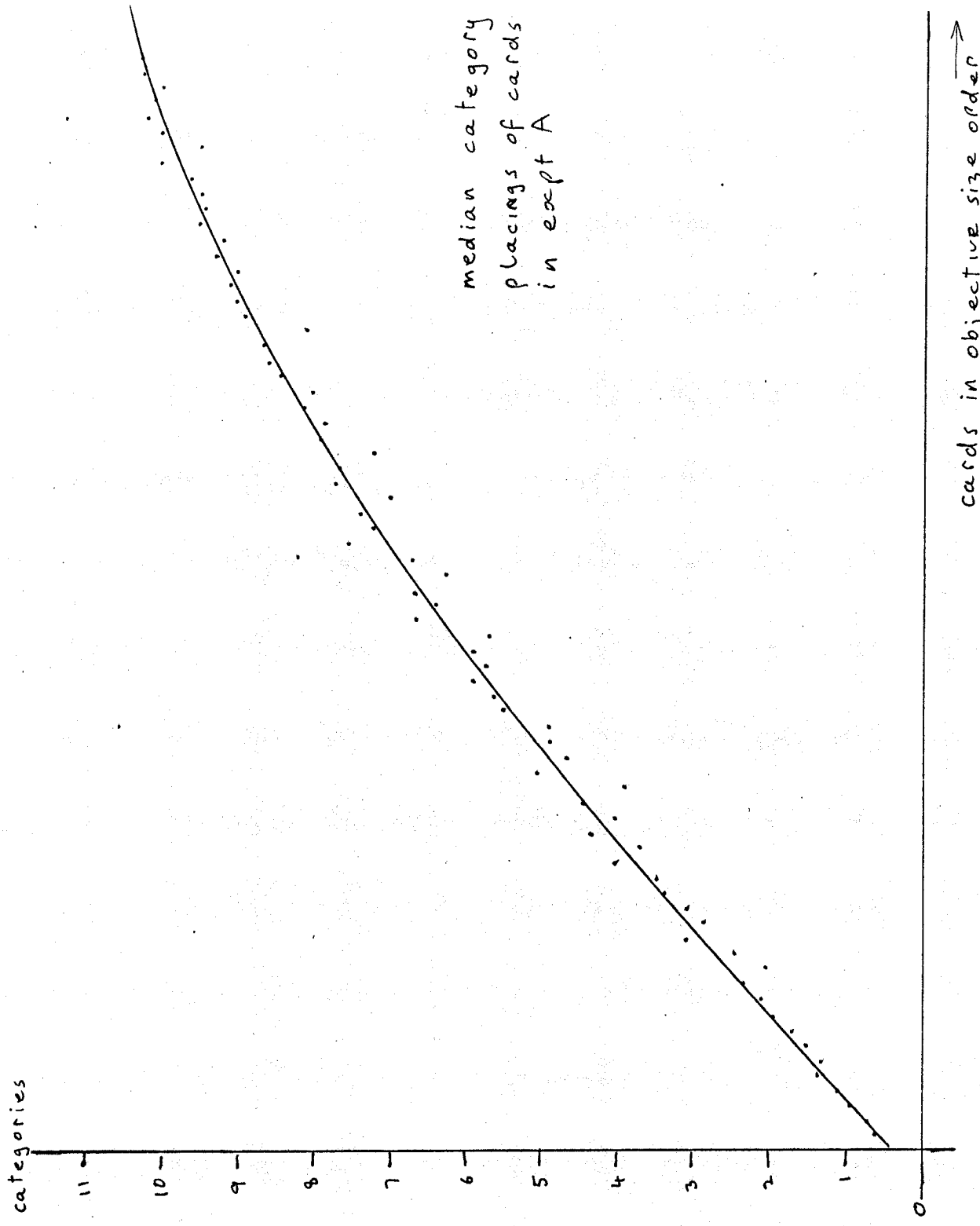


Card interquartile ranges, Expt A

Graph 3

Distribution of Likert's item selection statistic 't'.

T-test for the significance of the difference between the mean item category placings of the items in experiment A, taking the highest and lowest 'scoring' quartile subjects. Subjects'scores based on Likert's integral weighting of item category placings from 0 to 10.



Graph 4

Item median category placings, Expt A.

Source	Chi Square	df	P
Group sorting with largest card on the left	337.85	200	<0.0001
Group sorting with largest card on the right	270.28	200	0.0006
Total	725.26	410	$\ll 0.1^{8}$ (z=9.5)
Residual	117.13	10	$\lll 0.001$

It was therefore concluded that there was a considerable size bias.

Handedness bias was therefore ignored and the two groups combined for further analysis, ~~against intended order of size of the items.~~

From graph 2, it can be seen that central items are subject to higher interquartile range than extreme ones. This has a bearing on the Thurstone and the Edwards and Kilpatrick scaling techniques, to be discussed later.

Graph 3 is a plot of the statistic 't' used as part of the item selection technique in the Likert Method. This will also be discussed later.

Discussion

This experiment was necessarily simple because of the large number of subjects needed. As a result of it, more complex experiments were carried out and are reported below. The present experiment's bearing on the questions in hand will be left to the final discussion.

EXPERIMENT B

Experiment to examine the effects on a category scale of variations in the distribution of the items presented.

Apparatus

Two packs of 4" square cards with open circles inscribed one on each card. The circles varied from 5 to 7cm diameter. Unlike Experiment A the mean difference between adjacent size-coded items was a constant

CODE	ITEM	P_1	P_2
AA	1		
AB	2		
AC	3		
AD	4		
AE	5		
AF	6		
AG	7		
AH	8		
AI	9		
AJ	10		
BA	11		
BB	12		
BC	13		
BD	14		
BE	15		
BF	16		
BG	17		
BH	18		
BI	19		
BJ	20		
CA	21		
CB	22		
CC	23		
CD	24		
CE	25		
CF	26		
CG	27		
CH	28		
CI	29		
CJ	30		
DA	31		
DB	32		
DC	33		
DD	34		
DE	35		
DF	36		
DG	37		
DH	38		
DI	39		
DJ	40		
EA	41		
EB	42		
EC	43		
ED	44		
EE	45		
EF	46		
EG	47		
EH	48		
EI	49		
EJ	50		
FA	51		
FB	52		
FC	53		
FD	54		
FE	55		
FF	56		
FG	57		
FH	58		
FI	59		
FJ	60		
GA	61		
GB	62		
GC	63		
GD	64		
GE	65		
GF	66		
GG	67		
GH	68		
GI	69		
GJ	70		
HA	71		
HB	72		
HC	73		
HD	74		

TABLE 2 Relative distributions of packs P_1 and P_2 in experiment B.

proportional difference of approximately 0.457% rather than a constant size difference. This difference varied between 0.09% and 0.823%. The reliability of the final order of items could not be estimated as in the previous experiment, but it was probably of the same order.

The two packs consisted of:-

P_I 74 cards with circles representing the complete range as constructed.

P_2 45 cards distributed as in table 2. Thus the smallest and the central items of P_I were eliminated from P_2. The apparatus was otherwise as in experiment A.

Method

Sixteen subjects were used. Both packs were sorted by every subject. The packs were not presented in immediate succession, but two additional tasks were interpolated (see experiment C), this causing there to be an interval of approximately thirty minutes between the administrations of P_I and P_2.

Table 3

Division of subjects into groups

Group for expt B	Subjects sorting largest card on left (L) or on right (r) in expt B	Order of Presentation 1	2	3	4	No. of subjects	Group for expt C
A	L	P_I	R	P-C	P_2	2	X
A	R	P_I	R	P-C	P_2	2	X
A	L	P_I	P-C	R	P_2	2	Y
A	R	P_I	P-C	R	P_2	2	Y
B	L	P_2	R	P-C	P_I	2	X
B	R	P_2	R	P-C	P_I	2	X
B	L	P_2	P-C	R	P_I	2	Y
B	R	P_2	P-C	R	P_I	2	Y

The first and last tasks to be presented are those concerning us in this experiment. The two interpolated tasks are explained in experiment C.

The subjects were divided into two groups as in Table 3. Group A sorted P_I first and group B sorted P_2 first. Half of each group sorted the packs with the largest comparison card on their left and half with it on their right. For analysis the direction of sorting was ignored since experiment A showed that this had no effect upon the overall distribution of items over categories.

Unlike experiment A the comparison cards shown as defining the extreme categories were not exact replicas of the largest and smallest experimental circles, but were scaled up to half as big again with the proportions of circle diameter to card size the same.

Apart from emphasising the nature of the scaled up comparison cards the instructions and method were as in experiment A.

Results and discussion

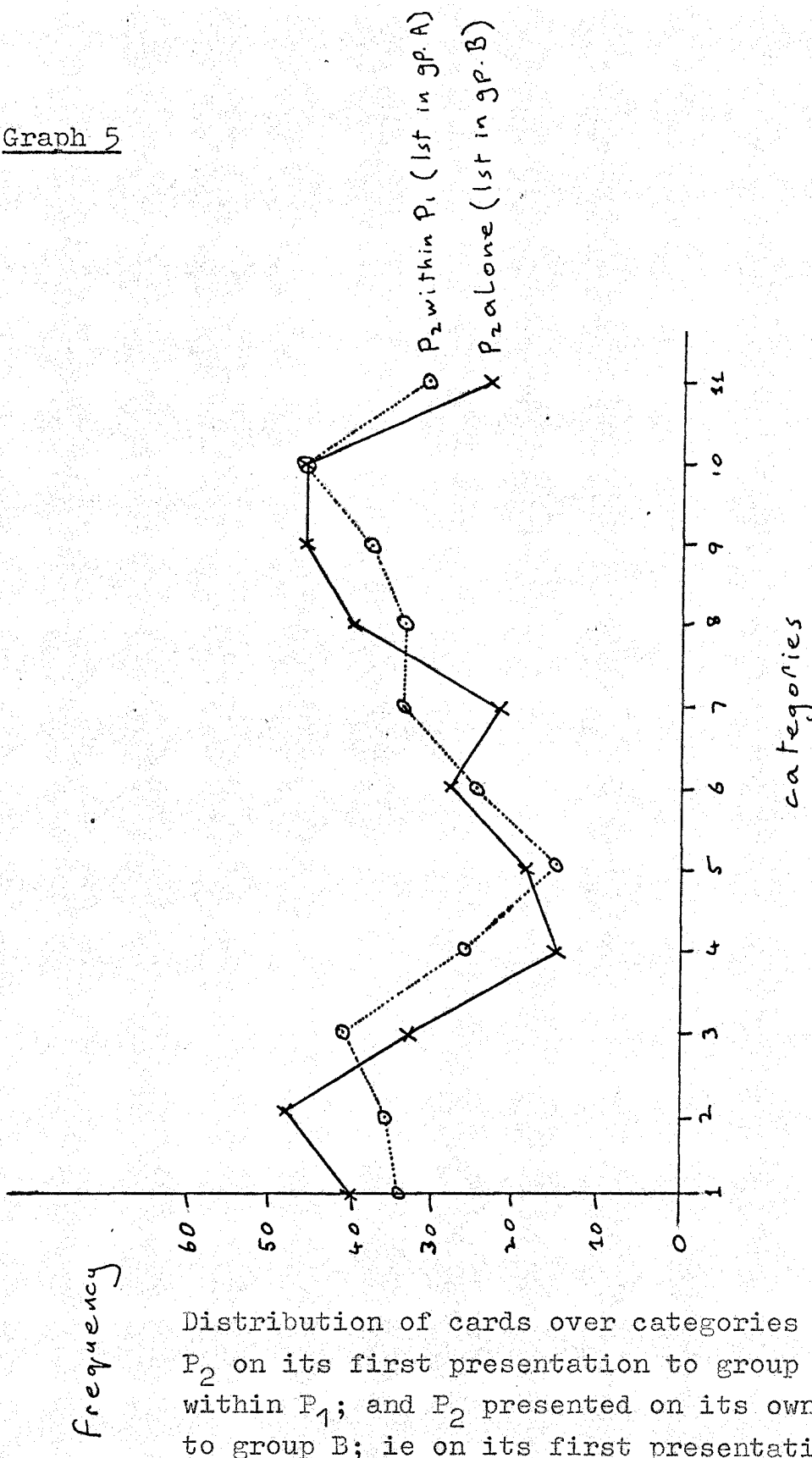
The distribution of cards over categories was examined in AP_1 to detect any 'handedness' bias.

Source	Chi-square	df	P
Subjects sorting largest cards on the left	63.6	24	0.001
Subjects sorting largest cards on the right	321.73	24	0.001
Total	388.72	56	0.001
Residual	3.39	8	0.9

No significant difference was found, and the left and right hand sorting groups were combined for further analysis.

Consider P_I as a combined pack of P_2 and 29 others, the others causing the whole pack to assume a random distribution between biggest and smallest. Ideally the distribution of P_2 over categories should be

Graph 5



Distribution of cards over categories for P_2 on its first presentation to group A, within P_1; and P_2 presented on its own to group B; ie on its first presentation to both groups.

substantially the same when presented on its own and within P_I. To test this the distribution of cards over categories was examined by partitioning chi square.

Firstly the distribution of P_2 presented on its own to Group B was compared to its distribution as part of P_I presented to Group A. That is on its first presentation to both groups. (Graph 5).

Source	Chi-square	Df	P
Gp A	110.68	42	$<< 0.001$
Gp B	142.89	42	$<< 0.001$
Total	274.28	90	$< 0.1^8$
Residual	20.71	6	0.0028

This analysis suggests that the form of the distribution of the items presented has an effect on the overall distribution of items over categories.

Next the distribution of P_I over categories in both groups was compared:- (Graph 6)

Source	Chi-square	Df	P
Gp A	172.13	70	$<< 0.001$
Gp B	165.72	70	$<< 0.001$
Total	356.35	150	0.1^9
Residual	18.5	10	0.05

and the same was done for P_2:- (Graph 7) 6a

Source	Chi-square	Df	P
Gp A	84.64	35	< 0.001
Gp B	80.58	35	< 0.001
Total	183.37	75	0.1^9
Residual	18.5	5	0.0049

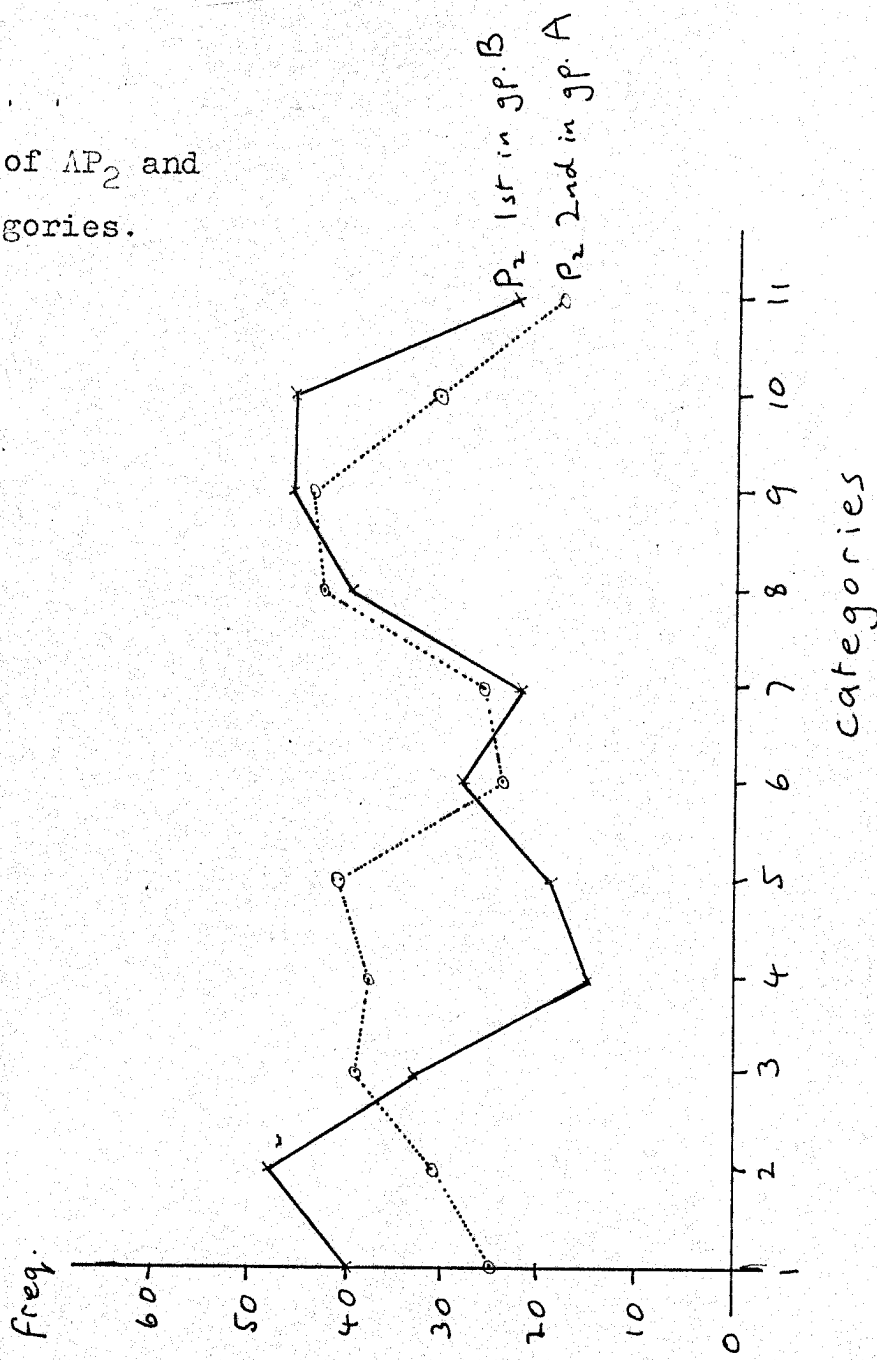
These two analyses examined the effect on sorting one pack of having already sorted the other. That is it examined the effects of 'carryover' from one pack to the next.

Graph 6

Distribution of AP_1 and BP_1 over categories.

Graph 6a

Distribution of AP_2 and BP_2 over categories.



In order to draw any conclusions from these analysis it was necessary to examine the distribution of the residual probabilities to decide whether they represented a random sample from the total population of possible residual probabilities. This was done by summing the residual chi-squares and df.s, producing chi-square = 57.36 df = 21 and $p < 0.001$. The residual chi squares were therefore taken as indicating truly significantly different factors.

(It will be noted and may be queried that in the above analysis chi-squares from the same source were different with different df.s. This was due to the combinations of categories required in order not to have too many low expected frequencies). It can be seen that in all the cases examined there was a significant difference in the distributions compared. However, the chi-square test is insensitive to direction of difference. The median category placings of the actual test circles were therefore examined empirically to determine the direction of any shift in median category placing. This analysis assumes that any shift in the distribution of cards over categories will be the result of a systematic shift in the median category placings of the cards themselves.

Table 4

Group	1st presentation	2nd presentation
A	P_I	P_2
B	P_2	P_I

(In the analysis to follow the notation will indicate the group and the pack considered, without reference to whether that pack was the first or the second presented to a group. AP_I will represent the median category placings of the circles constituting P_I when P_I was presented to group A &c. When both P_I & P_2 are expressed together it will indicate that only the circles common to both packs are treated. Reference to Table 4 will give the missing information.)

AP_I is taken as the 'ideal' set of median category values for the circles. This is basically because P_I was so constructed as to represent as fully as possible the total range of circles between the visible comparison cards. Also AP_I may be taken as being 'uncontaminated' by any outside influences not also common to all the other distributions. AP_I will be represented by the symbol 'I'.

BP_2 is taken as a contaminated estimate of P_2, the uncontaminated estimate being the distribution of P_2 within AP_I. The contamination of BP_2 is taken as being due to the asymetrical distribution of P_2 relative to P_I. BP_2 will be represented by 'I+e', where 'e' represents the contaminating influence ..

BP_I is taken as being equivalent to AP_I plus a 'carryover' effect from BP_2 and will be represented by 'I+c'.

AP_2 is taken as equivalent to BP_2 plus a further contaminating effect due to 'carryover' from AP_I and will be represented by 'I+e+B'.

To summarise:-

$$AP_I = I$$

$$BP_2 = I+e$$

$$BP_I = I+a$$

$$AP_2 = I+e+B$$

If these effects may be considered merely additive then it may be seen that:-

$$e = BP_2 - AP_I \qquad (I)$$

$$a = BP_I - AP_I \qquad (II)$$

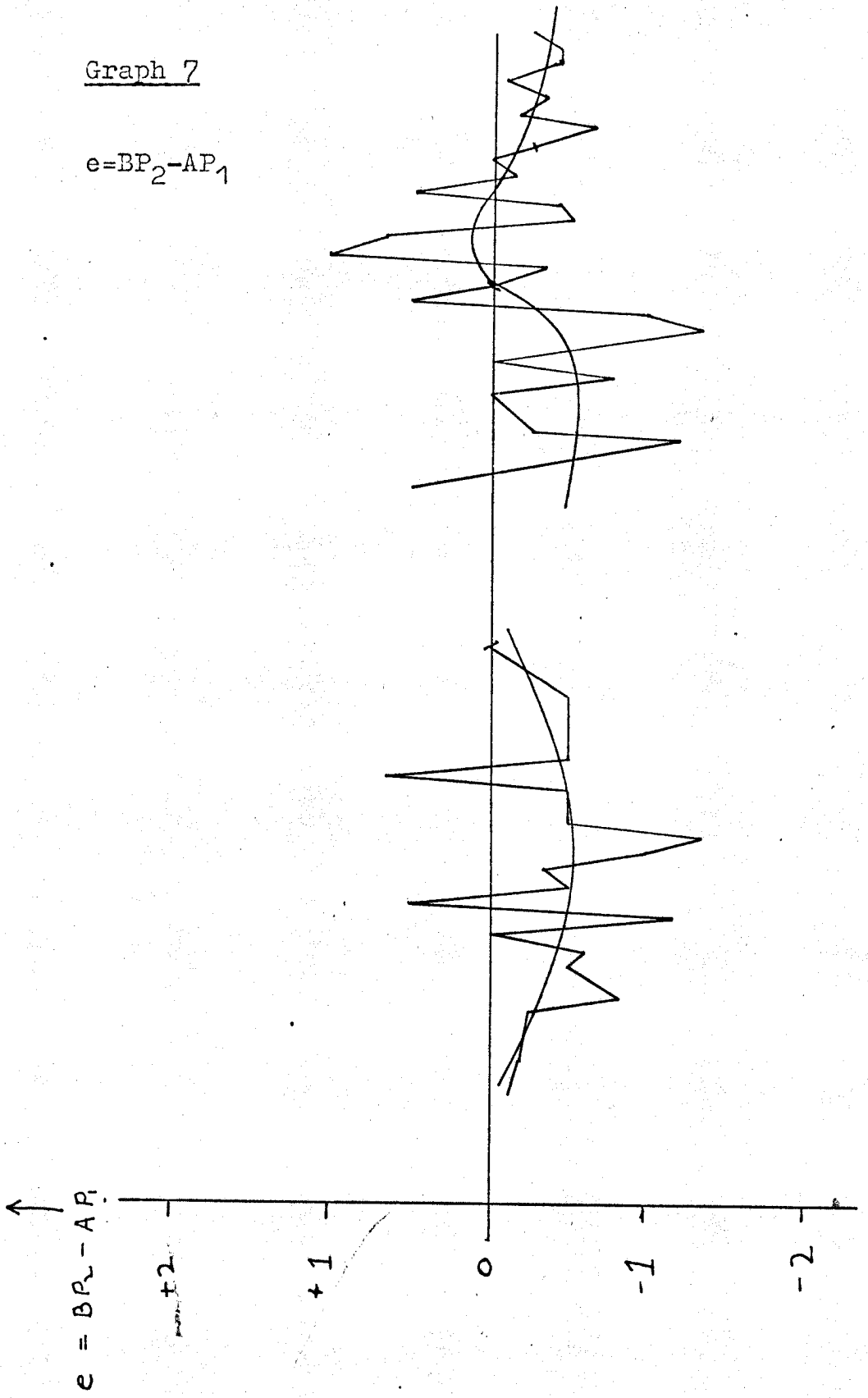
$$B = AP_2 - BP_2 \qquad (III)$$

In any case this form of analysis will demonstrate the forms of the shifts which produced the significant chi-squares above.

Consider graphs 7, 8 and 9, corresponding to e, a and B. These

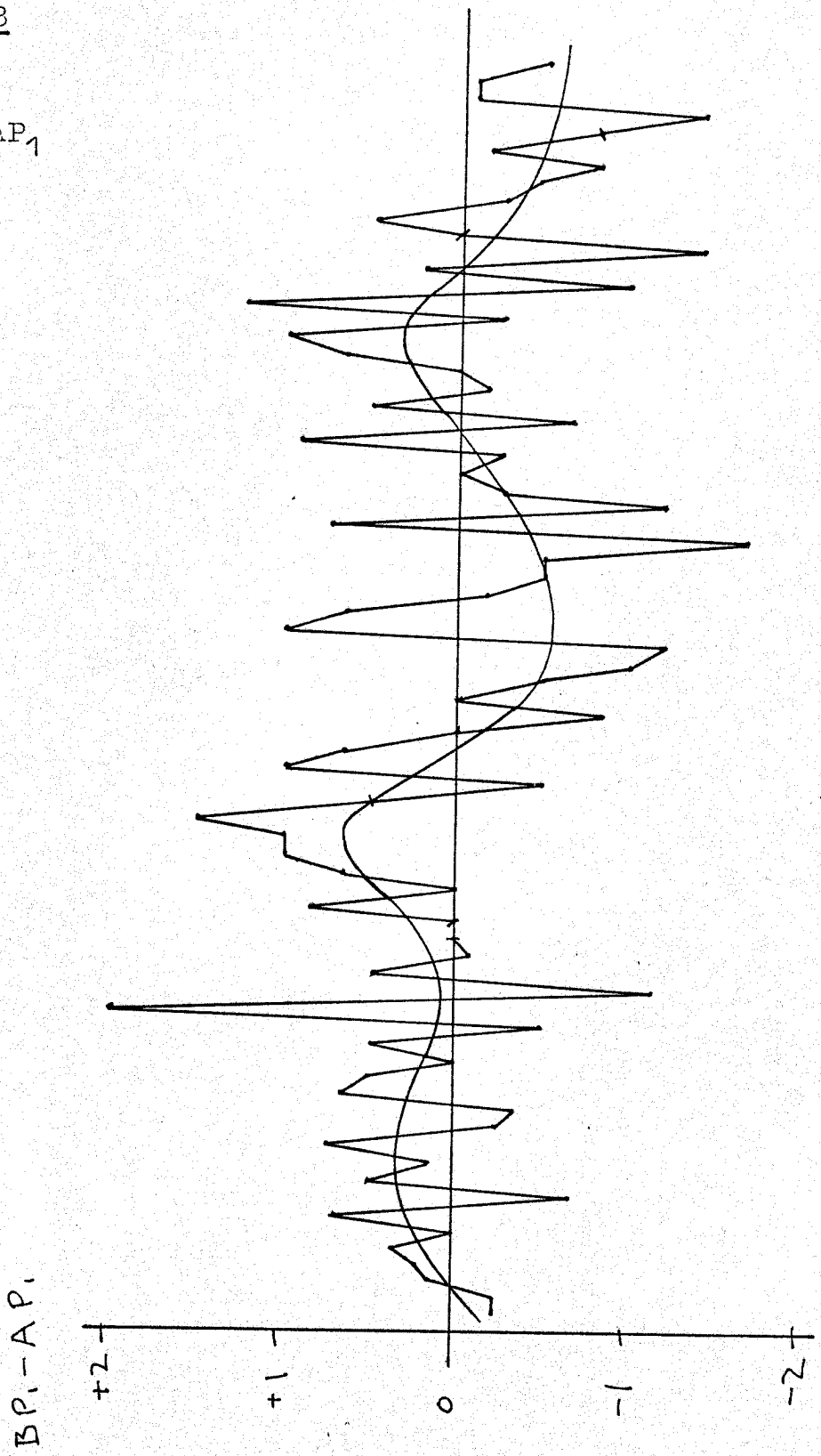
Graph 7

$e=BP_2-AP_1$

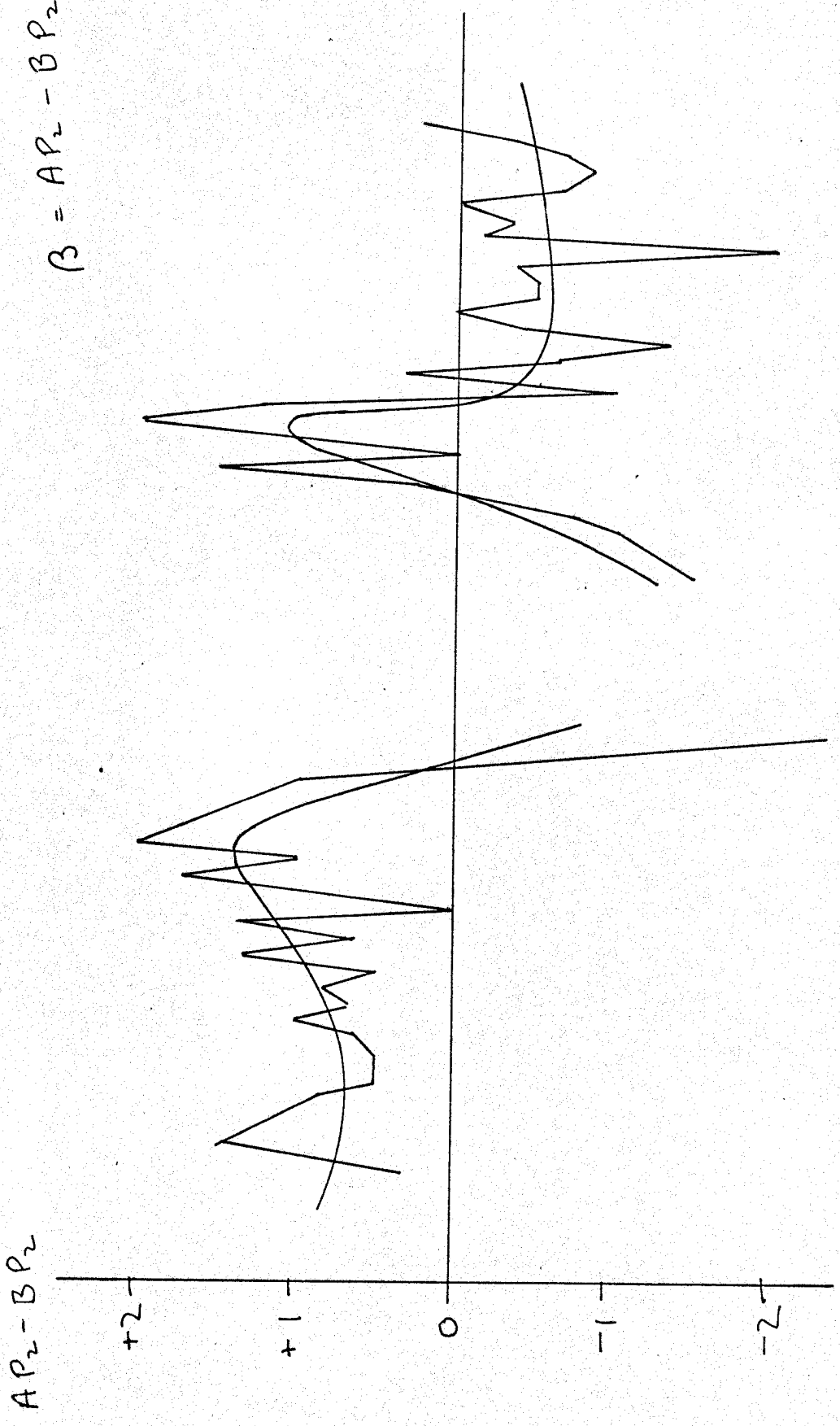


Graph 8

$\alpha = BP_1 - AP_1$



Graph 9



graphs were constructed by taking the differences in the median category placings of the items indicated. A positive difference indicates a shift towards the 'large' end of the response continuum caused by the factor indicated, and conversely a negative difference indicates a shift towards the 'small' end relative to the baseline value (AP_I).

A conjectural result was that for B in graph 9 which is taken as representing (I=e+B) - (I+e). As stated this assumes that all effects are additive. However, a case might be made for the non-existence of effect 'e' in AP_2 as this assumes essentially that the subject is unfamiliar with the general form of the items and in AP_2 this is clearly not the case. A better estimate of B might therefore be (AP_2 - AP_I) as in graph 10 which was constructed by plotting the median of the difference of the subjects' category placings of the items in P_2 and P_I. (It will be particularly noticed that the shifts evidenced in graph 9 are considerably larger than those in graphs 7, 8 and 10, suggesting that graph 9 might well be a biased estimate.)

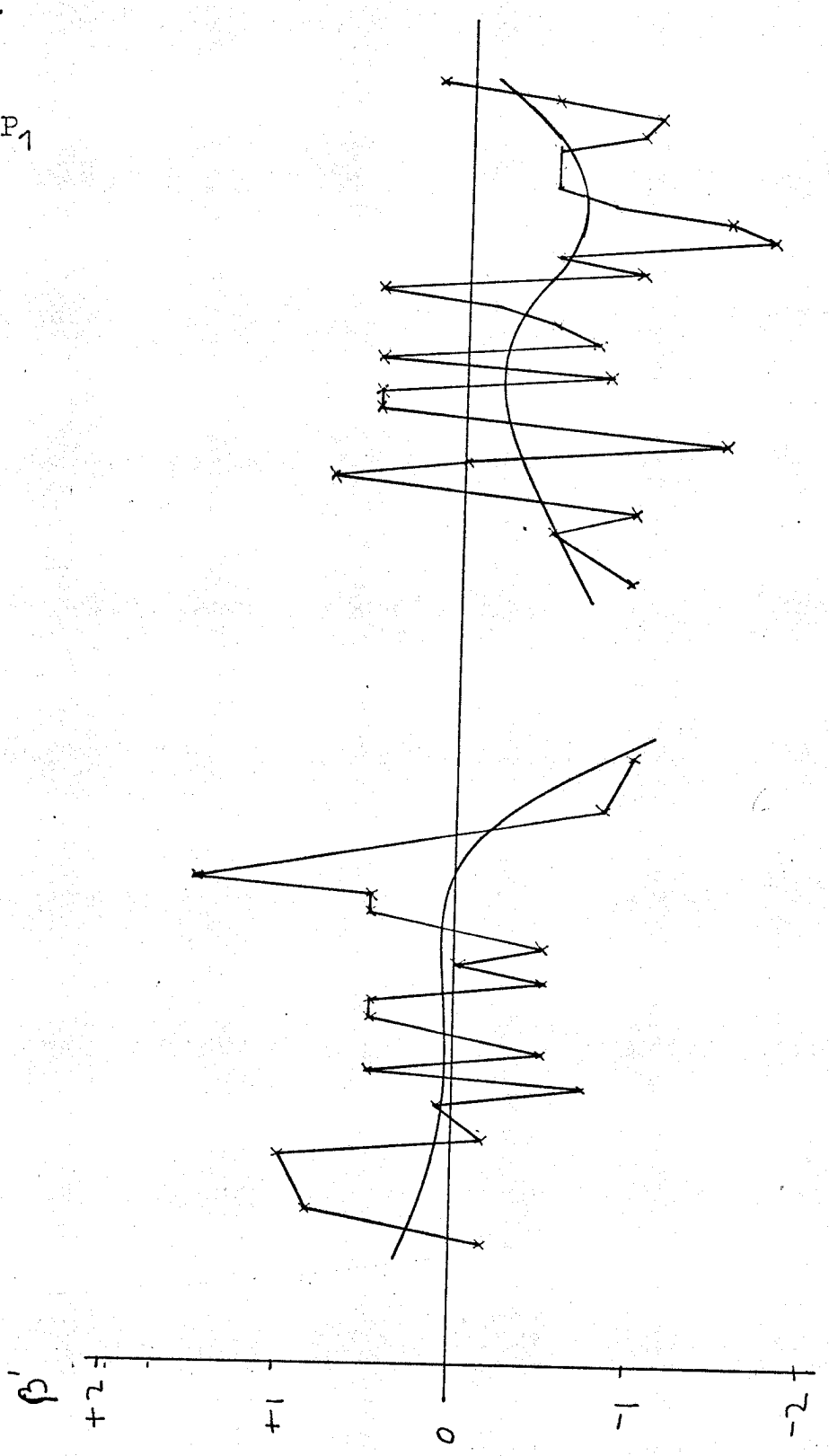
The partitioned chi-squares above do not treat the situation (AP_2 - AP_I). In order to examine the significance of the difference between the subjects' sortings of P_I and P_2 in group A the distributions of P_2 and P_2 within P_I for every subject in group A were examined by chi-square.

Table 5

Subject	Chi-square	df	P
I	8.00	6	0.26
2	17.47	5	0.0061
3	14.21	6	0.03
4	2.17	6	0.91
5	2.76	5	0.76
6	10.07	6	0.13
7	16.91	6	0.0094
8	11.91	6	0.059
Total	83.50	46	<0.001

Graph 10

$\beta' = AP_2 - AP_1$



Graph 11

$\beta' - \alpha$

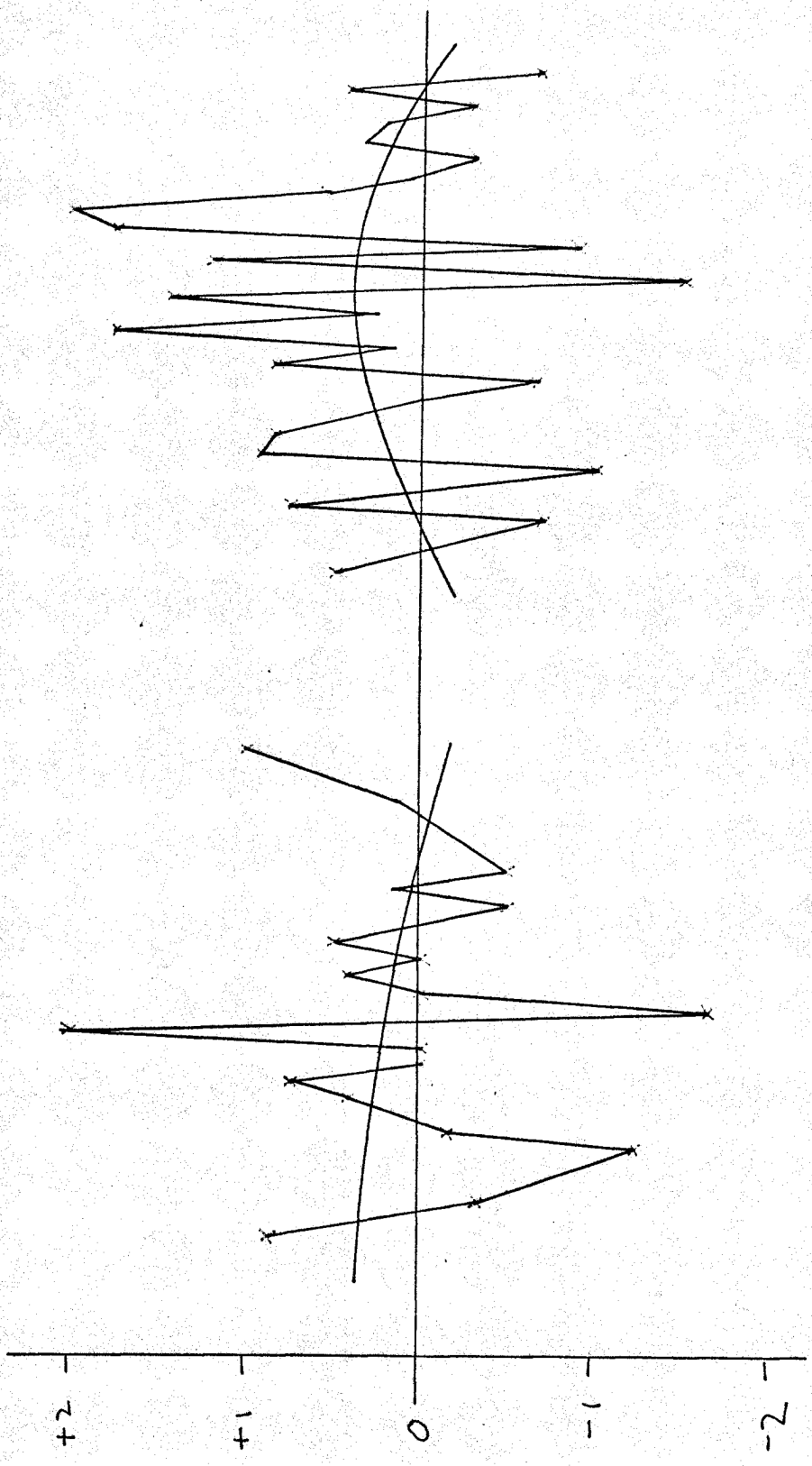


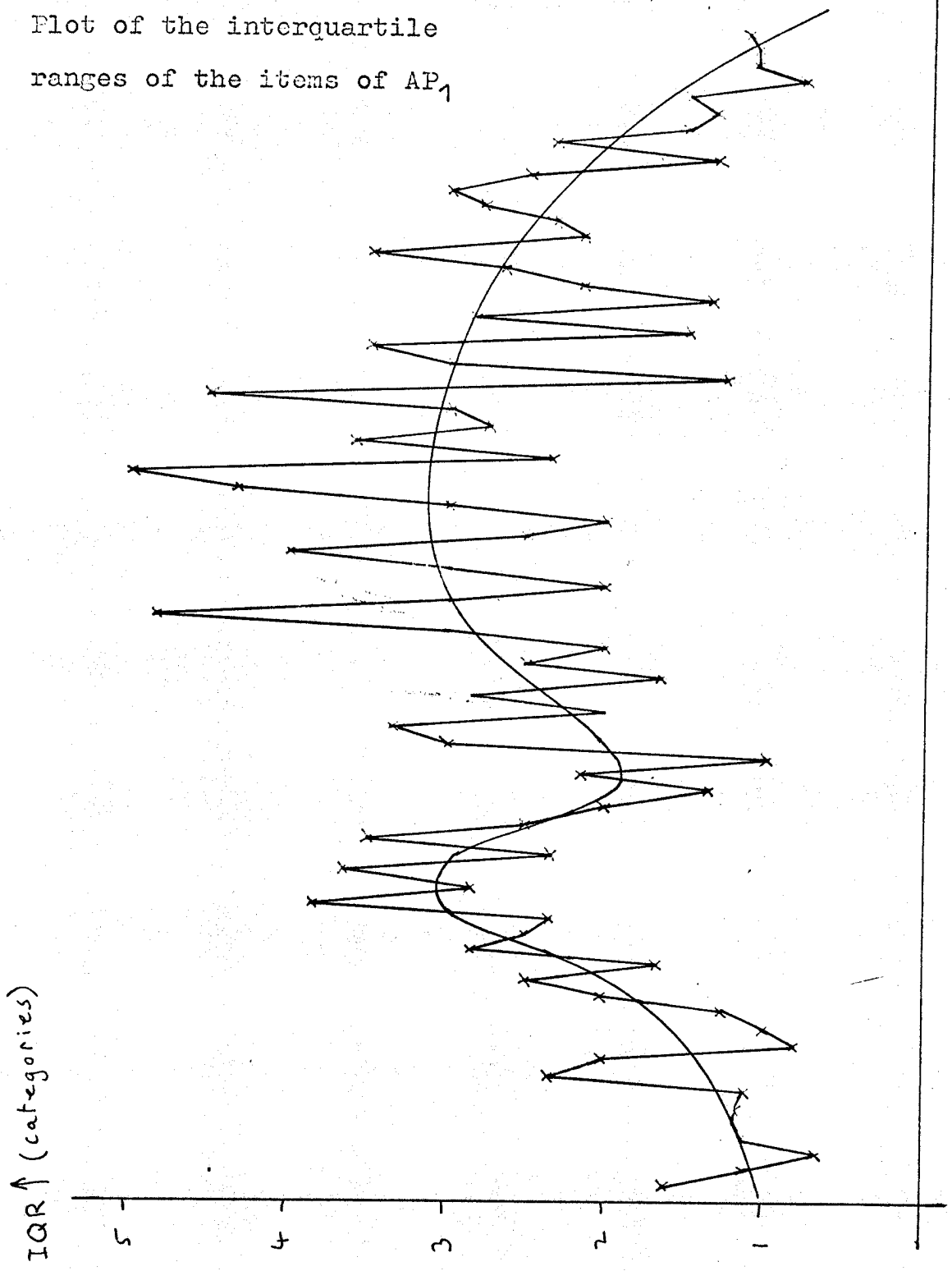
Table 5 shows the value of chi-square for each subject in group A and the total. This is taken as indicating that the general trend of the shifts exhibited in graph 10 are significant.

Examining graphs 8 and 9 a definite similarity may be observed in the general shapes and the positions of the peaks. This suggests that there was a common element in the carryover effects a and B. The strict application of the principles of Prägnanz, objective set or contrast would have predicted that the carryover effects appear mirror images of one another. We now hypothesise that the carryover effect consists of two additive effects. The common element we shall call 'C', and this will be assumed to be due to 'experience' or 'learning'. The carryover effect due specifically to having already sorted AP_I will be called 'a' and is equivalent to an uncontaminated estimate of a. Similarly the uncontaminated estimate of B' will be called 'b'. We may now say that the difference between the uncontaiminated carryover effects is (b-a) which equals (B'- a) = (C+b)-(C+a) = (b-a). Graph 11 represents B'- a. No chi-square estimate of the significance of this line can be gained, but from the above mentioned principles of Prägnanz, objective set or contrast it could be hypothesised that with the carryover effects being mirror images of one another, the difference between should be in the order of twice the individual effects. In order to test the obtained results a crude form of trend analysis was employed by comparing the variance of the data about i) the point of no difference, i.e., the line y = 0; and ii) the fitted line. Doing this the variance ratio was 1.28 with $df_I = df_2 = 44$ and thus $p = 0.2$. The data were therefore taken as indicating that no unbiased carryover effects could be detected, and it was concluded that no conclusions could be drawn concerning differential carryover effects from P_I to P_2 and from P_2 to P_1, and for further analysis a and B' were ignored.

The next analysis concerned the nature of the common element of

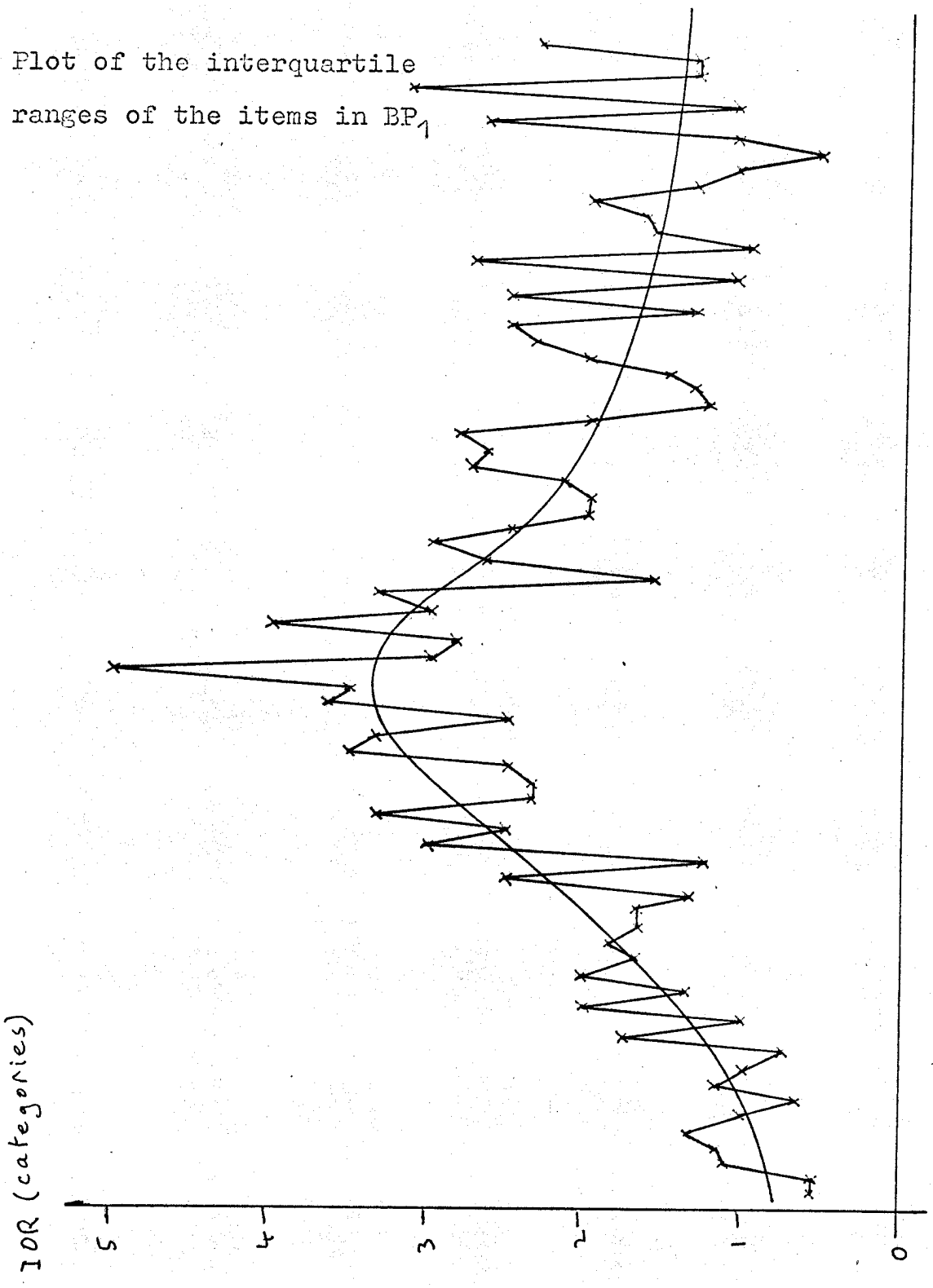
Graph 12

Plot of the interquartile ranges of the items of AP_1



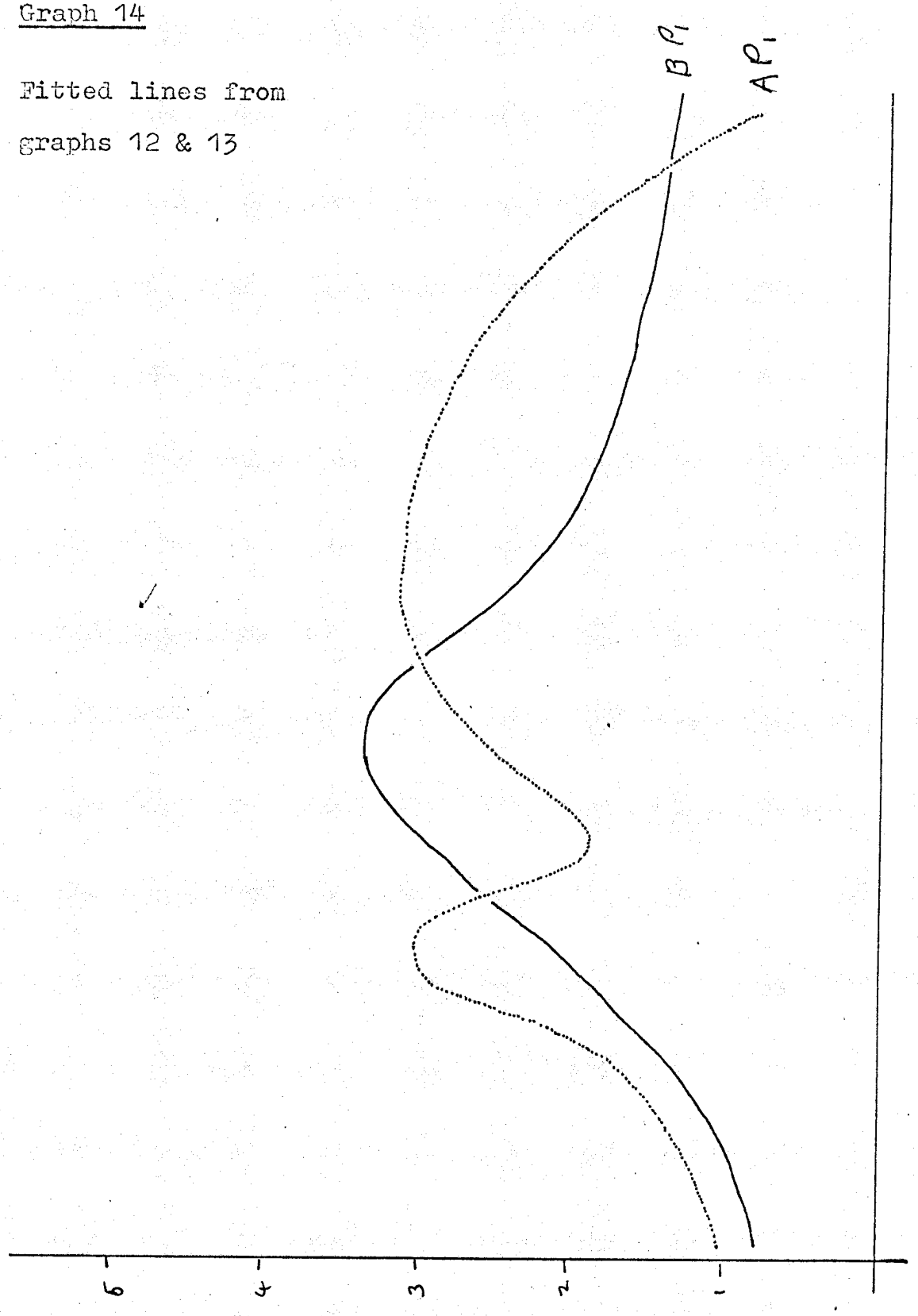
Graph 13

Plot of the interquartile ranges of the items in BP_1



Graph 14

Fitted lines from graphs 12 & 13



shift. To do this only one case was considered, namely that of AP_I and BP_I, since this provided more information than the case of P_2 consisting as it did of 74 points against the 45 of P_2.

Graphs 12 and 13 are plots of the Interquartile Ranges of AP_I and BP_I respectively. Graph 14 is a plot of the best fit lines, by eye, of the above two plots. This demonstrates the effect upon interquartile range of being already familiar with the general nature of the test items.

The writer suggests that these shifts may be understood in terms of the biasing effects of anchoring in a restricted response continuum such as this.

Examine the distribution of interquartile ranges in this experiment as compared with that in experiment A (Graph 2). Let us assume that the motivational states of the subjects were exactly the same in task AP_I and in experiment A. In both situations it was the first time the subjects had seen the circles involved. The differences in the tasks were thus:-

a) In experiment A the comparison cards were exactly related to the items sorted while in the present experiment they were only 'analogous'.

b) In experiment A the items were black circles while in the present experiment they were open circles.

c) In experiment A there was a constant mean size difference between adjacently size-coded circles, in this experiment this was a constant mean proportional difference.

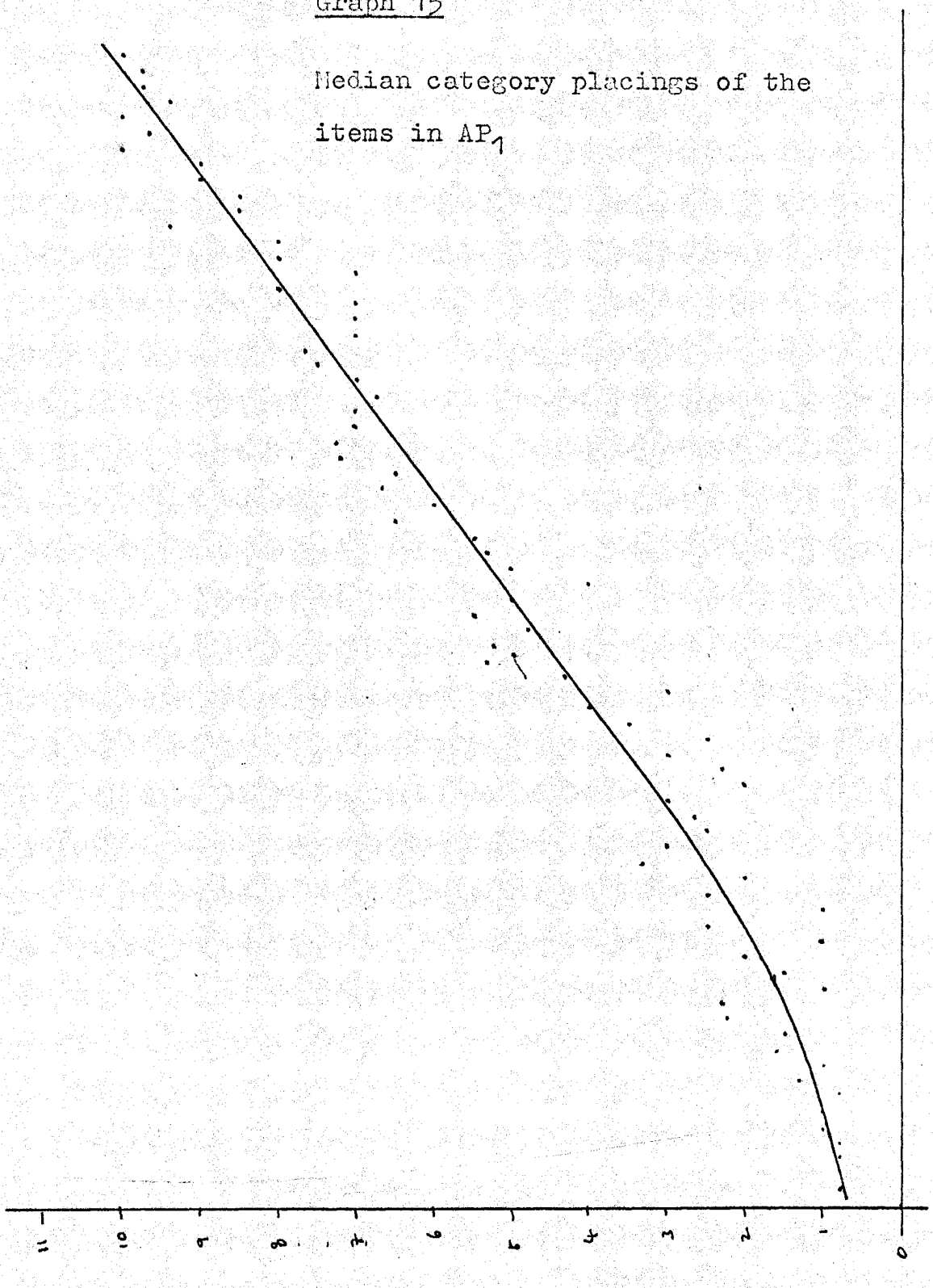
d) The ranges of circle diameters were different in the two experiments. With these in mind let us examine the effects of the differences.

Fitting a line to the item median category placings in the two experiments gives/in which the present experiment approximates far more

graph 16

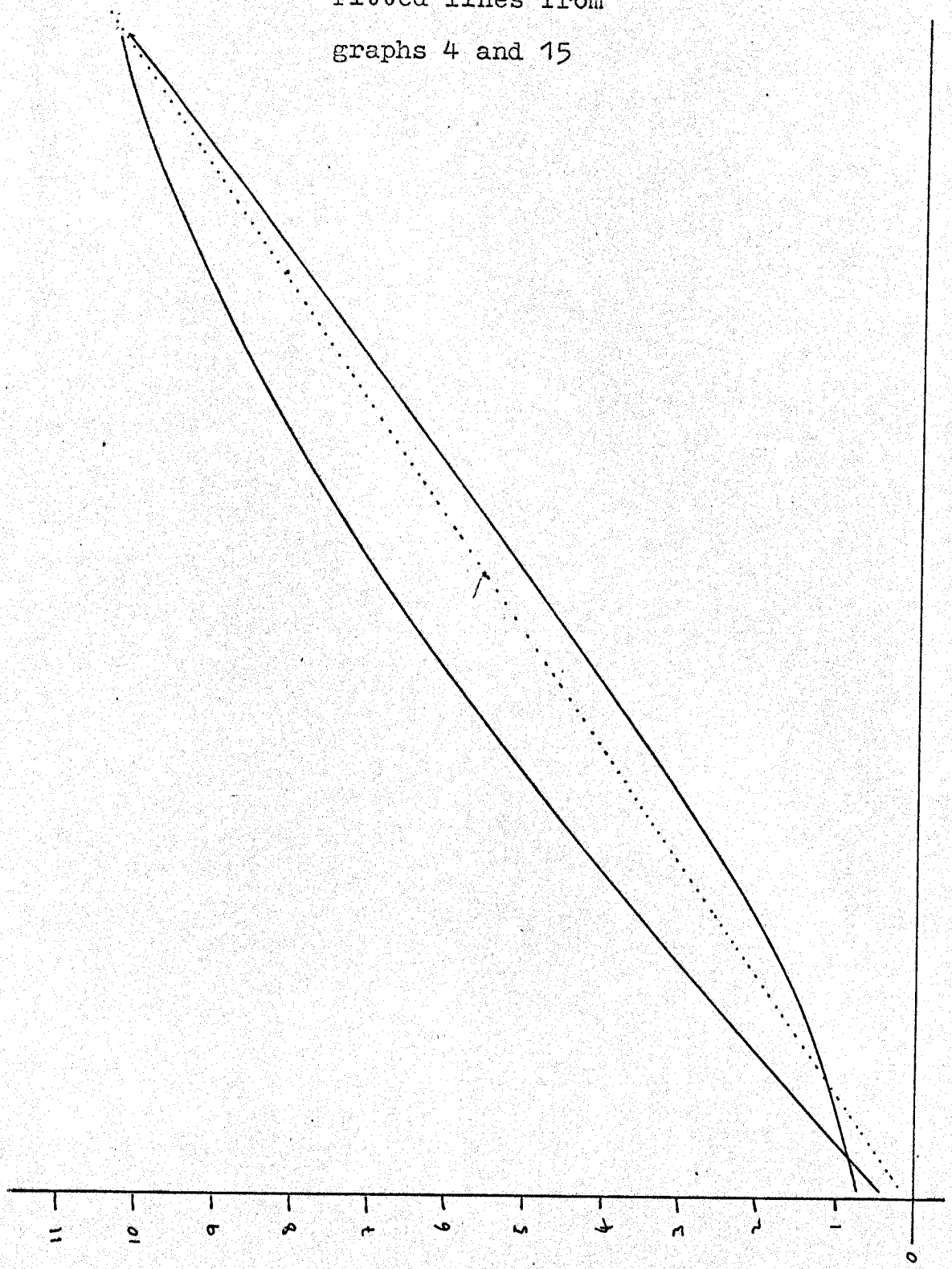
Graph 15

Median category placings of the items in AP_1



Graph 16

Fitted lines from graphs 4 and 15



closely to a straight line. This was partly the object of the present methodology which was designed after the results of experiment A had been analysed.

In motivationally analogous situations the distribution of interquartile ranges appeared radically different, while after a short period they tended towards similarity. See graph 13 and graph 2.

It thus appears that where the comparison cards were directly related to the items presented, the first trial achieved the 'optimum' distribution of items. But where the comparison cards were only analogous to the items presented the resulting distribution of interquartile ranges diverged from the 'optimum'. The 'optimum' here simply expresses the final stable values.

The writer offers the following hypotheses to describe these effects:-

I) The uncertainty with which a subject rates an object is a direct function of that object's psychological distance from an anchoring stimulus.

2) Where a subject is presented with an anchoring stimulus which either qualitatively or quantitatively does not fall near the range of the objects to be rated, the subject tends to establish his own anchor and to judge relative to that.

3) The 'potency' of a subjectively established anchor decreases as the subjects' familiarity with the relationship between the presented anchor and the objects to be rated increases.

The first hypothesis suggests that the sharp decrease of interquartile ranges near the centre of the range in graph 12 is the result of a subjective anchoring stimulus, since one characteristic of anchoring stimuli is that they tend to produce relatively invariable stereotyped responses in their vicinity (Upshaw 1962 and 1965).

The distribution of interquartile ranges in BP_I was similar to

that in experiment A apart from variations at the 'large' end, which suggests that after a period of time the sortings of the cards in the two different situations achieved a certain degree of mutual uniformity.

Graph 8 shows that about item 34 there was a tendency for more items to be massed in BP_I than in AP_I, and this coincides with the dip in interquartile ranges in AP_I. This is indicated by the negative slope of the fitted line crossing the x-axis. (As mentioned above this means that the items to the left of the crossing point will exhibit a shift to the right while the items to the right will exhibit a shift to the left, thus causing a 'relative' peak when compared with AP_I. This may be understood by comparison with the distribution of items over categories in experiment A where it may be seen that at the extremes of the response continuum, i.e., close to the anchoring stimuli, the subjects showed a reluctance to mass items. Thus in AP_I the subjects were reluctant to concentrate their responses close to their subjective anchoring stimulus.

Lastly the distribution of interquartile ranges in BP_I exhibited a tendency for relatively larger interquartile ranges at the large end than that in experiment A. This was probably due to the nature of the anchoring stimuli provided making it marginally more difficult to orientate with the anchoring stimuli than in experiment A.

EXPERIMENT C

Experiment to examine the relative discriminatory powers of paired comparisons and ranking in groups of three

Apparatus

25 items each consisting of two matched sets of cards. The sets of cards will be called the 'R' and 'P-C' sets. The R sets each consisted of three 4" square cards with an open circle inscribed on each card, and each P-C set consisted of three 4" x 8" cards with two

open circles inscribed symetrically on one side of each. The circles were exact copies of selected circles from experiment B and were identified with the same code. (see appendix)

Table 6 indicates the intended size difference of the pairs examined in proportional units (one proportional unit = 0.0914%). These size differences were chosen in an attempt to ensure as wide as possible a range of difficulty.

The sets were so constructed that each circle in R was exactly duplicated twice in P-C such that P-C consisted of the three constituent pairs of R. Schematically R may be represented by the symbols A, B, C, and P-C may be represented by AB, AC, BC.

Method

Sets R and P-C were used to examine sorting behaviour in Ranking and Paired-comparisons respectively.

To administer the R sets the subjects were told:-

"I've got some sets of cards here. I'm going to give you them a set at a time. I want you to look at the circles on the cards and to sort them into order of size. When you've sorted each set I want you to give them back to me in a pile with the biggest on the top and the smallest on the bottom and I'll give you the next to do right away. Do you understand?" Any queries were answered with paraphrases of the above instructions. The sorted orders were recorded.

To administer the P-C sets the whole pack of 75 P-C cards was placed in front of the subject face up and the subject was told:-

"In front of you are some cards with two circles on them. I want you to look at the circles carefully and decide which of them is bigger. If the circle on the right is bigger I want you to pick up the card without turning it and put it in a pile on the right. If the left hand circle is bigger I want you to put the card in a pile on the left. When you've finished you should finish up with two piles of cards with the

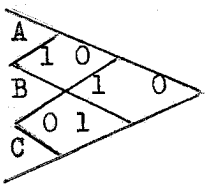
biggest circles on the outside away from you. Do you understand?" Any queries were answered with paraphrases of the above instructions and demonstrations with 'dead' cards. The object of the method was to preserve the cards as sorted in order to determine the judgement of each card and the numbers of cards moved to left and right.

Before each subject was treated the R sets were shuffled, and the P-C pile was sorted and shuffled to ensure a more or less equal number of objectively larger cards on the left and right and a different random order to each subject. The R sets were also presented in a different random order to each subject.

Half of the subjects were presented with the R sets first (group X) and half with the P-C sets first (group Y), as in Table 3. All of the R sets were presented in succession as were all the P-C pair cards; and the two groups of sets were presented with only a minute or so between them to prepare the next presentations.

Results

The results may be seen in Table 6 (in the appendix). Each section of the table is of the form:-



and consists of three cells, one for each pair of the set ABC. Within each cell are two marks. Each mark may be a 'I' or a '0'. An 'I' indicates that the subject reckoned the circle designated by the upper code defining that cell as bigger than the circle designated by the lower code defining the cell. The mark on the left of each cell refers to the subject's choice in the R set, and that on the right to his choice in the P-C set. Thus the example above represents the following discriminations:-

Ranking - subject sorted circles into order A > C > B, from which the individual pairs may be assumed to be discriminated thus; A > B, A > C, C > B.

Paired Comparisons - Subject discriminated thus; B > A, C > A, B > C.

A correct discrimination insofar as it is in the direction intended in the construction of the items is represented by a '0'.

In order to analyse these data a measure of the accuracy in each set was required. The measure taken was the number of pairs in each set correctly discriminated. Thus in the above example there were two correct discriminations in P-C against one in R.

For analysis the subjects were divided into their two groups X and Y in order to examine the effects of order of presentation of the sets, and each item was examined to discover whether either of the methods of discriminating produced more correct discriminations. Thus a 2(groups) x 8(subjects per group) x 25(items) x 2(sets per item) matrix was set up and analysis of variance done. The results were as in the following table:-

Source	Sum of squares	DF	Mean Square	F	P
Between Subjects	12.81	199			
A (groups)	0.10	24	0.0042	0.058	NSD
Subjects w. Groups (error a)	12.71	175	0.073		
Within Subjects	569.14	600			
B (items)	45.04	1	45.04	31.71	0.001
AB	15.99	24	0.67	0.47	NSD
B x subjects w. groups (error b)	248.60	175	1.42		
C (sets)	2.10	1	2.10	8.11	0.01
AC	0.03	24	0.0013	0.005	NSD
C x subjects w. groups (error c)	45.32	175	0.26		
BC	18.18	1	18.18	17.48	0.001
ABC	11.88	24	0.49	0.48	NSD
BCx subjects w. groups (error bc)	181.99	175	1.04		

This indicates a) no difference between the two groups of subjects, i.e., the order of presentation has no effect on the overall accuracy of discrimination; b) a significant difference in the accuracy with which the subjects sorted the items, i.e., some of the items were more difficult to sort or discriminate than others; c) a significant difference between the sets. This was the most important finding and indicates that the R sets produced significantly more correctly discriminated pairs (2.0875) than the P-C sets (1.985). d) there was a significant sets/items interaction. This was found to be due to a significant correlation between the number of correct discriminations in R and P-C ($r = 0.555$, $t = 3.2$, $df = 23$, $p < 0.007$).

In order to examine the sensitivity of item selection by category and paired comparison methods, firstly t-tests were carried out between the items of experiment A to decide how many 'easily distinguishable steps' could be detected in that method, i.e., to discover those items which from these data are all easily distinguishable from each other (at the 5% level, t-test for related samples).

Graph 17 represents the number of items from which each item (represented in size order on the x-axis) is not significantly different. From the fitted curve it can be seen that with some degree of reliability 22 items can be selected. This selection was done by dividing the number of items from which item X is deemed significantly different ($Y' = fX$) by two and selecting item $X + \frac{Y'}{2}$; the same procedure was then carried out for this selected item. It can be seen that on this criterion the distribution of items selected will be uneven over the defined stimulus continuum.

Next the data in the present experiment were examined, and the distribution of 'correct' and 'incorrect' discriminations on the simple paired comparisons method were plotted against the intended size order of the smaller of each item pair. By combining these results, the middle

Graph 17

The number of items in Expt A
from which each item was not
significantly different.
Data taken from multiple
t-tests for correlated means,
at the 5% level of significance.

approximately three-sevenths of the defined stimulus continuum was compared with the two extremes, and a chi-square analysis carried out on the numbers of 'incorrect' responses in these ranges. This produced chi-square = 0.853, df = 4, p = 0.9. Thus the data exhibited no end effect; a result which is by no means surprising considering the method by which the items were constructed.

The numbers of 'incorrect' discriminations were then plotted against the intended arbitrary proportional unit size differences of the pairs (see table 6). A correlation of r = -0.44, p = 0.001, was found. The regression of number of incorrect discriminations upon proportional unit size difference was found to be $Y' = 8.274 - 0.239X$, from which for $Y' = 4$ (p = 0.038 binomial test), $X = 17.88$. Now the total span of pack P_1 from which the item pairs were duplicated, was 361 proportional units, from which it may be concluded approximately 20 items may be selected on the stated criterion.

Considering the items selected from the category scaling application were from the pooled results of 42 subjects, each of whom had to sort some 72 items, the present paired comparison method represents a considerable saving in effort, representing as it does 75 simple discriminations by only 16 subjects. It is only fair to state however, that the data in the present experiment could not have been so easily gained without an accurate knowledge of the order of size of the items.

Had the triplet ranking data been used instead of the simple paired comparisons, or had a larger subject population been used and selection made at the 5% level as in experiment A, then the paired comparison method would have enabled the selection of a considerably larger number of items.

(The t-test item selection technique carried out above was not done on the data for experiment B, since for that experiment only the results for AP_1 could have been used, and thus only eight sets of results used, with the consequent loss of precision.)

EXPERIMENT D

Experiment to determine the relative temporal advantages of several methods of multiple comparison.

Apparatus

An electronic timer connected to a 'presentation shelf'. A stop clock. 200 4" square cards with circles of varying diameters inscribed one on each card. The cards were a combined pack from experiments C and B, and were more or less randomly distributed between the largest and the smallest.

Method

The cards were thoroughly shuffled and randomly divided thus:-

2 sets of 4 packs of 5 cards

2 " " 4 " " 9 "

1 set of 4 packs of 3 cards

1 " " 4 " " 7 "

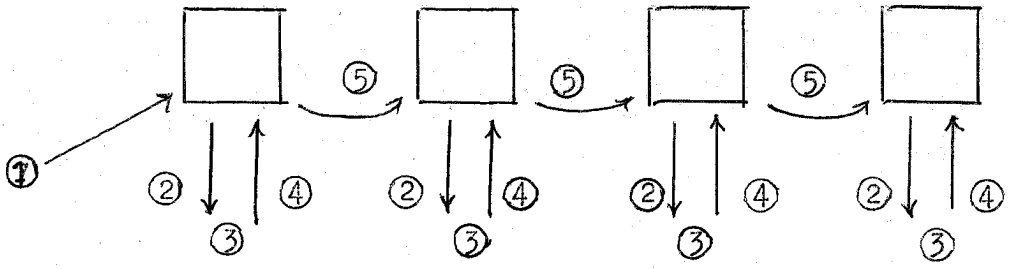
1 " " 4 " " 12 "

One set of packs of 5 cards and one of packs of 9 were used to accustom the subject to the method of sorting required. Half the subjects were presented with the 5 card packs first and half with the 9 card packs. The remaining 5 sets were presented in a different random order to each subject. All seven sets were presented in quick succession with a few minutes rest between each during which the experimenter noted results and set up the next set.

To administer each set the four packs of the set were placed on the presentation shelf side by side and the subject was instructed to sort each pack in order starting from the left. He was told to take the first pack, place the top card on the table before him, take the next card and place it to the left or right of the first as he judged the circle on it bigger or smaller than that on the first. He was to proceed with each succeeding card in a similar fashion placing them

among those already down to finish with a line of cards with circles of increasing size. At any time he could alter the order and when satisfied he was to gather up the cards with the largest on top, replace it on the shelf whence it came and move immediately to the next pack.

The sequence of actions carried out by the subject was as follows:-



1. Subject reaches out to select first pack.
2. Subject takes pack and brings it to sorting area.
3. Subject sorts pack.
4. Subject gathers up pack and replaces it.
5. Subject moves to next pack.

The presentation shelf was fitted with a microswitch beneath each pack. The first and last switches were used to record the time from the replacing of the first to the replacing of the last packs. Thus three complete cycles of events 5-2-3-4 were recorded automatically (T). Meanwhile the experimenter used a stop clock to record the total time for all four events 4 (t).

Also recorded for each sorted pack was a measure of the correlation between the true and the sorted orders in the form of $\leq d^2$ from Spearman's coefficient of rank correlation.

For each set of sortings the subject stood facing a bench 2' wide, 6' long and 3' high. On the opposite side of the bench was the presentation shelf. The area between him and the presentation shelf was the sorting area.

For subjects A and B, t was ignored due to faulty apparatus. For subjects C-F, t was corrected by a factor of 1.0123, and for subjects G-J by a factor 0.9958, due to variations in the timing clocks used.

Results

In Table 7 are recorded the raw results (see appendix).

Firstly, to see if there was any significant difference in the mean time to sort a pair between the methods of sorting, the raw results were converted thus:-

$$\text{Mean time per pair} = \frac{\frac{T}{3} - \frac{t}{4}\ (\text{x correction factor})}{\frac{N^2 - N}{2}}$$

$T/3$ = the mean time spent on each of the last three packs

$t/4$ = the mean time to pick up and replace a pack

$(N^2 - N)/2$ = the number of pairs that may be made from the N cards in each pack.

This gave the following results:-

Table 8

Subject	Sorting mode (pack size) N.				
	3	5	7	9	12
A	4.1729	2.7374	2.856	1.8039	2.1463
B	1.2597	1.088	0.7934	0.559	0.6015
C	2.0212	1.2829	1.0134	1.0742	0.8435
D	1.5876	0.9367	0.682	0.6561	0.8148
E	5.78	1.7475	1.5245	1.6819	1.2585
F	1.7766	1.4449	1.1876	1.1147	0.8876
G	1.5946	0.7369	0.5472	0.5357	0.5863
H	2.0786	1.9426	1.0258	0.9533	0.7038
I	1.9875	1.4169	1.5012	1.6386	2.1728
J	4.1046	3.328	3.1016	2.0013	1.7367

Mean time to sort a pair from the different sorting modes, (secs.)

Table 9

3	5	7	9	12
2.6363	1.6662	1.4233	1.2019	1.1752

Overall mean time per pair for each sorting mode

For subjects A and B,t was estimated from the mean t of the remaining 8 subjects despite the fact that analysis of variance showed a significant difference in t between subjects:-

	Mean square	DF	F	P
sorting modes	1462.4	4	99.568	<< 0.001
subjects	79.67	7	5.425	0.001
residual	14.69	28		

Subjects	Sorting Modes				
	3	5	7	9	12
a	0	0	7	2.1	-
	0	0	20.3	3.3	-
	0	0	26.5	3.3	-
	0	0	0	0	-
b	75	-	7	9.75	4.84
	0	-	13.8	12.9	6.87
	0	-	0	19	6.87
	0	-	0	0	8.22
c	0	19	7.02	3.31	1.4
	0	9.75	20.28	3.31	2.78
	0	0	0	6.56	5.52
	0	0	0	12.89	10.88
d	75	19	7.02	3.31	5.52
	75	36	20.28	6.56	8.22
	0	0	13.78	15.97	10.88
	0	0	0	15.97	16.08
e	0	19	7.02	3.31	1.4
	0	19	0	3.31	1.4
	0	0	0	0	5.52
	0	0	0	0	6.87
f	0	19	7.02	3.31	1.4
	0	19	13.78	3.31	2.78
	0	0	0	21.97	4.15
	0	0	0	24.89	5.52
g	0	19	7.02	3.31	2.78
	0	0	7.02	6.56	2.78
	0	0	20.28	9.75	10.88
	0	0	32.53	0	14.79
h	75	0	7.02	3.31	1.4
	0	0	13.78	3.31	2.78
	0	0	29.56	3.31	5.52
	0	0	0	14.44	21.97
i	75	19	13.78	3.31	2.78
	0	19	20.28	6.56	2.78
	0	0	0	6.56	6.87
	0	0	0	19.0	8.22
j	0	51	7.02	6.56	1.4
	0	0	7.02	0	5.52
	0	0	7.02	0	6.87
	0	0	20.28	0	10.22
Overall means	9.375	8.367	9.025	6.372	6.309

Table 10

Table 10

Sorting modes against random (unexplained) variance $(I - r^2) \times 100$ for each pack in each sorting mode.

A two way analysis of variance on Table 8 produced:-

	Mean square	DF	F	P
subjects	4.20	9	237.63	$<< 0.001$
sorting modes	3.76	4	212.71	$<< 0.001$
residual	0.018	36		

Thus subjects differ significantly in their mean sorting times and the difference in the sorting mode mean sorting times is significant.

A form of trend analysis on the sorting mode mean sorting times produced $z = 3.10$ suggesting the downward trend with increasing N was significant.

In order to assess the relative accuracy of the different sorting modes the rank correlation of each subject's sortings was calculated from Σd^2 and the measure $(I - r_s'')$ x 100 was taken as the percentage of the total variance involved in the correlation attributable to random factors.

This produced table 10.

Analysis of variance on these data produced:-

	Mean Square	DF	F	P
Subjects	329.05	7	1.80	0.0955
Modes	67.95	4	0.37	NSD
Interaction	177.17	28	0.97	NSD
Residual	182.86	120		

Thus there was no difference in the accuracy with which the items were sorted in the difference modes, and only 'subjects' approached significance.

The overall percentage of random error variance in the different sorting modes suggested decreasing error variance with increasing N. A form of trend analysis was carried out on the results but no trend

was found ($z = 0.582$).

The results were further examined for the relationship between $(I - r^2)$ and mean sorting time per pair. For each sorting mode each subject's mean sorting time per pair was correlated with the total error variance over the four packs sorted. These were then averaged using Fisher's z to produce $r = -0.328$. To find the significance of this r the t ratios were computed for the individual correlations and the probabilities were summed under the directional hypothesis of a positive true correlation. This produced $p = 0.4$. It was therefore concluded that time spent on discriminating did not correlate with accuracy of discrimination.

CHAPTER 6

Discussion

The present studies exhibited no evidence of a pure 'handedness' bias in the category scaling experiments, which conflicts with Matthews' (1929) finding. The most significant obvious bias demonstrated was a 'size' bias in which items tended to be massed at the 'large' end of the attenuated response continuum. The response continuum was so defined as to represent the total range of the items presented, and thus a significant underlying tendency was again demonstrated. The primary result of this was the non-linear distribution of item median category placings, the logarithmic relationship of psychophysics, which may be understood in terms of smaller proportional differences (approaching the least noticeable difference) with unit increase in size. Thus, while item median category placings show a powerful manifestation of end effect, when related to the defined response continuum the results are logical and comprehensible. However, when it comes to combining scales or scale items whose response continua cannot be shown to be coincident, problems of homogeneity may be met with. Ideally items or scales to be combined into a single scale should exhibit ranges of possible responses such that no attenuation effects should be obvious within the total range covered. Any item which shows manifest attenuation effects within the total range should not be accepted without some weighting of the attenuated responses such as the normal transformation described elsewhere. However, this transformation has the disadvantage of not coping rationally with the extreme category of any continuum, which is the

precise area where it is potentially of the greatest value.

A corollary of the effects of attenuation is the invalidity of the assumption of normal distribution of responses over categories outside the middle range of any attenuated response continuum. The adverse tendency of the Likert t-test item selection technique in graph 3 to select only middle range items is most probably a direct result of this; as also is the biasing of any scale derived from the scale discrimination technique of Edwards and Kilpatrick towards the extremes, since this method involves the elimination of the 50% of items with the greatest interquartile ranges, which from graphs 2, 12 and 13 can be seen to be concentrated within the midrange. The Likert method will produce a scale relatively less sensitive to subject scale values towards the extremes, while the Edwards and Kilpatrick method will be less sensitive to midrange subject scale values. The advantages of one technique over the other are difficult to establish, and may well depend on the purpose to which any derived scale is to be put. Of course the ideal scale would be equally sensitive to the whole subject scale value range, but in the present study this ideal appears to be elusive.

Thurstone selected items on the basis of item scale values with secondary consideration being given to item interquartile range. This would appear to be rather better than either the Likert or the Edwards & Kilpatrick methods, in that it is designed to produce an even density of items over the range considered. However, if the items are actually distributed objectively as in experiments A or B above, then in the first case, from graph 16 it will be seen that item objective scale value of the constructed scale will tend to be distributed approximately exponentially; while in the second case there will be under-representation of the lowest item objective scale values, which will in fact have the opposite effect to the former. Thurstone's second choice criterion is suspect on the grounds of the great variability in

both item median category placings and item interquartile range over categories. Where there is variability of this order in items of this kind, items of a more obtuse nature may well have any intrinsic invariability masked, and choice will be little better than random about each item scale value range. However, the nature of the items and the statistics derived justify the use of interquartile range, though under the circumstances, and notwithstanding the above remarks on the invalidity of the assumption of normal distributions in the forms of data studied here, means and standard deviations would represent considerably more stable statistics on which to base these choices.

So far we have considered the effects of response continuum attenuation. The present study also demonstrates the effects of stimulus attentuation, and suggests the confirmation of Upshaw's (1965) conclusions. Thus where items do not fully represent the total range of a subject's conceptual position within any defined anchors, i.e., where some area of a subject's understanding of the topic represented by the items presented is relatively less represented by items descriptive of that position, there is a tendency to 'fill in' with items representing adjacent points of view. Again this will tend to pervert a derived scale. Two circumstances may be described which fulfill this condition. Firstly, the span of items may not completely cover the whole span of a homogeneous population's understanding of the topic represented by the items; and secondly, a group may have a concept of the topic relatively attenuated at one 'end' and 'extended' at the other. Both these instances were demonstrated in a number of studies on the influence of attitude on scale values, most notably Kelley, Hovland, Schwartz & Abelson (1955), Hovland & Sherif (1952) and Upshaw (1962 & 1965).

Of special interest in this work is the confirmation of subjective anchors. These were shown to be temporarily established where no obvious

anchor existed, and to disappear with increasing familiarity with the items sorted. In fact this situation is probably closest to actual test construction situations in that the presented anchors are not completely related to the items to be scaled without actually being those items. One result is to nullify the above criticisms of Likert's and Edwards and Kilpatrick's item selection techniques since some midrange items will tend to be rejected from the former (assuming the apparent correlation between the t statistic and interquartile range seen in graphs 2 & 3 is meaningful and significantly positive), and selected in the latter.

The above sumarises the perturbing influences on item category scale values as demonstrated in the present experiments. They may be demonstrated as arising from the relative attenuation of response continua or items scaled. The elimination of such perturbations would be a significant first step towards the construction of an 'absolute' scale of attitude or semantics as discussed implicitly in the introduction above. However, even if it were possible to construct and present items representing every opinion from zero attitude to infinite attitude, stimulus attenuation by virtue of a subject's limited concepts and experiences would make many items meaningless, and thus in relation to the defined infinite response continuum those items which are meaningful would be subject to the effects of filling in with all the consequent perturbations in scale value and interquartile range (or standard deviation.)

A further solution would be to free the subject from questions of absolute stimulus value, and to consider instead relationships between items. It can more rationally be stated by a subject 'X has more of the attribute than Y', than 'X has P amount of the attribute'. Here the obvious solution of the problem is in terms of paired comparisons.

Thurstone based his law of comparative judgement on the assumption that all the items presented were unidimensional, and hence indicated

no means for item selection.

Gutman questioned this assumption (admittedly in the context of category scaling) but failed to show how items existing on different dimensions could be separated. Modern techniques and computer technology provide adequate models, such as cluster analysis, to accomplish this complex task. For reasons stated elsewhere the writer cannot accept Thurstone's assumptions in his case V of the law of comparative judgement.

Whatever method is used in analysing paired comparison data, there still remains the disproportionate increase in the number of pairs to be presented with increasing numbers of items to be examined $({}_nC_2 = \frac{N(N-1)}{2})$. The most promising method is an extension of Slater's multiple partial ranking. As already indicated, iterative means exist for generating sets of items such that p presentations of n items (${}_nC_2$ pairs per presentation) will represent all the constituent ${}_mC_2$ pairs of an item population m. The limitations of the relationships between p, n & m are such that many values of m exist which cannot be handled in this way, or for which n is so small that p is still unmanageable.

I will digress to explain a particular advantage of multiple ranking over single presentation methods or simple paired comparisons. Much of the variability in the above category experiments can be explained as being contributed by the first few items to be sorted, during which the nature of the items is only partly understood. Thus 'judgemental perspectives' change for each of the initial items sorted until an overall perspective becomes fixed. The distributions of interquartile ranges in AP_1 and BP_1 in experiment B above suggest that this final perspective was not established at all during the presentation of AP_1, if the low near midrange interquartile ranges can be taken as indicative of a form of primary judgemental perspective. This was substantiated in spirit in a subsidiary experiment not reported above

because of the specialised and extreme population used. In this experiment a number of theology students sorted P_1 in a non random order; i.e., every subject was presented with P_1 in a fixed random order. The only finding of note was the general decrease in item interquartile range with increasing familiarity with the items. Items presented at first had greater interquartile ranges than items approximately the same size presented at the end. A further finding of significant item mean category placing shift as compared to AP_1 could not be justified because of the nonhomogeneity of the two populations, though otherwise it would have supported the judgemental perspective model with the addition of a constant set between subjects.

As with the tangible unidimensional stimuli of the present experiment, so these varying judgemental perspectives will be even more strongly manifested where the items presented are not completely described by the presented anchors. Thus judgemental perspectives may shift with every item presented to represent only one aspect of the current stimulus item, within which the anchors become meaningful for that item. This is the situation which Gutman scalogram analysis was designed to detect.

Consider now paired comparison methodology. If two items are presented, then they must be judged, if any rational decision can be made at all, purely on some common aspect of their affective potential. Furthermore, the larger the number of items presented in multiple partial ranking, the more likely it is that all the item population will have been judged on the same aspect. Of course, for some items this common attribute might be by no means the major attribute; however this may be overcome in scale administration by presenting a number of the scale items to be assessed concurrently rather than discretely. The same problems will be partly overcome in category scaling procedures by allowing subjects to familiarise themselves with the items before assessing them.

To return to the projected extension of multiple partial ranking, where no adequately large n can be found to administer all the constituent pairs in m conveniently, it is projected that sets of sizes n, n-1, n+1, or any similar range, should be specified whichbetween them would accomplish the presentation of all the pairs. Furthermore, by specifying beforehand the level of probability at which any item pair will be deemed significantly different (binomial test), after a number of subjects have been presented with all item pairs, certain pairs may be excluded from subsequent subject presentations. With on-line computer techniques this process would be much simplified. The level at which item pairs would be accepted would be related to the number of items required in the final scale and the degree of 'misplacement' of item pairs which may be tolerated.

Implicit in the technique suggested here is an analysis of paired comparisons in no way related to the law of comparative judgement. If a set of items is unidimensional the above method will select a number of items, all of which are easily distinguishable one from the other (see Experiment C), and if multidimensional the data will be amenable to cluster analysis to separate them. This does not impute any relationship to any absolute scale value as in the law of comparative judgement case V, which ultimately says nothing about the relative ease of discrimination of items so scaled, though items selected by the present method could bear a direct relationship to case I of the law of comparative judgement.

With enough initial items, representing as many shades of opinion within the topic to be studied as possible, the present method will select items so distributed that where affect is relatively diffuse, in that any instrument based on absolute scale values would demonstrate relatively low between-item variance, selected item density would be relatively low. Thus the sharp cutoff criterion described in the introduction would be

fulfilled in that a precise transition between scale items would be observed, i.e., the scale would necessarily have a high Gutman coefficient of reproduceability.

All the present discussion on paired comparisons relies on the assumption that the degree to which a pair of items is distinguishable is a function of the items' discriminal dispersions and the difference between their modal discriminal processes (law of comparative judgement), and that the modal discriminal process is a simple function of absolute scale value. This has not been questioned elsewhere, and will be accepted here.

We may now discuss the relative ease of administration of category methods as opposed to the form of paired comparisons discussed above. Graph 18 indicates the mean time to discriminate a pair, when that pair is administered within a group of size N for ranking. This indicates a mean time per pair in simple paired comparisons in the order of 3.25 seconds, which accords with irregular and approximate times recorded during the administration of experiment C, and not recorded here ($\bar{m}$ = 2.9 sec).

During experiment D subjects showed no distress while sorting packs of 12, so let us assume that packs in excess of 12 will produce a t of 1 second. If all the items of P_1 had been sorted by multiple partial ranking, the time taken for a single subject to process all the constituent pairs would have been about 45 minutes, as against $2\frac{1}{2}$ hours in simple paired comparisons and 15 minutes for category sorting.

The question then is the work load it is acceptable to inflict upon a judge/subject. Whereas time taken in category sorting is probably linearly related to the number of items, in paired comparisons there is an exponential relationship, and the work load becomes intolerable as the number of items to be analysed increases. Of course some item pairs will be accepted after the first few subjects, and the

Graph 18

Mean times to sort a single pair when that pair is presented as a constituent part of N cards presented for ranking in Expt D. See table 9.

work load on subsequent subjects will be progressively reduced. However, if the work load on the first few subjects themselves is intolerable it may become expedient to administer only a proportion of the item pairs, to each initial subject, with curious results on the subsequent analysis due to irregular item pair sampling over subjects.

CHAPTER 7

Conclusion

Here we have, then, an attempt to return in social psychophysics to a measure of precision beyond that derived from category scaling and the application of case V of the law of comparative judgements. The form of notional analysis in terms of binomial distributions and just noticeable differences is not necessarily an attempt to establish an absolute scale of measured affect, since it is open to precisely the criticisms applied to category sorting above. That is, the just noticeable difference may bear a curvilinear or even quadratic relationship to any 'absolute' scale values.

However, insofar as the just noticeable difference from the above notional binomial model will represent significant changes in measured affect, it will eliminate much of the variance associated with confusion in the rank ordering of items selected by methods approximating to a function of absolute value.

It is to be regretted that the method of multiple partial ranking could not be expanded or tested in this paper. Despite the indulgence and infinite patience of members of Durham University Department of Mathematics and Durham University Computer Unit, and also extra-mural enquiries, the model of multiple partial ranking could not be formalised further than it was expressed above. In fact, the computer hardware for the on-line application was not even available at that time, and the software awaits the model's mathematical formulation.

The possibilities nevertheless exist for the application of this

techniques to the quantification of any affective principle, assuming that the form of item used in attitude methodology and semantic differentiation is analogous to the circles and discs analysed here. In fact these complex affective items cannot be assumed to be unidimensional. However, the principles of perturbation demonstrated here will nevertheless underly, and undermine, any measure taken with the techniques described.

BIBLIOGRAPHY

BIBLIOGRAPHY

ANAST, P. (1966)
JOURNALISM QUARTERLY 43 (4) pp.729-732.
PERSONALITY DETERMINANTS OF MASS MEDIA PREFERENCES

ANISFELD, H., BOGO, N., and LAMBERT, W.E. (1962)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY. 65 (4) pp.223-231

BARCLAY, J.E. and WEAVER, H.B. (1962)
JOURNAL OF SOCIAL PSYCHOLOGY. 58 pp. 109-120
COMPARATIVE RELIABILITIES AND EASE OF CONSTRUCTION OF THURSTONE AND LIKERT ATTITUDE SCALES.

BARCLAY, W.D. (1964)
JOURNAL OF ADVERTISING RESEARCH. 4 pp. 30-34.
SEMANTIC DIFFERENTIAL AS AN INDEX OF BRAND ATTITUDE.

BENNET, G.K. SEASHORE, H.G. and WESMAN, A.G. (1947)
DIFFERENTIAL APTITUDE TESTS MANUAL. NEW YORK: PSYCHOLOGICAL CORPORATION.

BENTLER, P.M. (1969)
JOURNAL OF PSYCHOLOGY 71. pp. 33-40.
SEMANTIC SPACE IS (APPROXIMATELY) BIPOLAR.

BEVAN, W., MAIER, R., and HELSON, H. (1963)
AMERICAN JOURNAL OF PSYCHOLOGY 76. pp. 464-469.
THE INFLUENCE OF CONTEXT UPON ESTIMATION OF NUMBER.

BEVAN, W. and TURNER, E. (1964)
JOURNAL OF EXPERIMENTAL PSYCHOLOGY 67. pp. 458-462.
ASSIMILATION AND CONTRAST IN THE ESTIMATION OF NUMBER.

A.B.
BLANKENSHIP, (1940a)
SOCIOLOGY AND SOCIAL RESEARCH 25 pp. 12-18.
THE CHOICE OF WORDS IN POLL QUESTIONS.

BLANKENSHIP, A.B. (1940b)
JOURNAL OF APPLIED PSYCHOLOGY 24 pp. 27-30
DOES QUESTIONAIRE FORM INFLUENCE RESULTS.

BLANKENSHIP, A.B. (1966)
JOURNAL OF ADVERTISING RESEARCH 6 pp. 13-17.
LET'S BURY PAIRED COMPARISONS.

BRINTON, J.E. (1961)
PUBLIC OPINION QUARTERLY 25 pp. 289-295
DERIVING AN ATTITUDE SCALE FROM SEMANTIC DIFFERENTIAL DATA.

BRUVOLD, W.H. (1969)
JOURNAL OF EXPERIMENTAL PSYCHOLOGY 81 pp. 230-234
CATEGORY AND SUCCESSIVE INTERVALS SCALES FOR RATING STATEMENTS AND STIMULUS OBJECTS

CATTEL, R.B. and DIGMAN, J.M. (1964)
BEHAVIOURAL SCIENCE 9 pp. 341-358.
STRUCTURE OF PERTURBATIONS.

COHEN, J. and HANSEL, M. (1956)
RISK AND GAMBLING.

COREY, S.M. (1937)
JOURNAL OF EDUCATIONAL PSYCHOLOGY 28 pp. 271-280
PROFESSED ATTITUDES AND ACTUAL BEHAVIOUR.

CRONBACH, L.S. (1950)
EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT 10 pp. 3-31.
FURTHER EVIDENCE ON RESPONSE SET AND TEST DESIGN.

DAS, J.P. and DUTTA, T. (1969)
ACT A PSYCHOLOGICA 29 pp. 85-92
SOME CORRELATES OF EXTREME RESPONSE SET.

DURBIN, J. (1951)
BRITISH JOURNAL OF PSYCHOLOGY IV pp. 85-90.
INCOMPLETE BLOCKS IN RANKING EXPERIMENTS.

EDWARDS, A.L. (1957)
TECHNIQUES OF ATTITUDE SCALE CONSTRUCTION.
N.Y. APPLETON-CENTURY-CROFTS INC.

EDWARDS, A.L. and KENNEY, K.C. (1946)
JOURNAL OF APPLIED PSYCHOLOGY 30 pp. 72-83
A COMPARISON OF THURSTONE AND LIKERT TECHNIQUES OF ATTITUDE SCALE CONSTRUCTION.

EDWARDS, A.L. and KILPATRICK, F.P. (1948)
JOURNAL OF APPLIED PSYCHOLOGY 32 pp. 374-384
A TECHNIQUE FOR THE CONSTRUCTION OF ATTITUDE SCALES.

FISHER, Sir Ronald A. (1967) (6th ed.)
STATISTICAL TABLES FOR BIOLOGICAL AGRICULTURE AND MEDICAL RESEARCH
OLIVER AND BOYD.

FRANCÈS, R. (1963)
JOURNAL DE PSYCHOLOGIE NORMALE ET PATHOLOGIQUE 60 (4) pp. 437-455
LIMITES ET NATURE DES EFFETS DE PRESTIGE II. NOTORIÉTÉ DE L'OEUVRE.

GARDNER, R.C. WONNACOTT, E.S. and TAYLOR, D.M. (1968)
CANADIAN JOURNAL OF PSYCHOLOGY 22(1) pp. 35-44.
ETHNIC STEREOTYPES: A FACTOR ANALYTIC INVESTIGATION.

GRANBERG, D. and ABOUD, J. (1969)
AMERICAN JOURNAL OF PSYCHOLOGY 82 pp. 221-227
A CONTEXTUAL EFFECT IN JUDGEMENTS OF VISUAL NUMEROUSNESS.

GREENBERG, B.S. (1966)
JOURNALISM QUARTERLY 43 (4) pp. 665-670.
MEDIA USE AND BELIEVABILITY: SOME MULTIPLE CORRELATES.

GREENBERG, M.G. (1965)
PSYCHOLOGICAL BULLETIN 64 pp. 108-112.
A MODIFICATION OF THURSTONE'S LAW OF COMPARATIVE JUDGEMENT TO ACCOMMODATE A JUDGEMENT OF "EQUAL" OR "NO DIFFERENCE."

GUILFORD, J.P. (1956)
FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION.
McGRAW HILL BOOK COMPANY INC.

GUTTMAN, L. (1944)
AMERICAN SOCIOLOGICAL REVIEW 9 pp. 139-150
A BASIS FOR SCALING QUALITATIVE DATA.

HOFMAN, J.E. (1967)
AMERICAN JOURNAL OF PSYCHOLOGY 80 pp. 345-354
AN ANALYSIS OF CONCEPT CLUSTERS IN SEMANTIC INTERCONCEPT SPACE.

HOVLAND, C.I. and SHERIF, M. (1952)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY 47 pp. 882-832
JUDGEMENTAL PHENOMENA AND SCALES OF ATTITUDE MEASUREMENT: ITEM DISPLACEMENT IN THURSTONE SCALES.

HUDSON, A.L. (1967)
NATURE 214 pp. 968-969
ARTS AND SCIENCE: THE INFLUENCE OF STEREOTYPES.

HUMM, D.G. and WADSWORTH, G. (1943)
THE INTERPRETATION OF THE HUMM-WADSWORTH TEMPERAMENT SCALE
LOS ANGELES: D.G. HUMM 1943.

HYMAN, H. (1944-1945)
PUBLIC OPINION QUARTERLY. WINTER 1944-45 pp. 557-559
DO THEY TELL THE TRUTH.

JONES, L.V. and THURSTONE, L.L. (1956)
JOURNAL OF APPLIED PSYCHOLOGY 39 pp. 31-36
THE PSYCHOPHYSICS OF SEMANTICS; AN EXPERIMENTAL INVESTIGATION.

KELLEY, H.H., HOVLAND, C.I. SCHWARTZ, M. and ABELSON, R.P. (1955)
JOURNAL OF SOCIAL PSYCHOLOGY 42 pp. 147-158
THE INFLUENCE OF JUDGES' ATTITUDES IN THREE METHODS OF SCALING.

KENDALL, M.G. (1948)
RANK CORRELATION METHODS.
CHARLES GRIFFIN & CO. LTD., LONDON.

KUTNER, B., WILKINS, C. and YARROW, P. (1952)
J. ABN. & SOC. PSYCH. XLVII pp. 649-652.
VERBAL ATTITUDES AND OVERT BEHAVIOUR INVOLVING RACIAL PREJUDICE.

LAMBERT, W.E., HODGSON, R.C., GARDNER, R.C. and FILLENBAUM, S. (1960)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY 60 pp. 44-51.
EVALUATIONAL REACTIONS TO SPOKEN LANGUAGES.

LaPIERRE, Richard T. (1934)
SOCIAL FORCES 13 pp. 230-237.
ATTITUDES VS. ACTIONS.

LINN, L.S. (1965)
SOCIAL FORCES XLII pp. 353-364.
VERBAL ATTITUDES AND SOCIAL BEHAVIOUR.

OPPENHEIM, A.N. (1966)
QUESTIONAIRE DESIGN AND ATTITUDE MEASUREMENT.
HEINEMANN.

OSGOOD, C.E. and LURIA, Z. (1954)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY 49 pp. 579-591.
A BLIND ANALYSIS OF MULTIPLE PERSONALITY USING THE SEMANTIC DIFFERENTIAL.

OSGOOD, C.E. SUCI, G.S. and TANNEMBAUM, P.H. (1957)
THE MEASUREMENT OF MEANING.
UNIVERSITY OF ILLINOIS PRESS, URBANA.

PHILIP, B.R. (1947)
CANADIAN JOURNAL OF PSYCHOLOGY 1 pp. 196-204.
GENERALISATION AND CENTRAL TENDENCY IN THE DISCRIMINATION OF A SERIES OF STIMULI.

ROBINSON, W.P. (1965)
PSYCHOLOGICAL REPORTS 16 pp. 419-422.
OWN ATTITUDE AND THURSTONE SCALE JUDGEMENTS.

RORER, L.B. (1965)
PSYCHOLOGICAL BULLETIN 63 pp. 129-156.
THE GREAT RESPONSE STYLE MYTH.

ROSS, B.M. and LEVY, N.A. (1960)
JOURNAL OF PSYCHOLOGY 49 pp. 133-137.
A COMPARISON OF ADJECTIVAL ANTONYMS BY SIMPLE CARD PATTERN FORMATION.

RUBIN, H.K. (1940)
A CONSTANT ERROR IN THE SEASHORE TEST OF PITCH DISCRIMINATION.
UNPUBLISHED MASTERS' THESIS, UNIVERSITY OF WISCONSIN.

RUGG, D. (1941)
PUBLIC OPINION QUARTERLY, MARCH 1941. pp. 91-92.
EXPERIMENTS IN QUESTION WORDING II.

RUGG, D. and CANTRIL, H. (1942)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY 37 pp. 469-495.
THE WORDING OF QUESTIONS IN PUBLIC OPINION POLLS.

RUNDQUIST, E.A. (1966)
PSYCHOLOGICAL BULLETIN 3 pp. 166-177.
ITEM AND RESPONSE CHARACTERISTICS IN ATTITUDE AND PERSONALITY MEASUREMENT: A REACTION TO L.G. RORER'S " THE GREAT RESPONSE STYLE MYTH".

SCHUCKER, R.E. (1959)
PSYCHOMETRIKA 24 pp. 273-276.
A NOTE ON THE USE OF TRIADS FOR PAIRED COMPARISONS.

SIEGEL, S. (1956)
NON-PARAMETRIC STATISTICS FOR THE BEHAVIOURAL SCIENCES.
McGRAW HILL BOOK COMPANY INC.

SIPOS, I. (1966)
STUDIA PSYCHOLOGICA VII pp. 286-305.
POKUS O EMPIRICKU ANALYZU.
SEMANTICKES NEURCITOSTI NICKTORYCH KATEGORII.

SIPOS, I. and KOLADA, S. (1966)
STUDIA PSYCHOLOGICA 8 pp. 162-164.
AN EXPLORATION INTO SEMANTIC ENTROPY.

SIPOS, I. (1967)
STUDIA PSYCHOLOGICA IX pp. 260-268.
POUZITIE PSYCHOFIZIKY NA
ODSTUPNOVANIE SEMANTICKEHO KONTINUA.

SIPOS, I. and ADAMICA, Z. (1967)
STUDIA PSYCHOLOGICA IX pp. 70-71.
SEMANTIC UNCERTAINTY AND VERBAL TESTS.

SJÖBERG, L. (1965)
SCANDINAVIAN JOURNAL OF PSYCHOLOGY 6 pp. 173-185.
A STUDY OF FOUR METHODS FOR SCALING PAIRED COMPARISONS DATA.

SLATER, P. (1961)
BIOMETRIKA 48 pp. 303 -312.
INCONSISTENCIES IN A SCHEDULE OF PAIRED COMPARISONS.

SLATER, P. (1965)
BRITISH JOURNAL OF MATHEMATICAL AND STATISTICAL PSYCHOLOGY 18
pp. 227-242.
THE TEST-RETEST RELIABILITY OF SOME METHODS OF MULTIPLE COMPARISON.

TARTER, D.E. (1966)
AN EMPIRICAL ANALYSIS OF ATTITUDE ACTION DISCREPANCY.
DISSERTATION PRESENTED TO THE GRADUATE COUNCIL OF THE UNIVERSITY OF
TENNESSEE IN PARTIAL FULFILLMENT OF REQUIREMENTS FOR DEGREE OF
DOCTOR OF PHILOSOPHY. UNIVERSITY MICROFILMS LTD. Aug. 1966.

TAVES, E. (1941)
ARCHIVES OF PSYCHOLOGY 265
TWO MECHANISMS FOR THE PERCEPTION OF VISUAL NUMEROUSNESS.

TERWILLIGER, R.F. (1962)
J. ABN. & SOC. PSYCH. 65 pp. 87-94.
FREE ASSOCIATION PATTERNS AS A FACTOR RELATING TO SEMANTIC
DIFFERENTIAL RESPONSES.

THURSTONE, L.L. (1925)
JOURNAL OF EDUCATIONAL PSYCHOLOGY XVI(7) pp. 433-451.
A METHOD OF SCALING PSYCHOLOGICAL AND EDUCATIONAL TESTS.

THURSTONE, L.L. (1927)
PSYCHOLOGICAL REVIEW 34 pp. 273-286.
A LAW OF COMPARATIVE JUDGEMENT.

THURSTONE, L.L. and CHAVE, E.J. (1929)
THE MEASUREMENT OF ATTITUDE.
CHICAGO: UNIVERSITY OF CHICAGO PRESS.

UPSHAW, H.S. (1962)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY 64 pp. 85-96.
OWN ATTITUDE AS AN ANCHOR IN EQUAL APPEARING INTERVALS.

UPSHAW, H.S. (1965)
JOURNAL OF PERSONALITY AND SOCIAL PSYCHOLOGY 2 pp. 60-69.
THE EFFECT OF VARIABLE PERSPECTIVES ON JUDGEMENTS OF OPINION STATEMENTS FOR THURSTONE SCALES.

VERNON, P.E. (1949)
JOURNAL OF ABNORMAL AND SOCIAL PSYCHOLOGY XLIV pp. 85-96
CLASSIFYING HIGH GRADE OCCUPATIONAL INTERESTS.

WAGENAAR, W.A. (1968)
PERCEPTION AND PSYCHOPHYSICS 3(5B)pp. 364-366.
SEQUENTIAL RESPONSE BIAS IN PSYCHOPHYSICAL EXPERIMENTS.

ZWAAN, E.J. (1964)
NEDERLANDSE TIJDSCHRIFT VOOR PSYCHOLOGIE EN HAAR GRENSGEBIEDEN 19 pp. 328-340.
DE AFHANKELIJKHEID VAN SUCCESSIEVE ANTWOORDEN IN DIVERSE GEDRAGSSITUATIES.

APPENDIX

Table 1, raw results expt. A

Table 2, raw results expt. B

Table 6, raw results expt. C

Table 7, raw results expt. D

Table 1

Raw results of experiment A.

Categories into which each subject sorted each card; cards numbered from smallest to largest.

Also distribution of cards over categories for each subject.

Subjects

LEFT SORTING — RIGHT SORTING

S13	S10	S30	S24	S1	S28	S26	S15	S21	S29	S17	S16	S7	S3	AS	DS	PE	TJ	C	SD	M	ITEM	NB	S2	S4	S5	S6	S8	S9	S11	S12	S14	S18	S20	S22	S23	S25	S27	ME	DP	AM	WH	Sim	IQR	Mdn
1	1	1	1	1	1	3	1	1	1	1	1	2	1	1	1	1	1	2	2	1	ca	1	1	1	2	1	3	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	0.6	0.61
1	1	1	1	2	2	1	1	1	1	2	1	2	2	1	1	2	1	1	1	1	cb	3	1	1	1	1	3	2	1	1	1	2	1	2	1	1	1	2	1	1	2	1	0.89	0.74
2	2	1	2	2	2	1	1	1	1	1	1	1	1	1	1	2	3	2	2	1	cc	2	1	1	2	2	5	2	2	1	2	1	2	1	1	1	1	1	2	1	2	1	1.05	0.94
2	1	1	2	1	4	2	3	1	1	1	1	1	2	1	1	2	1	4	3	2	cd	1	1	1	1	2	3	2	3	1	1	1	3	2	1	2	2	2	3	1	3	1	1.45	1.13
2	1	1	2	1	3	2	3	2	1	2	1	2	2	2	1	2	2	3	2	2	ce	1	1	1	1	2	5	2	1	2	3	1	3	2	1	2	1	2	4	2	3	2	1.11	1.38
2	1	1	2	1	3	1	2	2	1	2	1	1	1	1	1	2	4	2	2	2	cf	2	2	2	2	2	4	2	2	1	2	1	3	2	1	2	1	2	2	1	3	2	1.04	1.29
3	3	1	2	2	2	2	2	2	2	1	1	9	3	2	1	2	2	3	3	2	cg	1	2	1	2	2	2	2	2	2	3	1	3	2	1	2	2	1	4	2	2	2	0.88	1.52
3	2	1	3	1	4	2	2	2	2	3	2	2	2	2	2	2	3	2	2	2	ch	3	2	2	1	2	4	2	2	2	3	2	3	4	2	3	2	2	3	2	3	2	0.99	1.71
3	2	2	3	3	6	2	3	2	2	2	5	3	3	1	2	2	3	3	4	3	ci	2	2	2	2	2	5	2	3	2	4	2	4	3	2	2	2	3	4	2	2	3	1.29	1.98
2	2	3	2	3	2	8	3	2	2	2	1	2	2	2	3	2	3	3	4	3	cj	2	2	2	2	4	5	3	3	2	4	2	3	3	7	3	3	3	3	2	3	2	1.22	2.09
3	3	3	3	3	3	3	6	2	2	4	2	4	3	3	2	2	2	4	4	3	ck	2	2	3	2	2	6	3	4	2	3	3	3	3	2	2	3	2	4	4	3	2	1.21	2.36
3	3	2	2	3	5	2	4	2	2	3	2	2	3	2	2	2	2	3	3	2	cl	3	2	2	5	2	5	2	2	3	3	3	3	4	2	3	2	3	2	3	3	3	1.16	2.09
4	2	3	2	3	3	6	4	3	2	3	2	8	3	2	2	3	4	3	4	3	cm	2	2	2	2	3	6	4	3	3	4	3	3	5	2	3	3	5	4	2	3	2	1.44	2.5
4	6	2	2	5	3	3	4	2	3	4	3	9	3	5	4	3	3	4	5	3	cn	3	3	3	4	3	5	4	3	3	4	3	5	5	5	4	5	6	4	3	2	4	1.69	3.14
3	4	3	3	7	3	2	7	2	3	4	1	3	2	4	5	3	3	4	6	4	co	4	2	3	4	3	5	3	4	3	5	4	4	4	3	4	3	4	3	3	3	3	1.36	2.92
3	5	2	3	6	7	3	6	4	3	3	2	9	2	3	3	2	4	4	4	4	cp	3	2	3	5	3	6	4	5	3	6	4	4	6	4	3	5	4	4	3	3	3	1.85	3.14
4	5	3	3	4	4	4	8	3	2	3	8	4	3	4	4	3	5	5	4	3	cq	3	2	3	4	5	6	4	6	4	5	4	4	6	5	3	4	3	4	3	4	3	1.47	3.41
4	5	3	3	3	4	5	7	3	2	4	5	7	4	5	4	2	4	4	4	3	cr	5	4	3	4	4	5	4	4	3	4	6	5	6	2	4	4	6	4	3	5	3	1.49	2.56
4	5	5	4	5	6	4	6	6	3	5	3	4	3	4	3	3	5	4	5	4	cs	4	3	4	4	3	7	6	6	3	5	6	4	6	4	6	5	6	6	3	5	4	2.02	4.06
4	7	2	3	7	3	4	5	3	4	5	2	8	8	4	3	3	5	4	4	4	ct	4	3	4	2	4	5	5	5	3	6	4	4	5	5	4	6	5	5	3	7	4	1.7	3.75
4	4	5	3	8	3	5	8	3	5	6	5	9	4	5	3	3	5	4	7	4	cu	6	3	3	7	5	6	3	7	2	5	3	5	7	6	6	6	6	3	3	9	6	2.72	4.39
3	4	2	2	7	4	5	9	8	3	3	5	10	6	4	5	3	4	3	7	3	cv	4	3	4	5	5	8	6	5	4	5	4	5	7	6	4	6	7	3	3	5	4	2.26	4.05
4	5	6	4	8	4	5	6	6	3	5	3	3	4	5	5	4	6	3	6	5	cw	6	3	4	8	5	6	5	6	4	6	5	5	6	6	4	7	8	4	5	4	3	2.0	4.5
4	5	4	4	4	4	5	8	4	6	6	5	4	3	4	6	3	5	7	6	4	cx	5	3	8	4	3	8	8	4	4	4	6	5	7	4	4	6	5	4	4	6	4	2.04	3.97
3	6	3	6	6	3	3	9	4	5	7	5	10	4	5	6	6	5	6	6	5	cy	4	5	9	4	5	7	8	6	4	6	6	5	6	5	7	6	6	6	4	7	3	1.86	5.11
4	4	4	7	7	5	6	5	4	4	6	7	8	6	4	7	3	6	6	6	4	cz	4	4	4	5	4	6	5	5	4	6	4	4	6	6	4	7	7	6	4	4	6	2.16	4.7
4	7	4	5	8	4	3	7	3	4	6	8	8	5	3	5	5	6	4	6	4	aa	6	5	5	8	3	8	7	6	5	6	5	6	7	7	4	7	7	5	3	6	4	2.54	4.94
5	7	3	4	5	9	6	7	3	4	4	7	9	5	7	6	5	6	4	6	5	ab	4	4	5	5	3	9	9	5	4	6	6	5	7	8	8	7	8	6	4	4	5	2.61	4.95
3	8	5	6	7	9	5	9	4	8	7	3	9	4	4	7	9	6	6	6	6	ac	4	5	6	8	4	7	6	8	4	7	6	6	7	8	6	5	6	6	5	5	6	2.94	5.58
4	5	4	6	8	5	6	8	7	3	8	7	10	5	5	7	9	8	7	8	5	ad	6	4	3	4	5	8	5	7	5	8	6	6	7	6	5	6	7	6	5	7	9	2.5	5.69
6	7	7	6	9	5	8	6	3	5	8	7	10	6	4	8	3	7	4	5	4	ae	4	4	7	7	4	7	7	8	4	8	7	5	7	6	8	5	7	6	7	6	4	2.63	5.93
6	7	5	6	6	9	6	8	4	7	8	9	9	5	5	6	8	6	7	7	4	af	5	8	8	4	4	8	7	6	4	6	6	5	8	6	7	6	7	6	5	7	5	2.14	5.79
5	7	6	8	9	6	5	7	4	7	8	5	9	5	5	7	6	6	6	6	6	ag	5	6	8	9	4	10	8	7	4	7	6	5	7	6	8	8	8	8	4	9	5	2.5	5.95
5	8	6	4	8	7	7	8	4	7	7	5	6	4	6	8	5	5	6	7	5	ah	6	6	7	7	6	9	7	5	5	7	6	6	8	6	4	9	8	4	5	5	4	2.33	5.72
6	8	6	7	9	9	6	10	8	7	8	6	10	6	6	8	7	7	7	7	6	ai	5	5	5	5	8	7	8	8	5	8	6	7	8	9	8	4	7	7	8	8	7	2.0	6.08
6	7	5	8	9	7	9	10	5	6	7	6	10	6	7	8	6	4	4	7	6	aj	6	4	8	7	7	9	10	8	4	7	7	5	8	8	6	7	7	5	5	8	8	2.25	6.41
6	7	6	8	5	8	6	9	4	7	8	5	9	4	8	8	6	6	6	6	5	ak	9	6	5	9	4	9	10	8	7	7	7	6	7	7	9	8	9	7	9	6	8	2.44	6.69
6	9	7	8	8	5	7	7	7	6	8	4	9	5	6	9	6	7	7	7	5	al	5	4	5	9	6	10	7	6	5	9	7	6	8	8	9	6	8	5	6	8	5	2.46	6.28
5	7	6	8	8	8	8	9	5	7	9	8	10	6	6	8	6	7	7	7	5	am	6	5	6	9	6	9	7	5	8	9	7	8	8	6	8	7	8	7	8	7	9	2.06	6.75
4	8	9	9	8	6	8	10	4	9	8	8	9	7	8	9	9	7	7	8	5	an	6	5	8	8	8	10	9	9	6	9	8	7	9	9	8	8	8	8	7	8	6	2.67	7.59
5	9	7	9	8	6	7	8	8	8	10	6	10	5	9	9	10	6	7	9	5	ao	6	9	10	9	7	9	8	10	6	7	7	7	8	9	7	7	8	5	6	8	8	2.33	7.28
6	7	5	8	9	8	8	9	9	8	8	7	10	8	6	9	9	8	8	9	5	ap	6	6	8	6	6	9	8	9	6	8	9	6	8	8	8	8	7	7	6	9	8	2.1	7.41
6	6	7	8	8	7	8	7	8	4	7	4	10	5	6	9	6	8	5	8	6	aq	6	7	8	6	7	9	8	9	5	9	8	8	8	5	9	8	8	7	8	7	8	2.02	7.03
7	9	7	9	9	8	10	9	6	8	10	5	10	7	9	9	10	9	7	8	6	ar	6	5	4	10	6	9	9	7	7	9	8	6	9	9	8	7	8	7	9	10	7	2.36	7.75
6	9	7	8	9	9	10	9	8	9	9	8	11	9	8	8	10	9	8	8	6	as	7	6	6	9	5	9	8	10	6	8	7	8	8	5	9	7	9	8	6	9	6	2.04	7.71
8	7	6	8	8	8	9	10	4	8	7	5	10	8	6	10	6	8	8	8	6	at	6	8	7	8	7	10	8	7	6	10	7	7	8	7	10	8	9	7	8	7	9	1.7	7.25
7	9	8	9	10	6	10	8	8	9	10	6	10	6	7	9	7	10	7	8	5	au	8	7	9	9	5	10	10	9	7	8	9	7	9	7	9	8	9	6	7	10	9	2.28	7.93
6	10	8	9	9	6	9	10	8	9	8	8	10	7	8	9	10	7	7	9	7	av	6	11	10	7	8	10	9	8	8	9	8	8	9	8	11	9	8	8	10	8	8	1.7	7.9
8	9	7	9	10	9	9	9	5	9	9	10	11	7	8	10	9	7	6	8	7	aw	8	6	8	9	10	9	9	8	7	9	7	6	8	9	10	9	8	8	9	10	9	1.7	8.16
8	10	9	9	10	10	9	4	8	9	8	9	10	6	8	9	11	9	9	8	7	ax	8	7	6	9	9	10	11	9	7	7	8	8	8	8	10	8	8	7	8	9	8	1.56	8.04
8	10	10	9	9	8	8	9	10	9	8	8	10	8	8	10	11	10	8	10	8	ay	8	6	9	10	7	10	10	10	8	10	10	8	9	9	10	8	9	8	7	7	10	1.9	8.44
6	9	9	8	9	10	9	10	8	9	9	10	10	9	9	10	10	7	7	10	6	az	8	11	8	9	11	11	9	10	8	10	9	8	8	10	10	8	8	9	9	10	9	1.64	8.61
7	9	7	9	10	7	10	9	9	10	10	10	10	10	9	10	9	10	10	9	7	ba	9	9	5	10	9	11	9	9	8	9	8	8	9	10	10	9	9	9	9	9	8	1.20	8.66
7	9	8	9	10	8	8	10	9	9	10	9	10	6	8	10	7	8	9	9	8	bb	7	7	8	10	8	11	11	8	8	7	8	8	9	8	10	9	9	7	9	9	9	1.62	8.12
8	10	8	10	10	10	10	9	8	8	8	10	11	7	9	10	10	9	8	8	9	bc	11	10	7	10	11	9	11	9	8	9	8	9	10	8	10	10	10	7	10	9	10	1.83	8.94
9	10	8	10	10	10	11	10	10	10	9	9	11	7	9	10	9	10	10	10	8	bd	9	11	9	10	9	10	10	9	7	9	9	7	9	10	10	7	10	6	10	10	9	1.32	9.08
8	10	9	9	10	11	11	10	9	8	10	10	11	7	8	10	9	10	8	8	8	be	8	10	9	9	10	11	11	10	6	10	9	9	9	10	11	10	10	9	10	10	10	1.57	9.15
8	10	10	9	10	9	10	10	6	10	10	9	11	8	10	10	11	9	8	9	6	bf	8	9	9	10	9	10	11	9	9	10	10	10	9	10	10	8	9	8	9	10	10	1.38	9.03
8	11	9	9	10	10	9	11	10	11	11	8	10	7	10	10	11	11	9	8	8	bg	9	10	7	10	10	10	10	10	9	11	10	8	10	10	10	10	10	9	10	10	9	1.35	9.33
7	10	10	9	10	9	10	10	9	10	9	10	11	8	10	10	9	9	10	10	9	bh	8	10	10	10	9	10	10	8	9	10	10	9	10	10	10	10	10	9	10	10	10	1.02	9.25
10	11	8	10	10	11	10	10	10	10	10	9	11	8	9	11	11	8	8	10	9	bi	10	8	10	11	11	11	11	10	10	10	10	10	10	11	11	10	10	8	10	10	10	1.0	9.57
9	10	9	10	10	9	11	11	10	10	10	10	11	9	10	10	10	8	10	10	10	bj	11	10	10	9	10	11	11	11	8	10	10	10	9	10	10	9	10	9	9	10	10	0.88	9.48
9	10	10	9	10	10	10	10	10	10	9	10	11	9	11	11	10	11	10	10	9	bk	9	11	10	10	11	11	11	10	8	10	9	8	10	10	10	10	11	10	10	10	10	0.88	9.52
9	10	9	10	11	11	11	10	10	9	10	11	11	11	10	10	10	10	10	10	10	bl	10	10	11	11	8	11	9	9	10	9	10	10	10	10	10	11	10	10	11	11	11	1.14	9.66
10	10	9	10	11	9	11	11	11	11	10	10	11	11	11	11	9	10	8	9	9	bm	11	11	9	10	10	11	10	11	8	11	9	11	10	11	11	11	11	9	11	10	11	1.43	10.02
9	11	9	10	11	10	11	10	7	10	10	8	11	11	10	11	10	10	10	9	10	bn	10	10	11	8	11	11	10	10	10	10	11	9	10	10	10	9	10	8	10	10	11	0.91	9.54
10	11	10	10	11	11	11	10	10	10	10	11	11	11	10	11	11	9	9	10	10	bo	10	11	11	10	10	11	10	11	9	10	11	11	10	11	11	11	11	10	11	10	11	1.08	10.02
10	11	11	10	11	10	11	11	11	11	11	9	11	11	10	11	11	11	11	9	10	bp	9	10	10	10	11	11	11	11	9	11	11	10	11	11	11	11	11	10	10	11	11	0.99	10.24
10	11	8	10	11	11	11	11	9	11	10	10	11	9	11	11	11	11	11	11	10	bq	11	11	11	11	10	11	10	11	8	10	10	11	10	10	11	11	10	10	11	11	11	1.06	10.18
11	11	10	10	11	10	11	10	10	11	11	10	11	10	11	11	10	11	8	10	11	br	10	10	8	11	11	11	10	10	9	11	10	11	10	11	11	10	11	11	11	10	10	1.08	10.02
11	11	10	11	11	11	11	10	11	11	10	11	11	10	11	11	9	11	11	10	10	bs	11	11	11	11	10	11	11	11	10	10	10	11	10	11	11	11	11	10	11	11	11	0.82	10.29
11	11	11	10	11	11	11	11	10	10	11	11	11	11	11	11	10	10	10	10	11	bt	11	11	11	11	11	11	11	10	10	11	11	11	11	11	11	11	11	9	11	10	11	0.68	10.34

S13	S10	S30	S24	S1	S28	S26	S15	S21	S29	S17	S16	S7	S3	AS	DS	PE	TJ	C	SD	M	categories	NB	S2	S4	S5	S6	S8	S9	S11	S12	S14	S18	S20	S22	S23	S25	S27	ME	DP	AM	WH	Sim	IQR	Mdn
2	5	8	2	5	1	3	3	4	6	4	9	3	3	6	6	1	3	1	1	3	1	4	5	6	4	2	0	1	3	5	3	6	2	2	7	2	5	3	2	5	1	4		
5	5	6	10	3	4	7	3	10	9	5	6	6	7	6	6	13	4	4	5	6	2	6	11	6	9	9	1	9	5	7	2	4	1	5	6	7	5	6	3	6	5	7		
10	3	8	9	6	9	6	4	8	7	6	4	3	9	3	5	11	6	8	3	8	3	6	7	8	0	8	3	4	5	8	5	5	9	3	1	6	4	4	6	11	10	8		
13	4	4	5	2	8	3	3	11	5	5	2	3	7	9	3	1	6	13	8	10	4	9	7	6	9	9	2	5	4	12	6	6	7	3	3	10	3	2	12	6	4	8		
5	7	6	1	4	5	7	2	4	3	3	11	0	8	8	4	3	7	2	3	12	5	6	7	6	6	8	9	4	7	6	5	2	12	3	6	1	5	3	5	6	5	4		
11	3	7	5	3	7	6	5	4	3	4	4	0	10	7	4	8	9	7	10	10	6	14	7	4	2	5	7	3	7	7	8	11	9	6	10	4	8	6	10	4	4	5		
5	11	8	2	5	5	3	7	3	6	6	5	1	8	3	4	3	7	12	8	4	7	2	4	5	5	5	5	6	5	6	8	9	6	10	5	3	10	9	10	4	8	2		
8	4	7	10	11	7	9	8	9	7	13	9	5	6	9	6	1	7	10	12	5	8	8	3	11	6	5	5	9	9	10	6	8	11	15	9	8	11	14	9	4	6	8		
5	9	9	15	10	10	7	13	6	10	7	7	11	5	7	10	11	8	4	9	5	9	6	3	7	11	5	12	8	10	6	11	7	4	11	6	5	7	8	8	9	8	9		
6	12	7	12	14	9	9	18	10	10	15	11	21	3	8	14	12	9	8	12	7	10	5	9	7	13	9	12	12	11	5	13	10	5	12	12	16	7	10	6	10	17	9		
3	9	2	1	9	7	12	6	3	6	4	4	19	6	6	10	8	6	3	1	2	11	6	9	6	7	7	16	11	6	0	5	4	6	2	7	10	7	7	1	7	4	8		

Distribution of items over categories

Table 2

Raw results of experiment B.

Categories into which each subject sorted each card; cards numbered from smallest to largest; for both packs.

Also distribution of cards over categories for each subject, for each pack.

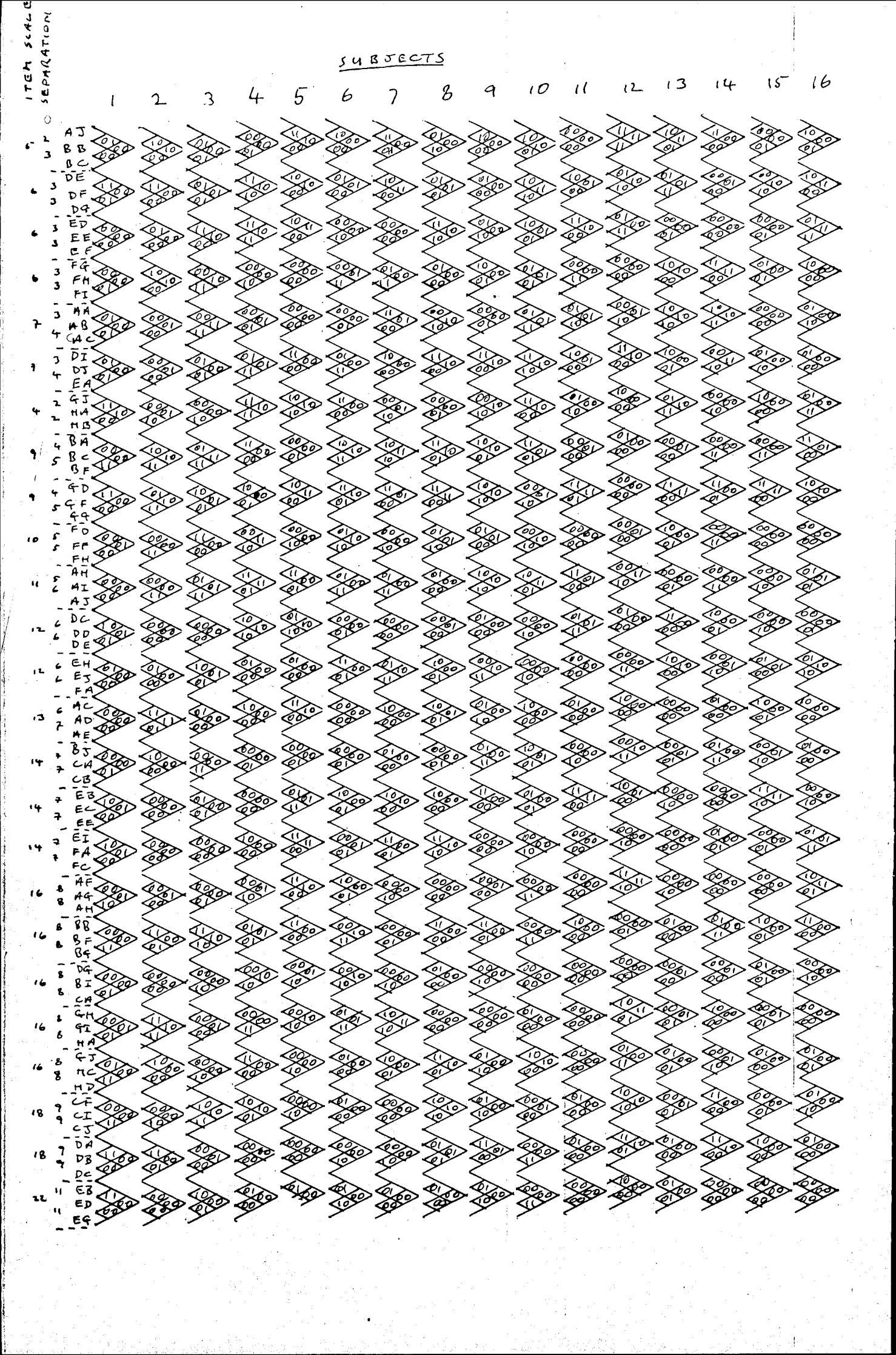
SUBJECT	1		2		3		4		5		6		7		8		9		10		11		12		13		14		15		16	
LARGEST CARD	L		L		R		R		L		L		R		R		L		L		R		R		L		L		R		R	
ORDER	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
PACK	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
AA	1		3		1		2		3		1		1		1		1		1		1		2		1		1		1		1	
AB	1		4		2		1		1		1		1		2		1		1		1		1		2		1		9		1	
AC	1		4		2		1		1		1		1		1		2		2		1		3		1		1		1		1	
AD	1		2		1		1		3		1		1		2		2		3		1		2		2		1		1		1	
AE	2		4		2		1		1		1		1		2		2		1		2		3		3		2		1		1	
AF	1		4		2		2		1		1		1		2		2		2		1		2		1		1		2		1	
AG	1	1	3	3	1	1	1	2	2	2	1	1	1	1	2	4	2	1	2	1	1	1	2	2	3	1	2	1	2	1	2	7
AH	1		3		2		3		2		1		1		2		4		2		1		2		1		1		2		1	
AI	1	2	3	3	1	3	3	2	3	3	1	3	1	2	3	4	2	2	2	2	1	1	3	1	2	1	2	3	3	1	1	1
AJ	2		4		2		2		3		1		2		2		2		3		1		4		2		2		2		3	
BA	2		2		1		3		2		2		1		3		2		3		3		3		4		1		3		1	
BB	2	1	3	5	5	2	2	4	2	3	2	3	1	3	3	1	2	3	1	1	1	2	3	3	2	2	2	2	2	2	4	1
BC	2	1	4	4	3	3	3	3	4	2	1	2	1	4	3	2	3	2	3	1	1	2	5	4	1	3	2	1	3	2	1	2
BD	1		3		4		2		1		1		1		4		3		3		1		2		3		2		2		1	
BE	3	3	3	4	3	1	2	2	2	2	1	1	1	1	3	3	3	3	1	2	1	1	3	3	3	2	2	1	1	1	4	1
BF	2	1	4	3	5	3	1	4	4	3	1	1	1	2	3	2	3	2	3	2	2	2	6	2	3	1	1	3	2	2	1	1
BG	1	1	3	4	1	5	2	2	4	4	1	2	1	3	4	2	1	1	3	1	1	1	4	3	3	3	2	3	2	1	1	2
BH	2	2	3	5	4	3	3	3	5	2	2	2	2	1	5	4	4	3	1	2	2	1	6	2	3	2	2	1	2	1	3	3
BI	1	1	5	4	5	3	4	2	5	5	1	3	1	2	1	3	4	3	5	2	2	3	4	1	3	1	2	2	3	1	4	3
BJ	1	1	3	4	5	1	3	4	4	3	1	2	1	5	4	3	2	6	2	2	1	2	4	3	3	5	1	1	2	3	1	2
CA	2	1	6	4	2	3	5	2	6	5	2	3	1	5	3	5	5	8	3	1	2	2	4	5	7	2	2	2	3	1	1	8
CB	2	2	10	4	4	7	4	3	4	4	2	2	2	1	6	4	4	5	3	3	3	2	4	2	5	7	2	3	4	2	4	3
CC	1	3	4	5	7	3	4	4	6	5	3	3	2	4	2	1	6	2	6	2	2	9	5	4	2	6	3	2	5	1	2	1
CD	2	1	7	4	6	1	5	4	3	4	2	2	1	2	3	3	4	3	5	4	2	2	7	2	3	3	2	2	3	2	1	4
CE	3		5		5		3		4		2		3		2		6		5		2		7		2		2		4		4	
CF	3	3	4	5	6	5	4	1	3	5	4	1	2	5	3	6	6	5	3	2	2	2	6	2	4	4	2	3	4	3	2	4
CG	1	2	4	4	5	4	4	4	2	5	2	2	2	6	3	5	5	3	3	3	5	2	5	5	3	2	3	4	2	3	1	5
CH	3	2	3	5	7	5	6	4	3	5	2	4	2	6	3	4	7	6	4	2	2	3	6	2	5	1	3	4	4	1	3	5
CI	2		8		3		4		6		3		2		5		4		7		5		6		6		2		3		2	
CJ	2		7		5		2		3		3		2		7		5		4		6		6		6		2		4		1	
DA	5		4		7		3		5		3		2		4		4		6		2		6		6		3		6		4	
DB	3	1	6	4	7	6	4	5	6	5	4	3	2	3	6	6	7	3	3	4	2	3	6	5	6	5	3	3	6	5	1	1
DC	3		4		4		2		5		4		2		3		4		6		2		8		7		3		5		3	
DD	3		7		5		6		5		3		1		5		7		4		5		6		8		2		6		1	
DE	6	3	5	5	7	7	7	2	6	3	4	3	2	6	6	5	8	6	6	6	3	3	7	7	6	7	4	4	6	6	1	2
DF	6		5		8		6		6		3		2		7		6		3		5		9		5		3		6		3	
DG	2		7		5		6		7		2		2		8		8		6		3		8		7		4		5		1	
DH	6		8		7		4		6		2		4		7		6		7		3		5		7		4		6		3	
DI	7		5		7		5		7		5		1		6		4		4		7		7		5		3		7		3	
DJ	4		5		8		7		4		4		7		4		9		3		3		4		3		3		7		3	
EA	6		8		8		7		4		3		1		5		4		7		5		7		8		4		7		6	
EB	4		6		6		7		7		4		1		6		6		6		8		10		7		3		8		5	
EC	7		6		8		6		7		4		5		5		8		6		6		8		5		4		6		1	
ED	7		8		7		9		7		4		1		7		8		7		6		8		5		4		7		3	
EE	6	3	8	5	8	7	6	8	8	7	3	4	3	6	7	5	8	8	6	7	6	7	8	6	6	8	4	5	6	10	4	6
EF	7		7		9		10		8		3		3		7		7		5		5		7		6		4		9		5	
EG	7		7		8		8		6		5		4		8		8		8		5		9		8		4		8		7	
EH	8	4	7	5	8	8	8	8	8	7	3	3	4	9	8	5	6	8	7	5	6	4	8	8	9	8	4	6	9	11	6	5
EI	6	5	8	7	8	7	10	7	8	6	5	4	2	7	8	9	9	6	7	8	6	6	9	6	8	10	5	7	9	8	6	6
EJ	9		8		8		7		6		4		5		9		8		7		7		9		8		6		9		6	
FA	9	7	9	5	8	9	4	8	7	8	4	5	10	8	7	8	8	9	7	8	7	6	10	7	7	6	7	5	8	9	6	8
FB	7	9	7	6	7	8	9	9	7	6	8	4	1	10	8	8	9	9	8	5	8	7	8	6	9	9	7	6	9	11	5	4
FC	9	6	7	8	9	7	9	6	6	8	6	4	3	9	9	8	9	9	6	4	7	7	9	9	7	9	6	7	7	7	4	6
FD	8		8		9		10		9		4		5		7		9		8		7		9		9		7		9		6	
FE	11	4	9	5	8	9	8	11	7	9	8	5	4	10	9	9	9	9	8	7	6	6	9	7	9	6	8	6	7	8	6	7
FF	6	8	8	6	9	8	7	8	9	8	3	4	6	11	9	7	9	10	8	1	4	9	9	3	6	8	8	6	7	10	6	6
FG	7	7	8	8	8	7	8	7	7	6	7	5	3	10	9	7	9	8	7	9	8	5	9	9	8	8	7	7	8	9	9	5
FH	10	4	9	8	8	9	7	8	9	8	7	3	5	7	7	9	9	9	8	7	7	8	9	7	9	10	5	6	9	10	6	4
FI	9	4	7	6	9	6	9	10	8	7	6	6	3	10	10	7	9	9	8	8	7	8	8	4	9	9	8	8	9	10	8	6
FJ	6	5	9	7	9	9	11	9	10	7	7	7	2	10	6	6	9	11	6	9	5	8	10	8	8	10	9	8	9	10	9	6
GA	10	9	4	6	8	8	9	9	7	9	7	5	2	10	9	9	9	6	8	9	8	10	6	10	9	9	7	9	10	10	6	6
GB	10	5	8	8	10	8	8	10	10	8	5	6	9	10	7	8	7	8	8	8	9	8	9	7	9	9	9	8	9	10	6	5
GC	10	9	10	6	10	9	10	10	11	10	7	6	6	11	10	6	10	10	8	8	8	9	9	8	9	10	5	9	10	10	8	10
GD	7	9	7	8	10	8	11	9	8	8	9	7	2	10	10	9	10	9	8	9	8	7	10	10	10	9	8	9	7	10	7	6
GE	9	6	10	8	9	9	11	9	10	8	7	5	3	11	8	9	10	7	9	9	9	10	10	9	11	8	5	8	10	10	9	5
GF	9	7	10	7	9	9	11	10	10	7	8	5	9	10	10	9	9	10	7	8	9	9	10	10	10	11	9	9	10	11	9	7
GG	10	9	8	9	9	9	11	9	9	9	7	5	11	10	11	10	9	9	9	8	10	9	9	11	9	9	9	10	9	10	9	10
GH	10	10	11	8	11	9	9	10	10	10	8	5	11	11	11	9	10	10	9	10	7	8	10	11	10	11	10	9	10	10	9	7
GI	10	7	11	11	11	8	10	10	9	11	8	6	10	11	11	10	11	11	8	9	8	8	11	10	10	11	10	10	11	11	8	8
GJ	11	9	10	8	11	9	11	11	10	10	6	8	11	10	9	10	10	11	9	10	8	9	10	11	10	11	10	10	10	10	9	7
HA	10	10	10	8	10	9	10	10	10	8	9	6	11	10	9	9	11	9	8	9	8	10	8	11	11	10	11	10	11	11	7	8
HB	10	9	10	9	10	8	11	11	10	9	9	8	11	11	11	10	11	10	8	9	10	9	11	10	10	11	10	10	11	11	9	9
HC	11	10	11	8	10	9	10	11	10	11	8	8	11	11	10	8	10	11	9	9	11	8	10	9	11	10	10	10	11	11	8	10
HD	10	10	11	11	11	9	10	11	11	11	7	8	10	11	11	10	10	10	8	10	8	11	11	10	11	10	11	11	10	11	9	6

1	14	10	0	0	6	4	6	1	5	0	17	4	24	4	4	2	3	2	6	6	15	5	1	2	5	5	9	5	6	9	23	6
2	13	5	2	0	7	0	9	7	6	4	11	7	19	4	9	3	10	4	6	9	13	9	6	7	7	5	19	5	11	5	4	4
3	7	5	11	3	3	7	8	3	8	5	11	9	7	3	11	4	4	7	14	2	6	4	6	5	12	3	10	6	7	3	9	3
4	2	4	12	10	4	1	9	7	8	3	11	6	4	2	5	5	10	0	4	3	1	1	7	3	2	1	10	3	5	0	8	4
5	1	3	6	10	9	3	3	1	5	7	4	8	4	3	5	6	4	2	4	2	8	1	5	3	6	2	4	2	3	1	3	6
6	8	2	4	5	3	0	6	1	10	3	5	2	4	6	3	6	3	10	1	7	3	7	3	8	3	2	5	8	1	11	9	
7	8	4	11	3	8	6	7	2	10	5	8	2	1	2	10	2	5	0	10	3	8	4	6	5	7	2	5	3	7	0	3	5
8	2	1	12	10	14	9	5	5	6	8	6	4	0	1	6	5	8	7	15	6	10	8	11	3	7	5	4	5	4	2	4	4
9	6	7	5	2	10	13	6	6	5	4	3	0	2	2	8	10	14	10	5	10	4	6	12	5	10	6	4	5	12	3	9	1
10	10	4	7	0	6	0	8	7	9	3	0	0	3	12	5	5	7	6	0	3	1	3	10	6	6	7	5	5	7	13	0	3
11	3	0	4	2	4	0	7	5	2	3	0	0	6	8	5	0	3	4	0	0	1	1	3	3	4	6	2	1	4	8	0	0

Table 6

Raw results of experiment C.

Also proportional unit scale separation of constituent pairs of each 'triplet'.


SUBJECTS
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
ITEM SCALE SEPARATION

EXPERIMENT D

RESULTS

TABLE 7

Raw Results

Subject	1	2	3	4	5	6	7
A	9 250.04 - -	5 118.5 - -	3 46.5 - 0,0,0,0	12 459.45 - -	7 199.62 - 0,2,6,8	5 97.13 - 0,0,0,0	9 222.85 - 0,2,2,2
B	5 59.68 - -	9 94.98 - -	5 47.65 - -	7 69.66 - 0,0,2,4	9 88.4 - 6,6,8,12	3 20.28 - 0,0,0,2	12 153.58 - 7,10,10,12
C	9 150.84 - -	5 - - -	9 144.1 37 2,2,4,8	5 55.19 22 0,0,2,2	12 202.69 47 2,4,14,16	3 28.44 13.5 0,0,0,0	7 86.62 30 0,0,2,6
D	5 74.44 - -	9 155.94 - -	12 190.93 39 8,12,16,24	9 92.88 29 2,4,10,10	3 22.64 11 0,0,2,2	5 39.87 15.5 0,0,2,4	7 59.65 22 0,2,4,6
E	9 193.8 - -	5 83.76 - -	7 116.54 27 0,0,0,2	3 60.75 11.5 0,0,0,0	9 206.7 33 0,0,2,2	6 66.85 19 0,0,2,2	12 28.44 42.5 2,2,8,10
F	5 75.04 - -	9 145.85 - -	7 95.32 27 0,0,2,4	5 58.53 20 0,0,2,2	9 145.44 33 2,2,14,16	12 210.66 46 2,4,6,8	3 25.1 12 0,0,0,0
G	9 132.61 - -	5 39.34 - -	9 87.73 40 0,2,4,6	7 49.41 20 2,2,6,10	12 145.22 39 4,4,16,22	3 21.82 10 0,0,0,0	5 33.31 15 0,0,0,2
H	5 59.43 - -	9 129.55 - -	5 76.95 25 0,0,0,0	12 187.9 65 2,4,8,14	7 87.03 30 0,2,4,8	9 138.81 48 2,2,2,9	3 27.67 12 0,0,0,2
I	146.51 43 - -	5 69.9 - -	3 26.84 12 0,0,0,2	9 208.31 42 2,4,4,12	7 114.74 27 0,0,4,6	5 60.43 24 0,0,2,2	12 466.8 49 4,4,10,12
J	5 108.12 - -	9 369.68 - -	12 373 39 2,8,10,15	5 114.03 19 0,0,0,6	3 46.64 13 0,0,0,0	9 243.03 36 0,0,0,4	7 214.82 26 2,2,2,6
	training trials				test trials		

In each cell is recorded

a) N number of cards per pack (sorting mode)
b) T (time for three cycles) in secs.
c) t (total pickup & return time) in secs.
d) Σd^2 (4 results per set).