Testing a 'Trilemma' Instrument for Vocational-interest Assessment

David Bimler & John Kirkland
Massey University

How an individual ranks items by preference, or on any other form of subjective scale, can be assessed efficiently by presenting the items three at a time for three-way, forced-ranking decisions. We propose that these choices are most informative if each selection of items is guided by their empirical locations within a multi-dimensional model. We designed a questionnaire on this principle to elicit vocational preferences from secondary-school students. Ninety-nine items, each specifying a vocational interest or activity, appear in 66 "trilemmas", so each item is ranked twice. Because the choices are not constrained by a pre-determined theory of occupational cognition, the data are not restricted to a single form of analysis, though here they are interpreted using multivariate regression within the multi-dimensional model. In a first stage of validating the questionnaire, responses were simulated on the basis of existing data. The scores accruing to each item from each participant's simulated trilemma responses were compared against actual item rankings. In a second stage, the questionnaire was administered to 299 students, who found it easy to complete. The underlying structure of the responses matched the results of earlier research. The results for individual participants were meaningful, and replicated data from 17 students who had previously ranked the same items with a more conventional task.

The process of constructing a Likert scale is a standard part of a university psychology curriculum. Psychometric tests of one form or another are ubiquitous. But when a self-report questionnaire asks untrained participants to assign numerical values to each of a list of statements (e.g. 1 to 5 on a scale of 'strongly disagree' to 'strongly agree'), the task does not come naturally. Response-style problems such as amenability bias and halo effects may affect the data, even when the task is stripped of numerical nuances and reduced to the simplicity of binary choices (i.e. when the questionnaire is a checklist of items to be endorsed or rejected).

Pairwise comparisons (forced binary choices) are one procedure adopted as a remedy for these problems; participants consider every pair of items, and must choose one, rejecting the other (e.g. Bechtel, 1976). Summing the number of 'acceptance' decisions for a given item, over all such pairs, gives it a numerical value on a fine-grained scale. However, the number of choices grows rapidly as the number of items increases: for 99 items there are 4851 pairs. Thus Schucker (1959) suggested presenting the items three at a time to be ranked for preference, so that in each triad one is accepted, one is rejected, and the third is in the middle. This is equivalent to eliciting three pairwise comparisons at once. All possible comparisons among 99 items are covered by 1617 of these three-way forced-ranking choices.

This is a *complete* design. Each item would appear 49 times, being ranked against different alternatives each time. A logical extension of this line of thought is to reduce the number of three-way forced-ranking choices by using *balanced incomplete* designs, in which each item occurs in (for instance) two of these 'trilemmas', in a different context each time. Such a design would provide scores for an item ranging from -2 (if the participant rejects it on both occurrences) up to +2 (if it is accepted both times). The procedure implies that the score for a given item is only meaningful relative to other items from the same participant. There are no absolute scales, on which a participant's score for the item (or a combination of items) can be compared to scores from other participants. The data are *ipsative*, in other words; we return in the Discussion to the questions raised by this.

Elsewhere we have argued that such choices should be optimised to make each one as stark and as informative as possible (Bimler et al., 2005). A trilemma is non-optimal if raters' preferences for two of the items in it are highly correlated. The trilemma is forcing a choice between two items of very similar content. But a highly negative correlation between items is not desirable either: a correlation of -1 implies that those two items are diametrically opposite, so rejection of one follows automatically from a high preference for the other, simplifying the three-way choice but reducing its information content. Each item should appear in selected contexts where it is not judged against other items that are either very similar in meaning or content, or diametrically opposite. In this case the optimisation makes use of a multidimensional scaling solution (MDS) for the set of items.

A solution from MDS is a geometrical model, in which items are represented by points within a low-dimensional space (Jones & Koehly, 1993). In the case of only two dimensions, the model is simply a map. The points are located so that distances between them reflect as accurately as possible the dissimilarities between the corresponding items. The dissimilarities can be obtained empirically from a corpus of preference values from a substantial number of participants, from which it is easy to calculate the covariance between the values assigned by each person to any given pair of items. Alternatively, participants may have directly provided subjective judgements of inter-item similarity. The result is the same; the three items comprising a trilemma are represented by points marking out a triangle within this multidimensional map of 'item space'. We propose that such a trilemma is optimal when this triangle is as large and close to equilateral as possible. This minimises the correlations among ratings or rankings assigned to those three items, due to the nature of the map's construction.

Validity testing

The objective of this report is to test the validity of a trilemma questionnaire. Several stages of testing are involved. We begin by assuming ideal participants who are conscious of a preference value for each of the items, and who work through the questionnaire reliably, using those values to respond to each trilemma without error. We further assume that these preference values do not fluctuate in the process, and that they are accessible to us as a basis for evaluating the questionnaire's outcomes. The question is whether the questionnaire is capable of capturing those known values without distortions or artefacts.

If the first phase of testing is a success, the next stage is to analyse real data, collected with a standard preference-rating procedure as well as the questionnaire. The question is whether the two sets of data possess equivalent internal structures; that is, whether the scores are distributed in comparable ways, so that comparable sets of factors can account for them.

The final test is more stringent and involves participants who followed both procedures. We expect a reasonable correlation between the direct preferences for the items as expressed by a given participant, and the scores that items accrue from that participant's trilemma rankings.

Vocational-interest assessment

Our secondary objective is to develop a vocational-assessment instrument which vocational counsellors can use to elicit a job-seeker's preferences among job-related activities (but also suitable for self-administration), and to expand any specified occupational niche into its associated pattern of activities. Our item set, the VOC-99, consisted of 99 vocational activities ("verbs") and objects of activity ("nouns"). Secondary-school students used the trilemma procedure to rate the items according to personal preference.

Many instruments for vocational assessment are already available, with well-documented properties. Most are checklists (non-ipsative), where interests can be Liked or Disliked (or Appealing or Unappealing). Others share the ipsative quality of the trilemma procedure: for instance, the Kuder Career Search (Kuder, 1977) offers 180 interests, presented in 60 forced-choice triads. The Jackson Vocational Interest Survey consists of 289 pairs of work-related activities; Jackson (1977) used the forced-choice format to eliminate response biases, arguing that these otherwise accounted for as much as 35% of response variance. The Career Interest Test (Athanasou, 2002) is another example of pair-wise choices.

But in the majority of instruments, the selection of items and the system for summarising responses were driven by pre-existing theories of occupational cognition. For instance, the Unisex American College Test or UNIACT (Swaney, 1995) embodies a categorization of jobs and vocational personalities into Realistic, Investigative, Artistic, Social, Enterprising and Conventional types. Thus one motivation for the VOC-99 was to create an open-ended (and non-proprietary) research tool with no prior expectations and few restrictions about the kind of information it would capture (explained at more length in Bimler & Kirkland, 2005). The generation of items was an iterative process, beginning as a pilot study in collaboration with Career Services New Zealand. Early versions of the item list were 'mapped' with MDS, and refined by eliminating nearly-synonymous items while generating new ones where the MDS solution displayed gaps. It is possible that the authors' speculations about the appropriate dimensions for early versions became entrenched, despite our intentions.

Previous analysis of the VOC-99 set found that the distinctions made between items and the dissimilarities that informants perceived among them could be explained by locating them along three major dimensions or continua: 'People / things', 'Physical / cerebral', and 'creative / routine' (Bimler & Kirkland, 2005). A fourth minor axis, 'self-reliance / external', also reliably emerged. Here it is convenient to focus on the three major dimensions, though the fourth may become the subject of future study. For comparison, Day and Rounds (1998) applied MDS to the 90 items comprising the UNIACT, and reported a three-dimensional solution. Their third axis was uninterpretable; the first two are subsumed by the first and third axis of the present scheme. Note that this frame of reference is not absolute – since a MDS solution can be rotated to new axes – and is simply a tool for interpretation. The dimensions are not used as criteria for selecting items into combinations or scales. An item is never reduced to its coordinate on a single axis; its full description requires all three coordinates.

Table 1 lists examples of the items, and their coordinates along the three major axes.

Once responses have been collected for an exhaustive set of items, it is helpful to have some way of summarising them as a manageable number of indices, since a list of undigested scores for each item is likely to convey little meaning. The present discussion of the VOC-99 item set uses three summary indices, one for each of the first three dimensions of the MDS solution. The result is a score or location for each participant along three orthogonal, bipolar scales ('People / things', 'Physical / cerebral', and 'Creative / routine'). By treating these as coordinates, individual participants can be represented, like items, as points within the 'vocational interest space' spanned by these three axes. This scheme is not set down as the only one for rendering down VOC-99 data, nor is it necessarily the best, but it is convenient for present purposes. It provides one method of comparing consistency across alternative procedures for data collection.

Table 1. Examples of items from the VOC-99 set, and their coordinates in the MDS solution, from the extremes of each axis.

x_{i1}	x_{i2}	x_{i3}	Text of the item
6.72	0.63	0.90	Working with machines
6.60	1.15	0.48	Servicing heavy equipment
6.51	1.26	1.55	Operating big machinery
6.00	1.65	0.48	Making household repairs
5.97	1.56	-1.02	Working with small hand held tools
5.96	2.31	-0.86	Working on buildings or structures
5.84	-1.80	-1.57	Working with electronics
-5.72	-1.02	0.25	Making a positive difference to others' lives
-5.77	0.39	2.78	Working with the elderly
-5.79	0.95	2.45	Working with young children
-5.81	-1.82	1.52	Getting others "back onto their feet"
-5.84	-0.45	2.64	Working with sick or injured people
-5.95	-1.00	1.68	Caring for others
-5.99	-0.10	-0.82	Working with people (rather than with "things")
-1.33	6.32	1.14	Being a professional sports person
-0.90	6.23	-0.16	Looking for adventure
1.69	6.01	2.00	Working mostly outside
-0.79	5.92	1.28	Pushing my body to the physical limits
-0.03	5.60	-1.64	Working in different places (e.g. being "on the road")
-1.36	5.43	-0.25	Working with recreational activities
1.74	5.39	2.84	Working in natural environments
2.00	-5.20	-0.59	Using my brains as much as possible
0.58	-5.20	2.14	Becoming a member of an established firm
0.70	-5.46	-2.18	Organising information or things
-0.90	-5.58	0.49	Providing others with financial (money) advice
-1.77	-5.64	0.55	Providing others with legal advice
1.17	-5.90	-1.59	Working with documents
0.00	-5.93	-0.83	Working with private information
-0.56	-0.76	6.78	Doing as my elders/parents suggest
-0.05	0.15	6.45	Keeping to traditional ways
0.56	1.71	6.16	Continuing the family business (eg farming, fishing, selling, etc.)
-1.16	-1.66	6.11	Following known pathways, walking in others' footsteps
-2.52	-0.56	5.87	Respecting the views of older people (including parents)
2.44	-0.63	5.44	Following orders
0.54	-2.17	5.43	Being well-disciplined
-0.58	2.89	-5.40	Being creative
-2.92	1.75	-5.42	Using my voice (eg singing, speaking)
0.29	-1.99	-5.92	Working on printed text or stories
-2.30	2.44	-5.94	Performing in the entertainment industry
-0.41	2.48	-6.00	Working with visual arts or crafts
0.53	1.05	-6.00	Working with video and film
-0.70	1.01	-6.09	Creating material which can be used for entertaining others

Values of the three summary scales are obtained by multivariate regression, as outlined below in the Analysis section of Study 1. Each is essentially a weighted sum of the values for individual items, with the weights designed to extract a desired global feature of the distribution of values. They can be interpreted as components of a vector within the MDS solution (Jones & Koehly, 1993). The vector traces the direction in which item-values increase (from items on one side of the solution, to which that participant assigned the lowest values, ranging up to highest values on the other side). For instance, a participant who responded to the items primarily on the basis of their location along the 'people / things' continuum – preferring the latter quality and rejecting the former – will be represented by a vector which is nearly parallel to the first axis of vocation space. The first summary scale will be high and positive while the other two are small.

STUDY 1. Imputed data: Re-phrasing ranking responses in trilemma format

This section describes how the VOC-99 items were arranged in a 'trilemma' questionnaire, which we applied to a list of exactly-known preferences, to test how accurately they were recorded. It is not essential for the preference data in this first stage of testing to come from real situations. Hypothetical, synthesised preferences could be used for the dual task of simulating trilemma responses, and gauging their accuracy. However, these may lack realism in their distribution and internal noise. Actual data happen to be available for the current application: in the course of constructing the VOC-99, vocational interest preferences were collected from 137 students at four New Zealand secondary schools, as reported in more detail by Bimler and Kirkland (2005). Here their data were reanalysed to simulate how they would have responded to the trilemma task. Questionnaires of different lengths were simulated (66, 33 and 16 trilemmas).

Method

Participants:

137 students were recruited from four secondary schools. Ages ranged from 15 to 18. Sex was not always recorded, but numbers of males and females were broadly equal.

MOSS technique:

Students described their vocational preferences by ascribing a value to each item, in the range -2 to 2 (from least- to most-preferred). They did so by sorting items into five rank-ordered piles, ranging from most appealing, congenial or preferred items (which go into the pile at one extreme) to those least preferred (which go into the pile at the other extreme). This ranking procedure is the 'Method of Successive Sorts", a variant of Q-sorting (Block, 1961). The sorting is a two-step process. The participant begins by creating three piles containing items that are more preferred, neutral, and less preferred. Next, the first pile is subdivided into piles containing *most* and *more* preferred items, while the third pile is subdivided into piles of *less* and *least* preferred. Items can be shifted between piles if the participant reconsiders. Often a researcher using Q-sorting sets the number of items in each pile, but a forced distribution was not necessary here.

A given item, specified by an index i ($1 \leq i \leq 99$), can be labeled by the shorthand E_i. A second index q specifies the participant. Let v_{qi} represent the number of the pile to which E_i was assigned by the q-th participant, from +2 (most preferred) down to -2 (least preferred). It is convenient to treat these 99 values as the elements of an array, which we label $\mathbf{v}_q$.

Trilemmas:

A three-dimensional model of vocation space was earlier obtained by a multidimensional scaling analysis of the VOC-99 set (Bimler & Kirkland, 2005). Each E_i is located by a point in this model, with coordinates (x_{i1}, x_{i2}, x_{i3}). We can write these coordinates as entries in a 99-by-3 matrix, labelled $\mathbf{X}$ to refer to the entire 'map'. The model is a roughly spherical cloud of item-points, and it turns out that all points are about equally distant from the origin: that is, the cloud is a spherical shell, surrounding an empty void. However, the distribution of points across the surface of this shell is fairly even. This model was used to select 66 combinations of three items, with a computer program applying the "maximum-area triangle" criterion mentioned in the Introduction (see also Bimler *et al.*, 2005). Figure 1 shows examples.

Each item appeared once and once only among the first group of 33 trilemmas, then a second time within the second group of 33. To ensure that a given item is judged against different alternatives on its two appearances, a second constraint prevented a given *pair* of items from appearing together in any trilemma more than once. Subject to these constraints, each trilemma $\{E_i, E_j, E_k\}$ was selected such that the triangle marked out in $\mathbf{X}$ by the corresponding points had a greater area than other available combinations. This criterion penalises trilemmas which contain two items with relatively similar meanings, because the corresponding points are close together in the model, resulting in elongated acute triangles with restricted area. It also penalises any trilemma where two items are opposite in meaning, because the corresponding points should be at opposite extremes of the model – separated by an entire diameter of the sphere – forming an obtuse triangle, again with less-than-maximal area. It encourages trilemmas where the points are arranged in an equilateral triangle, each pair separated by 120° (seen from the centre of the spherical cloud).

Ideally each item would appear twice in the completed questionnaire. However, the number of appearances of E_i may be reduced in some cases by omitted or spoiled responses, so for generality we write n_i to label this number. As noted earlier, the procedure requires the participant to choose the most and least-preferred items from each trilemma. These receive scores of 1 and -1 respectively (with a score of 0 for the middle item). There are n_i such scores for E_i, and the final value for E_i (which we label u_{qi}) is simply the average of those scores. Together, the 99 values of u_{qi} comprise an array $\mathbf{u}_q$.

MOSS responses from the q-th participant were used to simulate how he or she would have responded to the trilemma questionnaire, according to the following procedure:

For a given trilemma in the questionnaire consisting of items E_i, E_j, E_k, the corresponding MOSS values were v_{qi}, v_{qj}, v_{qk}. Then a score of 1, -1 or 0 is assigned to E_i depending on whether v_{qi} is the largest, smallest, or middle of the three values. The same rule is applied to E_j and E_k. Special consideration is required for cases where v_{qi}, v_{qj}, v_{qk} were not all different. If $v_{qi} = v_{qj} = v_{qk}$ (all items fell in the same pile), then that trilemma's contributory scores for E_i, E_j, E_k are all zero. Alternatively, if $v_{qi} > v_{qj} = v_{qk}$, the respective scores are 1, -½, -½. If $v_{qi} < v_{qj} = v_{qk}$, the scores are -1, ½, ½. Whenever an participant skipped some items in their MOSS data, any trilemmas involving the missing items are omitted in turn, since no prediction can be made for them. The scores for E_i are summed over trilemmas and divided by n_i. Since some of the same participants later provided actual trilemma data (u_{qi}) in Study 2, here we use a different label w_{qi} to represent these imputed values.

Analysis:

For each participant, the 99 values w_{qi}, predicted as responses to the trilemma procedure, were compared against his or her MOSS responses v_{qi} on which those predictions were based. A first index of agreement between the two sets of values is the Pearson correlation r_q.

Next we summarised each participant's actual and simulated responses by representing them as a pair of three-dimensional *vectors* within vocation space, $\mathbf{b}_q$ and $\mathbf{c}_q$ respectively. The vector $\mathbf{b}_q$ with components $\{b_{q1}, b_{q2}, b_{q3}\}$ shows the direction of a gradient within the 'map' $\mathbf{X}$, running from least- to most-preferred items. Both vectors were found by multivariate regression. That is, the components of $\mathbf{b}_q$ were chosen to provide the best solution to the approximation:

$$v_{qi} \cong b_{q0} + b_{q1} x_{i1} + b_{q2} x_{i2} + b_{q3} x_{i3} \quad (1)$$

Thus the participant's actual

responses v_{qi} served as the dependent variable. There were three independent variables: the coordinates x_{i1}, x_{i2}, x_{i3} of the items on each dimension in turn. Recall that these coordinates are identified as the item's connotations on the three bipolar continua of 'People / things', 'Physical / cerebral', and 'Creative / routine'. For instance, b_{q1} describes how much the q-th participant's preference for each item is determined by its location x_{i1} along the first dimension. That is, it expresses the importance of the 'people / things' distinction to that participant. The offset b_{q0} is usually close to 0.

We are assuming that each participant's preferences will follow a spatial gradient, with a trend to prefer items on one side of the model, while rejecting items on the other side (the evidence supports this assumption: see Bimler & Kirkland, 2005). The regression calculation also yields a multivariate correlation R_q, which measures how far the q-th participant's preferences fit this assumption, and can range up to 1.

A second regression calculation uses the trilemma-imputed values w_{qi} as the dependent variable. It summarises them by a vector $\mathbf{c}_q$, with components c_{q1}, c_{q2}, c_{q3}:

$$w_{qi} \cong c_{q0} + c_{q1}\,x_{i1} + c_{q2}\,x_{i2} + c_{q3}\,x_{i3} \qquad (2)$$

The crux of all this is the angle ϕ_q between the two vectors. In standard vector algebra,

$$\phi_q = \text{arc-cos}\ \{ (\mathbf{b}_q \bullet \mathbf{c}_q) / (\|\mathbf{b}_q\|\ \|\mathbf{c}_q\|) \} \qquad (3)$$

Small values of ϕ_q indicate that the content of the actual MOSS responses are not being distorted by their translation into trilemma terms, and provide confidence that the trilemma technique is not imposing artefacts.

Results

The multivariate correlation R_q is an index of the compatibility between the locations of items within the spatial model **X**, and the q-th participant's preferences among them. The approximation (1) is far from perfect, but it is adequate: the mean value of R_q across the 137 participants was 0.43. In other words, on average the three parameters of each $\mathbf{b}_q$ capture about 20% of the variance within that array of values $\mathbf{v}_q$. The other 80% is a mixture of information which could be captured by a better model (perhaps one with more dimensions); meaningless noise; and variance arising from individual models of 'vocation space' which differ from **X**.

We can expect to find an association between ϕ_q and R_q. If the preferences $\mathbf{v}_q$ are particularly compatible with 'vocation space' (so that R_q is high), it follows that a trilemma questionnaire will capture those preferences quite faithfully as $\mathbf{w}_q$, because the trilemmas were optimised with **X** in mind; so ϕ_q will be small. Conversely, if the pattern of preferences is *not* compatible with the model, they may be distorted by the questionnaire, leading to a high ϕ_q. As expected, the correlation between R_q and ϕ_q was significant, at -0.51.

Mean values of r_q and ϕ_q are shown in Table 2. Clearly the process of transformation through the trilemma procedure has not radically altered preference ratings.

Table 2 also shows mean results for further simulations where the predicted responses used only the first 33 trilemmas in the questionnaire (so that each item only appears in a single trilemma, and contributes only once towards its final predicted value); and again for 16 trilemmas (so that more than half the items accrue no score at all, and must be treated as 'missing data'). Though the truncation to 33 trilemmas does impact on predicted responses, the distortion is not excessive. We conclude that there is a substantial degree of redundancy among the items comprising the VOC-99, making the instrument quite robust to missing data and to the possibility that specific values are only approximations of the true preference. The summarised gist is the same, however the data were obtained.

STUDY 2. Practical application

This section reports the use of the VOC-99 trilemma questionnaire in practice, rather than simulations. It was field-tested on secondary-school students. We analysed their responses in the context of the spatial model **X**, and examined how the individual summary values were distributed in comparison to the earlier data set. Some students provided data for the previous study as well as this one, allowing a direct test of the accuracy of the questionnaire.

Method

Participants:

299 secondary-school students were recruited, from six classes at five schools. In age they ranged from 15 to 19, around a mean of 16.4 years. There were 158 girls and 136 boys (five unrecorded). Seventeen of the students from one class had taken part in Study 1, in which they ranked the VOC-99 items in order of appeal, applying a five-pile MOSS procedure. Some weeks had intervened since then, so the problem of familiarity with the item set (and learned responses) was not seen as serious.

Procedure:

The 66-trilemma questionnaire generated in Study 1 was printed as an eight-page booklet. The three options comprising each trilemma were written out in full, side-by-side in three boxes, their order being randomised. These three boxes were labelled left-to-right from 1 to 3. A space was marked below these boxes, at the left, into which

Table 2. Outcome of Study 1. Mean similarity between original item responses and imputed trilemma responses, across 137 informants, for questionnaires with varying numbers of trilemmas. Similarity is measured in two ways: correlation r_q between original / imputed response arrays $\mathbf{v}_q$ and $\mathbf{w}_q$; and angle ϕ_q between vectors $\mathbf{b}_q$ and $\mathbf{c}_q$.

Trilemmas	Mean r_q	Mean ϕ_q
66	0.85	7.7°
33	0.78	9.6°
16	0.80	18.3°

participants were instructed to copy 1, 2 or 3 according to which of the three offered activities (or objects of activity) appealed to them *least*. They were further instructed to pick an activity from that trilemma which appealed to them *most*, and copy the appropriate number (1 to 3) into a space at the right. There remained a third space for the number of the neutral activity (see Figure 1).

The questionnaires were administered within a standard classroom context, during a 45-minute class period. This was overseen by each class's regular teacher, rather than the authors. Participation was voluntary. As partial compensation for the students' time and presumed concentration, they were promised some sort of feedback from their responses. To reconcile this promise with the preservation of confidentiality, students identified their booklets with a self-chosen four-letter code. Within each class, we grouped the results of the analyses into bands and made a set of summary analyses available to the class – one per band – along with a list of identifying codes belonging to that band, so that only the students knew which analysis applied to them.

Analysis:

Each trilemma was related back to the identifying numbers of the 99 items, so that the participant's most-preferred, neutral and least-preferred choices received scores of 1, 0 and -1 respectively. The participant's responses were thus converted to an array $\mathbf{u}_q$ of item values u_{qi}.

Multivariate regression was applied once again to summarise and represent each $\mathbf{u}_q$ in the geometrical model as a three-component vector $\mathbf{a}_q$. We expect these vectors to follow a similar distribution to the $\mathbf{b}_q$ vectors obtained in Study 1, if the trilemma procedure is to be plausible as a field instrument.

A direct repeated-measures comparison is possible for the 17 participants who had earlier provided MOSS data. The similarity between their responses $\mathbf{v}_q$ and $\mathbf{u}_q$ (and between the summary vectors $\mathbf{b}_q$ and $\mathbf{a}_q$) affords an estimate of the reliability available through the VOC-99. Of course this task confounds three possible sources of variability and unreliability (within the trilemma and MOSS tasks separately, as well as between the two test administrations), so a low similarity index would be inconclusive, but a high index would establish whether informants of this age can follow the trilemma task's demands and use the items informatively to delineate their vocational preferences.

A final test of trilemma reliability involves the empirical pattern of associations among items. That is, a 99-by-99 matrix of correlations was calculated, where each entry was the Pearson correlation between the scores u_{qi} and u_{qj} for each pair of items E_i and E_j, with q ranging across all 299 participants. One way to extract the latent structure from this matrix would be Factor Analysis, but here it was subjected to MDS, to remain comparable with earlier analyses of the data of Study 1. As reported elsewhere (Bimler & Kirkland, 2005), we had earlier applied MDS to the matrix of inter-item correlations within those 137 arrays of MOSS data, resulting in a solution $\mathbf{X}^M$. For the trilemma data, the structure among the items is summarised as the MDS solution $\mathbf{X}^T$, which we compare with $\mathbf{X}^M$ in several ways.

Results:

Students were generally happy with the length of the task and its level of difficulty. We received no feedback that the task was excessively onerous, or that a 45-minute class period was too short

Figure 1. Specimen trilemmas from the questionnaire used in Study 2.

1 Working with machines | 2 Performing in the entertainment industry | 3 Respecting the views of older people (including parents)

Appeals **least** ( ) ( ) ( ) Appeals **most**

1 Working with documents | 2 Continuing the family business (e.g. farming, fishing, selling, etc) | 3 Using my body in "creative arts" (e.g. acting, dance, modelling, etc.)

Appeals **least** ( ) ( ) ( ) Appeals **most**

1 Getting others "back onto their feet" | 2 Using computers | 3 Working with livestock

Appeals **least** ( ) ( ) ( ) Appeals **most**

1 Working on printed text or stories | 2 Working with the elderly | 3 Operating big machinery

Appeals **least** ( ) ( ) ( ) Appeals **most** to complete it (it was reported as taking on average about 30 minutes).

Occasionally a triad was skipped or spoiled by an invalid response (such as giving the same number for most- and least-preferred options). In light of the redundancy within the VOC-99, noted above, we decided that a participant's data set could be safely accepted if eight or fewer trilemmas ($\leq$ 12.5%) were omitted or spoiled.

Some participants reported that additional response options would have allowed them to record their preferences more reliably. Specifically, they requested a relaxation of the forced-ranking requirement, allowing them to abstain from choosing a least-preferred activity (if the two remaining after the most-preferred choice were equally unattractive); or from making a most-preferred choice; or from ranking them at all (if the three activities of a trilemma were all equally appealing or unappealing). Tied-preference alternatives have already been described, within Study 1, where they were unavoidable when imputing responses from five-point ratings. Here, however, we judged that the extra explanations required for augmented response options, and the greater complexity of data entry, outweighed the possible benefits.

The vector summaries followed the same distribution as the vectors $\mathbf{b}_q$ summarising MOSS data (Study 1). There was a significant gender difference in the a_{q1} vector components. Mean a_{q1} was 0.037 (s.d. = 0.234) across boys, and -0.286 (0.228) across girls ($p < 0.0005$). In other words, there was a weak trend for boys to be drawn towards activities with positive coordinates on the 'people / things' axis of **X**, while girls were drawn more strongly towards activities with negative coordinates. They conformed to gender stereotypes by orienting towards 'things' and 'people' respectively. The two distributions overlap, but of the 99 participants with $a_{q1} \geq 0$ (a preference for 'thing'-oriented activities), there were 81 boys and only 18 girls, plus one unrecorded. This difference is hardly surprising or novel, and was encountered with earlier versions of the VOC-99 (Bimler & Kirkland, 2005). For the MOSS data of study 1, combining boys and girls, mean b_{q1} was -0.211 (s.d. = 0.281).

No significant gender difference emerged for the other vector components. Mean a_{q2} (cerebral / physical orientation) was 0.049 (0.255) for boys and 0.067 (0.214) for girls. Thus there is a tendency to prefer physical or outdoor activities, but it is only slight: 110 participants had $a_{q2} \leq 0$, preferring cerebral or indoor activities. This is consistent with the MOSS data, where mean b_{q2} was 0.055 (0.304). Mean a_{q3} (creative / routine orientation) was -0.058 (0.198) for boys and -0.100 (0.199) for girls. The tendency is to prefer creative activities, though 106 participants had $a_{q3} > 0$, preferring routines. The corresponding average from the MOSS data was mean b_{q3} = -0.076 (0.234).

The mean of the multivariate correlations R_q (i.e. the mean compatibility between each participant's vocational preferences and the MDS solution **X**) was 0.52. This is acceptable; indeed it is higher than the mean R_q for MOSS data. Not too much can be read from this, since the criteria for trilemma selection make it difficult for them to collect data which are <u>not</u> compatible with the MDS solution. Specifically, if two items are sufficiently dissimilar to be included as elements of a trilemma (i.e. sufficiently distant in **X**), then they cannot receive identical values. This rules out the kind of data incompatibility where $u_{qi} = u_{qj}$ even though items E_i and E_j are distant in **X**.

Turning now to the 17 two-time respondents, the correlations r_q between their arrays of MOSS and Trilemma responses $\mathbf{u}_q$ and $\mathbf{v}_q$ have a mean of 0.63. This level of replicability is heartening. Between each pair of summary vectors $\mathbf{b}_q$ and $\mathbf{a}_q$, the angle ϕ_q (provided by equation 3) had a mean value of 14.7°.

Finally we consider the correlations <u>within</u> the trilemma data. We applied MDS to the matrix of inter-item correlations. A solution with three dimensions was chosen, for consistency with the three-dimensional solution $\mathbf{X}^M$ obtained earlier for the MOSS data (Bimler & Kirkland, 2005). We label the present solution as $\mathbf{X}^T$. According to the badness-of-fit measure ($Stress_1$ = 0.136), $\mathbf{X}^T$ accommodated the correlation matrix with relatively little distortion.

The similarity between two MDS solutions can be measured in a number of ways. One index is the Procrustes distance, g_l. This involves superimposing the two solutions, rotating and rescaling them to minimise the total distance between corresponding pairs of points. Here g_l = 0.068 (g_l would be 0 if the solutions were identical). A second measure is the correlation between corresponding pairs of inter-item distances: r_{TM} = 0.71. Both values indicate a high degree of congruence, in our experience, but they are not amenable to tests of significance.

Significance testing requires Canonical Correlation or CANCORR (e.g. Johnson, 1995). CANCORR compares two sets of coordinates by extracting a linear combination from each, such that the correlation between them R_c is maximal. It goes on to extract further pairs of linear combinations (each new combination being orthogonal to those previously extracted from its respective coordinate set), providing correlations R_2, R_3. The three canonical correlates between $\mathbf{X}^M$ and $\mathbf{X}^T$ were R_c = 0.92, R_2 = 0.86, R_3 = 0.82. According to a χ^2 test of Wilk's Λ statistic, all are significant at $p < 0.0001$. In other words, three pairs of mutually-recognizable linear combinations can be extracted from the coordinate sets, or to put it more loosely, the two solutions can be rotated so that each of the three dimensions from either solution has a recognizable counterpart in the other.

Discussion

As noted, trilemma data (like Q-sorting) are ipsative, so that any increase in the score given to one item requires a decrease in the score for some other item. It has been argued that this forces the participants to make more informative choices, and facilitates comparisons between them, by eliminating the problem of 'response biases' (these include the possibility that a participant may tend to check many interests as Appealing, or just as indiscriminately tend to reject them). But at the same time, the constraints of ipsativity reduce the correlations among items, which affect assessments of the reliability of scales derived from such data. Thus Bartram (1996) has argued that response biases should be shown to be a genuine problem in a given situation, before choosing an ipsative over than a normative (non-ipsative) instrument.

Because of those reduced correlations, the validity of applying Factor Analysis (FA) to ipsative data has been a particular focus of debate (e.g. Closs, 1996). Cattell and Brennan (1994) argued that any general factor found in normative data is likely to vanish if data are ipsatised (since it captures the response biases), but factor interpretability is otherwise robust. Similarly, Baron (1996) surveyed the literature and concluded that ipsativity had less impact on the robustness of scales than had generally been feared. The debate has largely featured theoretical issues, or simulated data (Saville & Willson, 1991), rather than empirical cases. In this study, where MDS was applied rather than FA, the concern is not so pressing.

The use of an ipsative instrument certainly demands some awareness of these issues. Meade (2004) pointed out that as well as the general interdependence inherent to all form of ipsative data, a forced-choice questionnaire introduces interdependence between the pairs of items presented together in each multiple-choice question. But having advised caution in how forced-choice data are interpreted, Meade (p. 549) concedes that "measures of vocational interest seem ideally suited for ipsative measurement as the intent with these measures is to determine which career field may be most appropriate for each given respondent (rather than making predictions across respondents)."

Conclusion

We applied the trilemma technique to the specific goal of accessing the domain of vocational aspirations and interests. A range of instruments for this task already exist, but the majority of these were designed as expressions of some existing model of occupational cognition. Typically each item is designed to instantiate one of six, eight, ten or more types. Such typologies determine the interests presented in each choice (when forced-choice questions are involved) and how item-scores are combined into scales to measure each type. In contrast, with the VOC-99, responses are analysed according to the empirical structure of relationships among the items. To the extent that this structure matches that of an existing instrument, the support for the latter is completely independent. Thus in one application of the VOC-99 we are 'mapping' a similar (but proprietary) item set, for a rigorous comparison.

The VOC-99 is primarily a research tool. Other on-going projects involve the use of the trilemmas to elicit descriptions of particular occupations: the interests required to fit those vocational niches. The data can come from existing members of the profession, as self-descriptions, or from naïve informants relaying a 'received opinion' (stereotype). The outcome is to specify the occupation along the same dimensions as the items (and to locate it as a point in the same geometrical model). One test of the usefulness of the VOC-99 will be whether those descriptions are meaningful.

In the absence of direct access to a participant's views (i.e. the true position of each item along his or her subjective scale), it is impossible to establish whether a trilemma questionnaire captures those actual views more accurately than ranking or rating instruments (any more than ranking and rating can be compared directly). Only indirect evidence can be mustered in support of the accuracy of trilemma data. Study 1 at least ruled out the possibility that the procedure of three-way forced-ranking choices necessarily distorts the data it is designed to capture. Further, the results of Study 2 show that if trilemma data do contain artefacts, then they are equally present within MOSS ranking data. Seventeen students who were presented with the items on two separate occasions, following these quite different procedures, each time provided very similar patterns of preference.

Just as important as this specific application is the general principle of optimising the items for forced choices. The approach lends itself to the wider situation of asking an informant about the relative positions of items along a subjective dimension or scale. The general question is whether one item dominates or is dominated by another, and the broad term 'dominance data' is often used (e.g. Nishisato, 1978); specific cases include personal preferences and descriptive accuracy of statements.

These inter-item relationships can be inferred from rating procedures in which the participant ascribes a numerical value to each item (ranging from 1 to 5, perhaps, or -3 to 3). The effect is to impose a metric scale on the subjective dimension. Alternatively, relationships can be obtained explicitly, by comparing the items with ranking procedures (such as Q-sorting) where the participant sorts them into a ranked sequence of groups. Some early studies argue for the superior psychometric properties of ranking data (Block, 1956). However, Q-sorting procedures are not amenable to unassisted unsupervised administration. The physical act of sorting into groups requires desk-space, while items must be printed as physical tokens which can be lost. In contrast, rating procedures involve only a pencil and a booklet of bipolar scales. Ranking can be performed with a computer interface, if the screen is sufficiently high resolution, but 'virtual sorting' does not remove one further disadvantage of the task of ranking items by sorting, which is the dauntingly detailed set of instructions.

The trilemmas described in this report seem to combine the psychometric advantages of ranking with the physical convenience of rating. Their self-explanatory pencil-and-paper nature is suitable for self-administration at the participant's own time and convenience. A questionnaire based on the trilemma approach was administered easily and completed rapidly, with few errors. Of course this is equally true of other forced-choice instruments, but we argue that the principle used here to design the choices is particularly suitable in situations where the items are not clearly clustered (with items in each cluster being interchangeable instances of an underlying 'type' or factor). Thus the general technique has many potential applications.

References

Athanasou, J.A. (2002). *A brief, free and standardised assessment of interests for use in educational and vocational guidance – career interest test (Version 3.0).* Occasional Paper 14, University of Technology, Sydney.

Baron, H. (1996). Strengths and limitations of ipsative measurement. *Journal of Occupational & Organizational Psychology*, *69*, 49-56.

Bechtel, G.G. (1976) *Multidimensional preference scaling* (Methods and Models in the Social Sciences 6). The Hague: Mouton & Co.

Bertram, D. (1996). The relationship between ipsatized and normative measures of personality. *Journal of Occupational & Organizational Psychology*, *69*, 25-39.

Bimler, D.L., Kirkland, J., Yuhara, N., Kurosaki, M. and Coxhead, E. (2005). "Trilemmas": Characterising the Japanese concept of "amae" with a three-way forced-ranking technique. *Quality & Quantity*, *39*, 779-800.

Bimler, D.L. and Kirkland, J. (2005). From circumplex to sphere: Perceptions of vocational activities, explored and applied. *Human Resources Psychology*, *3*, 29-40.

Block, J. (1956). A comparison of the forced and unforced Q-sorting procedures. *Educational & Psychological Measurement*, *16*, 481-493.

Block, J. (1961). *The Q-sort method in personality assessment and psychological research*. Springfield, IL: Charles C Thomas (reprinted 1978, Palo Alto, CA: Consulting Psychologists Press).

Cattell, R.B. and Brennan, J. (1994). Finiding personality structure when ipsative measurements are the unavoidable basis of the variables. *American Journal of Psychology*, *107*, 261-274.

Closs, S.J. (1996). On the factoring and interpretation of ipsative data. *Journal of Occupational & Organizational Psychology*, *69*, 41-47.

Day, S.X and Rounds, J. (1998). Universality of vocational interest structure among racial and ethnic minorities. *American Psychologist*, *53*, 728-738.

Jackson, D.N. (1977). *Jackson Vocational Interest Survey*. Research Psychologists Press, Ontario.

Johnson, L. (1995). A multidimensional analysis of the vocational aspirations of college students. *Measurement & Evaluation in Counseling & Development*, *28*, 25-44.

Jones, L.E. and Koehly, L.M. (1993). Multidimensional scaling. In G. Keren & C. Lewis (eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 95-163). Hillsdale, NJ: Erlbaum.

Kuder, F. (1977). *Activity interests and occupational choice*. Science Research Associates, Chicago.

Meade, A.W. (2004). Psychometric problems and issues involved with creating and using ipsative measures for selection. *Journal of Occupational & Organizational Psychology*, *77*, 531-552.

Nishisato, S. (1978). Optimal scaling of paired comparison and rank order data: An alternative to Guttmann's formulation. *Psychometrika*, *43*, 263-271.

Saville, P. & Willson, E. (1991). The reliability and validity of normative and ipsative approaches in the measurement of psychology. *Journal of Occupational & Organizational Psychology*, *64*, 219-238.

Address for correspondence:
David Bimler
Health and Human Development
Massey University
Private Bag 11-222
Palmerston North
New Zealand
Ph. 0064-4-3800151
email: d.bimler@massey.ac.nz