Initial development of a Physics Goal Orientation survey using factor analysis

Christine Lindstrøm and **Manjula D. Sharma**, School of Physics, The University of Sydney, Australia
clind@physics.usyd.edu.au sharma@physics.usyd.edu.au

Abstract: *This paper presents the first stage in the development of a Physics Goal Orientation survey - a survey identifying students' beliefs about how to be successful in physics studies. The analysis method used is exploratory factor analysis, a powerful statistical method requiring subjective decision making. Instead of taking a 'black box' approach, which can easily lead researchers to draw incorrect conclusions, we have provided the mathematical basis for principal components analysis, the most common type of exploratory factor analysis.*

Introduction

Goal orientation theory forms part of the motivation literature, and is perhaps the most prominent theory today (Urdan, Kneisel and Mason 1999). It focuses on students' *reasons* for engaging in academic tasks, as these affect important educational outcomes such as types of cognitive strategies used, and how well newly learnt material is retained (Anderman, Austin and Johnson 2002). Studies of high school students' motivation in the general settings of 'classroom' and 'sports' have identified four different goal orientations, each associated with a certain belief in how success is achieved (Duda and Nicholls 1992; Skaalvik 1997). Task orientation is associated with the belief that success is a product of effort, understanding and collaboration. Ego orientation describes the belief that success relies on greater ability and attempting to outperform others. Cooperation oriented students value interaction with their peers in the learning process; and lastly, work avoidance describes the goal of minimum effort – maximum gain. A similar study in physics, however, has not been found, so the first aim of the paper is to develop a Physics Goal Orientation survey.

Factor analysis has become an increasingly popular statistical method over the past few decades, primarily due to the ease of use with statistical packages such as the *Statistical Package for the Social Sciences (SPSS)*. Whereas the availability of such analysis has the potential to improve work in science education, it is a double edged sword if a solid understanding of the underlying statistics does not accompany its use, as shown by Preacher and MacCallum (2003). Unfortunately, however, the literature on factor analysis is seemingly divided into the thoroughly mathematical and the purely practical. Therefore, the second aim of this paper is to provide adequate mathematical insight to support decision making in the process of using the most common statistical approach to exploratory factor analysis, principal components analysis. The mathematics requires familiarity with vectors or linear algebra.

Research method

In developing a new survey, statements are written or adapted from previous surveys and accompanied by a Likert scale. Each underlying construct has statements, each measuring a different aspect of the construct. Some statements will need to be removed, and a minimum of four statements must be retained for each factor. The requirement on sample size is not clear. In general, the conceptual basis of the statements (theory driven) and results from factor analysis (data driven) are useful guides.

In 2006, 125 first year physics students at The University of Sydney completed the Physics Goal Orientation survey. For each of the 20 statements students responded on a 5-point Likert scale ranging from strongly disagree (1) to strongly agree (5). All statements were adapted from Duda and

Nicholls' (1992) surveys to suit tertiary physics education (see Table 1).

Table 1. Statements on the Physics Goals Orientations Survey

I feel really successful when…

Item 1	I know more physics than other people
Item 2	what I learn in physics makes sense
Item 3	the other students in my tutorial group and I manage to solve a tutorial problem together
Item 4	I don't have to try hard to do well in physics
Item 5	I get a high exam mark
Item 6	I solve a problem by working hard
Item 7	I do my very best
Item 8	I work in a group on physics problems
Item 9	I can complete an assignment without really having understood the answers
Item 10	others get physics problems wrong and I don't
Item 11	I can answer more physics questions than other students
Item 12	a group of us help each other
Item 13	I learn something interesting
Item 14	I can copy an assignment off somebody else
Item 15	I am in a group and we help each other figure something in physics out
Item 16	others know more than me so they can answer the questions
Item 17	something I learn makes me want to find out more
Item 18	I do better than others in physics
Item 19	I have somebody else to discuss physics problems with
Item 20	I know I can pass the exam without studying too hard

Theory of factor analysis

Factor analysis is a data reduction method, allowing a reduction in the number of variables in a data set, while retaining a large fraction of the information. In science education factor analysis is commonly used with surveys that measure some psychometric construct, which cannot be measured directly (such as self-efficacy or students' study strategies). Respondents indicate on a Likert scale their level of agreement with several statements that focus on different aspects of the construct. Factor analysis is then used to evaluate whether the statements indeed measure aspects of the same underlying construct, and finally give each individual respondent to the survey an overall score on the construct.

Two different types of factor analysis exist. Exploratory factor analysis is used to identify underlying structure in the data. Confirmatory factor analysis is used in hypothesis testing, and is the only method for confirming whether modeled factor structures are compatible with the data. Only exploratory factor analysis is discussed in this paper. Please note that normally distributed variables are only required if the data are used to generalise findings (Field 2000). The novice user will find Field (2000) helpful, whereas Gorsuch (1983) and Floyd and Widaman (1995) provide fine detail. The brief discussion below bridges the gap.

The correlation matrix

The basis of factor analysis is that people show a pattern in their responses to groups of statements or variables. From Table 1, respondents would be expected to indicate a similar level of agreement with Items 1 and 18. A scatter plot of responses should therefore produce a strong, linear correlation. The Pearson's r correlation coefficients between each pair of variables are presented in the Correlation matrix or R-matrix in the *SPSS* output of a factor analysis; a $k \times k$ matrix for k variables. All further analysis of the data is based on this matrix; individual responses are no longer considered. However, before the analysis can proceed, several assumptions on the Correlation matrix must be met.

Firstly, no two variables must correlate too strongly. Since the purpose of a factor analysis is to identify underlying concepts using statements that target *different* aspects of a concept, two almost identical statements do not satisfy this requirement. Therefore, the determinant of the Correlation matrix is required to be greater than 10^{-5}. If this condition is violated, correlations with $r > 0.8$ should be eliminated by removing one item at the time until the determinant is satisfactory.

The second test is Bartlett's test of sphericity, which reports how similar the Correlation matrix is to an identity matrix. The statistical significance of the similarity is quoted, and since the Correlation matrix is required to be considerably dissimilar to an identity matrix, which has no intervariable correlation, the p-value must be less than 0.05.

The last test is the Kaiser-Mayer-Olkin measure of sampling adequacy, or KMO. This measure predicts whether the data is expected to factor well. Its value should be greater than 0.5 for an adequate sample, but the greater the value, the better. In the Anti-image matrix, the diagonal elements are individual KMOs, whose average is the sample KMO. Variables with individual KMOs lower than 0.5 should be considered removed as they show an unacceptably high level of multicollinearity (see Hutcheson and Sofroniou 1999, for more detail).

Constructing the vector space

The remaining factor analysis will be explained invoking multi-dimensional vector spaces, where each variable is considered a unit vector. The correlation, r, between two variables is represented in vector space according to $r_{12} = x_1\, x_2 \cos\theta_{12}$, where θ is the angle between the two vectors. However, since each variable is a unit vector, this simplifies to $r = \cos\theta$. In this representation r is the fractional length of one vector projected onto the other. Note that r^2 represents the variance shared between the two vectors.

The following procedure will build up a k-dimensional space dimension by dimension. Let $\boldsymbol{x_1}$ represent the first variable, its base defining the origin of the vector space. The direction of $\boldsymbol{x_1}$ defines the first dimension. The second variable, $\boldsymbol{x_2}$, is placed at the origin at an angle θ_{12} to $\boldsymbol{x_1}$ according to r_{12}, thus introducing the second dimension. All remaining variables are introduced in the same way, ensuring that each new variable is positioned at the correct angle to all previously introduced variables until a k-dimensional space is constructed (assuming each variable introduces some unique variance).

The subsequent task is to introduce a coordinate system with k orthogonal axes. Introducing one axis at the time, the first axis is placed in the direction which maximizes the sum of squares of all vector projections onto the axis. The remaining axes are introduced according to the same condition, subject to the additional requirement of being orthogonal to the previously introduced axes. That is, the m^{th} coordinate axis is positioned so as to maximize E_m, given by

$$E_m = \sum_{n=1}^{k} \cos^2 \phi_{m,n} = \sum_{n=1}^{k} r_{m,n}^2$$

where $\phi_{m,n}$ is the angle and $r_{m,n}$ is the correlation coefficient between the n^{th} vector and the m^{th} axis.

Identifying and extracting factors

Much of the *SPSS* output in a factor analysis is direct reporting of variables described above. Each coordinate axis represents a factor, and E_m is the eigenvalue of the m^{th} factor, which is found in the *SPSS* output Total variance explained. In the same table, the Percentage of variance explained by the m^{th} factor is given by $\frac{E_m}{k}$. The Scree plot displays eigenvalue as a function of component number (factor).

Based on these outputs, the number of factors to extract is decided. Recall that the purpose of factor analysis is to maximize the amount of variance explained in the data with the minimum amount of factors. There are two methods to decide on the number of factors, which should be used in tandem: Kaiser's criterion and the Scree test. Kaiser's criterion states that all factors with an eigenvalue greater than 1 should be kept. Each factor accounts for $\frac{1}{k}$ of the information, but $\frac{E_m}{k}$ of the variance in the data. Consequently, factors with $E_m > 1$ account for a larger proportion of the variance explained than information retained. However, the Scree plot should also be consulted before the final decision is made. The plot consists of two parts: a steep decline at the first few factors, and a relatively flat plateau at higher order factors. The inflection point occurs immediately before the plateau, which represents factors containing mostly uninteresting, noisy variance. The factors prior to the inflection point stand out as they contain more variance per factor than those in the plateau, and we associate this with the underlying constructs. Generally both Kaiser's criterion and the Scree plot produce the same number of factors, but when this is not the case care should be taken to extract a sensible number of factors based on knowledge of the data set (see the next section for an example).

Once the number of factors or dimensions (f) has been chosen, all variables are effectively projected onto this f-dimensional sub-space. The squared length of each projected vector is the variance explained by the extracted factors collectively. These values are reported in the Communalities table. The resulting 'unexplained' variance is therefore simply the information discarded along with the discarded dimensions. The coordinates of each vector are referred to as the loadings onto each factor (or axis), and are reported in the Component matrix. When the coordinate axes are orthogonal the factor loadings correspond to the r-values for each variable-factor pair. Generally, only factor loadings greater than 0.4 are quoted for ease of table interpretation.

The current solution is referred to as the unrotated solution. The variables loading heavily onto one factor form a cluster of vectors intersected by the corresponding axis. However, due to the way the coordinate system was generated, this cluster intersection may not be optimal. Therefore, to optimize the individual factor loadings the entire f-dimensional coordinate system can be rotated. The criterion used is that each variable should load strongly onto only one axis (that is, the variable belongs to one underlying construct only). In an orthogonal rotation the axes are required to remain orthogonal, whereas an oblique rotation allows the axes to move independently of each other. The resulting angles between axes reflect correlations between the factors, which are presented in the Component correlation matrix.

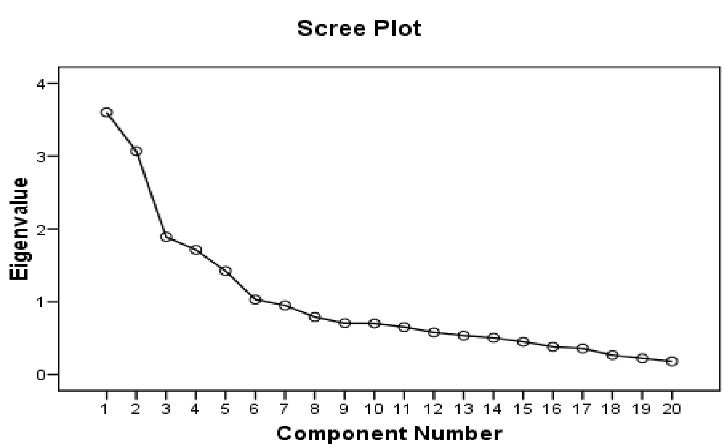

Figure 1. Scree Plot produced by SPSS for the Physics Goal Orientations survey

After rotation, the total variance explained by the factors remains the same since the projection of each variable onto the sub-space (i.e. the communality) is unrelated to the position of the coordinate axes. The factor loadings, however, have changed, and are presented in the Rotated component matrix for orthogonal rotations and in the Pattern matrix for oblique rotations. Note that after an oblique rotation the factor loadings are no longer equivalent to the variable-factor correlations. The correlations are presented in the Structure matrix, but this is generally ignored since a correlation in a non-orthogonal vector space includes information that is not unique to the particular variable-factor pair.

Analysis and interpretation

From the *SPSS* output the data were found suitable for factor analysis (determinant = 0.001, Bartlett's test: $p = 0.000$, and KMO = 0.664). All individual KMOs were > 0.5, except for two variables which had values of 0.484 and 0.483. However, being very close to 0.5, the variables were kept to consider their overall contribution to the analysis.

Kaiser's criterion initially extracted six factors. Investigation of the Scree plot (Figure 1), however, suggested retention of five factors only. The Component matrix supported this, as the sixth factor only contained one variable, hardly satisfying the critrion as a factor.

The analysis was therefore rerun specifying extraction of five factors. Note that the following tables and figures were unaffected by the number of factors extracted: Descriptive statistics, Correlation matrix, KMO and Bartlett's test, Anti-image matrices, and the Scree plot. The Total variance explained and Component matrix only saw the sixth factor removed. The Pattern matrix, Structure matrix, and Component correlation matrix did change, however.

Having decided the number of factors, the type of rotation was chosen. An oblique rotation (Direct Oblimin) was performed first to allow the data itself to reveal any correlations between factors, which were indeed observed. Had there been none, an orthogonal rotation (Varimax) could have subsequently been performed.

The Pattern Matrix (Table 2) revealed that variable 8 did not contribute strongly onto any of the extracted factors since it had no factor loadings greater than 0.4. This was not surprising as the variable showed a factor loading of 0.638 onto the initially extracted sixth factor, which was discarded. The variable was therefore removed.

Considering that the purpose of the Physics Goal Orientation survey is to obtain statements that collectively give indications about underlying psychological constructs, variables 1 and 4 were problematic. By loading onto two different factors, both variables targeted elements of two constructs simultaneously. The variables were therefore discarded.

Communalities reflect how much of the information in a variable is retained by the factors. Generally, a sample of less than 100 is acceptable if all communalities are above 0.6, and 100-200 is acceptable for communalities in the 0.5 range. Alternatively, if a factor has four or more factor loadings greater than 0.6 it is reliable. With an average communality of 0.58 after extracting five factors, the sample size was considered adequate. Since a reliable factor should have a minimum of four factor loadings greater than 0.6, only factors 1 and 5 currently satisfy this criterion.

Table 2 The Pattern matrix showing the factor loadings after an oblique rotation

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Item 6	.860				
Item 3	.701				
Item 2	.614				
Item 7	.612				
Item 5	.417				
Item 8					
Item 11		.890			
Item 10		.802			
Item 18		.789			
Item 1		.584	.469		
Item 13			.822		
Item 17			.778		
Item 16				.709	
Item 9				.660	
Item 14				.650	
Item 20					-.660
Item 19					-.646
Item 15					-.634
Item 12					-.607
Item 4		.416			-.416

As demonstrated above, factor analysis is not a clear cut process. Decisions have to be made and these are often not presented in research articles. The subjective nature makes it even more important that one has an understanding of the mathematical basis when practicing factor analysis or relying on studies that use factor analysis. As seen in this paper, the factors identified by Duda and Nicholls (1992) could not be reproduced in a physics setting. For a first trial of an adapted survey the structure is very promising, but addition of items and a retrial of the survey are necessary before it is fully developed.

What does Table 2 tell us? First, factor 1 reflects task or mastery orientation and this is clearly demonstarted both conceptually and in the data. It is interesting to note that item 3 on ‘group work in tutorials’ is in this factor reflecting the focus on constructive meaning making in learning physics. Factor 2 represents the ego orientation and factor 4 is clearly work avoidance. We have called factor 3 the interest orientation, but having only two items more will need to be added for the second trial of the survey. Factor 5 is the cooperation orientation, but it also contains an item (number 20) which does not conceptually belong with the rest of the items, even though all the items group mathematically. Item 20 will therefore be removed from the survey. This highlights one of the most important aspects of factor analysis: the mathematical sophistication of the analysis is of little worth if it is not accompanied by a critical mind.

Conclusion

This paper has demonstrated that surveys used within one area may not be directly applicable in another area. However, certain constructs do emerge clearly despite the change in discipline area. In our case task orientation, ego orientation and work avoidance were readily identifiable. The paper also aimed to give an insight into principal components analysis, and how subjective decisions need to be made when carrying out factor analysis. It is the hope of the authors that this will inspire fellow science education researchers to develop a more profound understanding of this complex statistical method.

References

Anderman, E.M., Austin, C. and Johnson, D. (2002) The development of goal orientation. In A. Wigfield and J. Eccles (Eds) *The Development of Achievement Motivation.* San Diego, CA: Academic Press, 197–220.

Duda, J.L. and Nicholls, J.G. (1992) Dimensions of achievement motivation in schoolwork and sport. *Journal of Educational Psychology*, **84**(3), 290–299.

Field, A. (2000) *Discovering Statistics Using SPSS for Windows.* London: Sage Publications Ltd.

Floyd, F.J. and Widaman, K.F. (1995) Factor Analysis in the Development and Refinement of Clinical Assessment Instruments. *Psychological Assessment*, **7**(3), 286–299.

Gorsuch, R.L. (1983) *Factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum.

Hutcheson, G. and Sofroniou, N. (1999) *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publications.

Preacher, K.J. and MacCallum, R.C. (2003) Repairing Tom Swift’s electric factor analysis machine. *Understanding Statistics*, **2**(1), 13-43.

Skaalvik, E.M. (1997) Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology*, **89**(1), 71–81.

Urdan, T., Kneisel, L. and Mason, V. (1999) Interpreting messages about motivation in the classroom: Examining the effects of achievement goals structures. In M. Maehr and P. Pintrich (Eds) *Advances in Motivation and Achievement.* Greenwich: JAI Press, 123–158.

Lindstrøm, C. and Sharma, M.D. (2008) Initial development of a Physics Goal Orientation survey using factor analysis, In A. Hugman and K. Placing (Eds) *Symposium Proceedings: Visualisation and Concept Development*, UniServe Science, The University of Sydney, 60–66.