Exploratory and Confirmatory Factor Analysis: Developing the Purpose in Life test-Short Form (PIL-SF)

Erin M. Buchanan

Missouri State University

Kathrene D. Valentine

Missouri State University

Stefan E. Schulenberg

University of Mississippi

Abstract

This article discusses the development of a short form of the Purpose in Life test (PIL) by using exploratory and confirmatory factor analysis. We describe why the creation of a useful, separate scale for measuring only meaning in life was desirable. The analyses used here, while normally complex, are broken down into manageable research questions that will show you how to interpret factor analyses and use these results in future endeavors. We first analyzed the original, 20-item version of the Purpose in Life test to determine which items would be the most beneficial at assessing meaning. Those items were examined in a second set of studies to analyze reliability of the proposed short form, which was revealed to be a handy tool for evaluating a person's perceived meaning in life. Meaning (or lack thereof) is an important concept, which is related to positive and negative variables, such as happiness, life satisfaction, depression, and

drug use. You will learn about the challenges and advantages to using factor analysis, and why it is an essential statistical procedure for researchers who are interested in scale development, validation, and refinement. This article is an extension of material originally presented by Schulenberg, Buchanan, and colleagues.

Contributor biographies

Erin M. Buchanan is an Assistant Professor of Psychology at Missouri State University. She has an undergraduate degree in psychology from Texas A&M University, and her master's degree and Ph.D. from Texas Tech University. Her research specialties include applied statistics with a focus on scale development and validation, as well as research on new statistical procedures and their implementation in the social sciences. She mainly teaches undergraduate and graduate statistics courses that cover the whole range of types of statistics, including structural equation modeling. Finally, she also is interested in understanding the underlying structure of our language systems and how those systems interaction with our ability to make judgments about the relationships between words.

Kathrene D. Valentine is currently an Adjunct Instructor at Missouri State University. She has an undergraduate degree in psychology and a master's degree in experimental psychology from Missouri State University. Her research interests include the importance of properly applied statistics in the social sciences with an emphasis on non-null hypothesis significance testing techniques, power and effect size, replication, and new statistical technologies. She teaches undergraduate introductory statistics courses.

Stefan E. Schulenberg (Ph.D., Clinical Psychology-Clinical/Disaster Specialty Track, University of South Dakota, 2001) is a licensed psychologist in the state of Mississippi and an Associate Professor in the University of Mississippi's Psychology Department. Dr. Schulenberg teaches graduate courses in cognitive assessment, personality assessment, and disaster mental health, as well as undergraduate courses in disasters and mental health, positive psychology, psychology and law, abnormal psychology, and tests and measurements. His research interests include clinical-disaster psychology, meaning and purpose in life, positive psychology, psychological assessment, serious mental illness, and adolescent psychopathology in the legal context. Dr. Schulenberg co-organizes Out of the Darkness community walks with the American Foundation for Suicide Prevention and serves as a disaster mental health volunteer in the American Red Cross. He was a mental health consultant for a research grant issued in response to Hurricane Katrina, and recently conducted evaluation research funded by the Mississippi Department of Mental Health relating to the effects of the Gulf Oil Spill. Dr. Schulenberg also serves as the Director of the University of Mississippi's Clinical-Disaster Research Center (UM-CDRC), an integrated research, teaching, training, and service Center with a primary emphasis in disaster mental health and a related emphasis in positive psychology.

Relevant disciplines

Psychology

Academic levels

Intermediate Undergraduate, Advanced Undergraduate, Postgraduate

Methods used

Exploratory factor analysis, confirmatory factor analysis, structural equation modeling, Cronbach's alpha

Keywords

Factor analysis, structural equation modeling, meaning in life, logotherapy, positive psychology, reliability, validity

Link to research output

DOI: 10.1007/s10902-008-9124-3

Schulenberg, S.E. & Melton, A.M.A. (2010). A confirmatory factor-analytic evaluation of the Purpose in Life test: Preliminary psychometric support for replicable two-factor model. *Journal of Happiness Studies, 11,* 95-111.

DOI: 10.1007/s10902-010-9231-9

Schulenberg, S.E., Schnetzer, L.W., & Buchanan, E.M. (2011). The Purpose in Life test-Short Form: Development and psychometric support. *Journal of Happiness Studies, 12,* 861-876.

Learning Outcomes

By the end of the case, students should:

- Know the purposes of exploratory and confirmatory factor analysis
- Be able to discriminate between exploratory and confirmatory factor analysis

- Know how to achieve simple structure in exploratory factor analysis and examine if that model is appropriate
- Know how to verify model results from exploratory factor analysis with confirmatory factor analysis

Purpose in Life test – Short Form (PIL-SF)

The 20-item Purpose in Life test (PIL) was designed by James Crumbaugh and Leonard Maholick as a means of assessing a person's perceived sense of meaning and purpose. It was initially based in the logotherapy framework originated by Viktor Frankl, but has been used extensively by those interested in the concept of meaning, regardless of the theoretical orientation of the researcher. Simply stated, logotherapy focuses on the paramount importance of perceived meaning and purpose in life to the human condition. Meaning, in terms of applications and assessment, has become increasingly popular in recent years, in large part due to its emphasis in the positive psychology movement. Meaning has been defined in a number of ways over the years. By way of example, Michael Steger has conceptualized meaning as "the web of connections, understandings, and interpretations that help us comprehend our experience and formulate plans directing our energies to the achievement of our desired future. Meaning provides us with the sense that our lives matter, that they make sense, and that they are more than the sum of our seconds, days, and years." (p. 165). Meaning is related to many positive variables, such as well-being, happiness, and life satisfaction, and meaninglessness is related to psychological distress, such as depression and anxiety.

One concern in the scientific study of meaning relates to how best to assess the concept in a reliable and valid fashion. The 20-item PIL, while one of the earliest and most influential measures of meaning, has come under scrutiny in recent years over concern about whether some items were appropriate for the scale (i.e., whether some items assessed related but distinct concepts, such as depression). Our research was designed to investigate the structure of the scale – we wanted to know if this scale was a useful way to measure meaning. These items include short sentences (e.g. "I am") followed by a seven-point rating scale with different anchors for each item (i.e. 1 – completely bored to 7 – enthusiastic). Item ratings are summed to derive a total score of 20 to 140 where higher scores indicate more perceived meaning in life. For this article, we incorporate and integrate several research projects to explain how we assessed and developed a scale that is reliable (shows consistent scores) and valid (measures what it is expected it to measure). First, before data collection, we analyzed power for an adequate sample size for our studies. Maxwell and colleagues have suggested guidelines ranging from 10 participants per item or simply 200+ participants. In our particular studies, we have used very large samples ranging from approximately 270 participants to over 900 participants to be able to determine the structure of the PIL. Second, we screened the data in each analysis for accuracy, missing data, outliers, and statistical assumptions, such as normality, linearity, and multicollinearity.

Exploratory Factor Analysis (EFA)

Exploratory factor analysis (EFA) is akin to a mathematical coin-sorting machine. Traditionally, EFA is analyzed to sort out individual items on a scale, but numerous types of

variables can be used as long as they have some expected similarity. The main goal of an EFA is to group variables into meaningful categories called factors, which are considering underlying constructs to the data. For instance, Vocabulary, Arithmetic, and Picture Completion task scores are all influenced by the primary intelligence of a person causing them to group together into a factor. When performing an EFA, there are several distinct questions to examine to determine the structure of the scale:

1) How many factors should I have?

2) How do I achieve simple structure?

3) Is my solution adequate and interpretable?

Number of factors

Often, a researcher will have an idea of the number of factors to expect when looking at a group of items on a scale. We thought the PIL scale would have one or two factors because 10 or more studies have analyzed the items to determine its underlying constructs. We cannot underestimate the importance of previous research and theory when trying to determine how many factors to expect. After a thorough literature search and discussion on the content of the PIL's items, we expected two factors to emerge that would be based on ratings of items about an exciting life (i.e., items like 2, 5, and 7) and a purposeful life (i.e., items like 3, 8, and 20). The specific question content can be found in our article about the short form. We did not expect all the items to group together, as some of them just did not seem to fit with the rest, but it was good to have an idea of a starting point. However, one of the cons of EFA is ending up with

unexplainable factors. This result often signals the need to rethink target items or explore why our expected factors did not appear in the data.

Thankfully, there are several other ways to estimate the number of factors to examine in EFA, and we will highlight three of them: eigenvalues, scree plots, and parallel analyses. First, eigenvalues are a mathematical representation of the amount of variance (the differences in scores across participants) that each factor accounts for in the data. Let's say we have sorted out our coins through the EFA machine. Eigenvalues would represent the size of the coin piles, where larger piles amount to explaining more of the underlying construct we are investigating. The Kaiser criterion is a guideline wherein the number of eigenvalues over one is used to determine the number of factors. Eigenvalues vary in size depending on the data; so many researchers have now shifted to examining a scree plot instead of a simple cut-off rule. A scree plot is a line graph that shows the factor number on the x-axis and the eigenvalue size on the y-axis. One confusing aspect of these graphs (and eigenvalues in general) is that there will always be the same number of eigenvalues as items on the scale. Even though we only wanted two factors, the variance is sorted into the same number of piles as items. We were looking for a large drop off between eigenvalues, where the rest of the points appear to be about the same size. In Figure 1 (Caption: Scree plots denoting the size of eigenvalues for the PIL 20 question scale. The left plot shows an example of a one-factor model, while the right plot shows a two-factor model.), the left scree plot would indicate a one-factor model, while the right scree plot would indicate a two-factor model. Unfortunately for us, our theory (two-factors) does not match the scree plot (one-factor). However, we can check our scree plot results with a parallel analysis, where each eigenvalue is examined to determine if it is bigger than a random assortment of the data. To calculate a parallel analysis, we used a free EFA program called FACTOR by Urbano

Lorenzo-Seva and Pere Joan Ferrando because IBM SPSS Statistics, the popular statistical analysis program that many researchers use, does not compute parallel analyses. We found that two factors were better than chance, and therefore, we decided to test a two-factor model. If the different criterions disagree, we could have analyzed several models to determine which one is best by using the simple structure rules described below.

Simple structure

Once we decided on the number of factors to analyze, we needed to determine which type of mathematical fitting estimation and rotation option to use. First, the fitting estimation is how the coin-sorting machine determines in what way to make the piles of variance for eigenvalues. Maximum likelihood and principle axis factoring are common ways to estimate factor loadings for normal data; that is, we assume that the data are continuous and have a normal distribution. We used maximum likelihood because our data fit this assumption, but it is important to be sure to check the data's distribution to pick the right type of estimation function. Second, we focused on the type of rotation. Rotation helps the researcher achieve simple structure, where the piles of coins created by the EFA sorting machine are easier to interpret. Rotation types fall into two categories: orthogonal and oblique. Orthogonal rotations force the factors to be unrelated, while oblique rotations allow factors to be correlated. In psychology, it is often unrealistic to use uncorrelated rotations because we find that many of our underlying constructs are related to each other. Also, when an oblique rotation is used and factors are truly unrelated, we would have found the same solution as if we had used an orthogonal rotation. Therefore, we used an oblique rotation called direct oblimin, which is one of several oblique rotations commonly seen in

research articles.

Table 1 shows the estimation of our two-factor model given maximum likelihood estimation and the oblique direct oblimin rotation. This table contains the factor loadings, which are the relationship between the items and the factor. Basically, these values told us if the particular coin matches the general coin pile. We wanted to find high values for loadings (up to 1.00) to indicate that our items were strong indicators for that factor. Most researchers use a cut-off score of 0.300 as a guideline to suggest that the item is loading onto that factor. We have bolded those factor loadings in Table 1. However, to keep things simple, we want each coin to be in only one pile. Therefore, simple structure occurs when items load onto one and only one factor and the combination of items for each factor are interpretable. We did not find simple structure on the first pass of our two-factor model, as shown in the decision column of Table 1. We found several items for factor 1 and factor 2 that clearly were associated with only one factor. Then several items were considered "bad" items because they did not load onto either factor (items 7, 13, 14, 15, 18) or split loadings between two factors (items 11, 12, 17, 19). This result is fairly normal and to achieve simple structure, we removed the bad items and reran the two-factor model without them (Table 2). In this second analysis, we found that only item 4 split between two factors. The third analysis showed no bad items, and we found simple structure with each item loading onto one and only one factor (Table 3). However, item 4 turned out to be a bit of an unexpected result, but in this particular dataset, it did not "want" to load strongly onto only one factor.

Adequate solution

The last stage of working with an EFA is to examine if the sorted simple structure fits the data. The coin sorter could show nice, clean piles of coins but, in reality, the coins do not match the pattern of coins in the real world. This match is called model fit. Fit indices can appear to be a bit of an alphabet soup, but we will talk about the ones we normally use for research papers involving EFA and confirmatory factor analysis. The common fit indices fall into two categories: residual statistics and goodness of fit statistics (alternatively: absolute or incremental). Both fit statistics are measurements that look at how well we fit our proposed model to the real model of the data (i.e., number of factors, loadings) and measure the discrepancy between the two models. For residual fit indices, we wanted to see very small numbers indicating that our error (i.e., residuals) was very small. For goodness of fit indices, we wanted to find larger numbers that indicate a better fit to the model, since a fit of 1.00 would indicate a perfect match to the data. Table 4 includes a list of common fit indices and their criteria for good fitting models. The root mean square error of approximation (RMSEA) and standardized root mean residual (SRMR) are residual error measurements, while the normed fit index (NFI), Tucker-Lewis index (TLI or non-normed fit index NNFI), and comparative fit index (CFI) are all types of goodness of fit indices. Other fit indices exist, such as the goodness of fit index (GFI), but are not recommended because they have been shown to be biased estimators (e.g., they show very high values). Additionally, predictive fit indices are often used for comparing different models such as the Akaike fit index or expected cross validation index where lower values indicate a better model over the comparison model. In our PIL EFA two-factor model, the fit indices: RMSEA (0.07), SRMR (0.04), TLI (0.93), and CFI (0.96), were good and excellent fit indices. This finding indicated that our simple structure found in Table 3 was probably a good indication of the underlying relationships between items.

Another check for model fit could be to examine the reliability of each factor using Cronbach's alpha. This measure tells us how consistent these items are together, and scores over .80 are recommended. Our factor 1 (0.83) and factor (0.82) showed good internal consistency reliability scores. The last facet of an adequate solution is whether the factors are interpretable – that is, could we understand what the underlying construct might be based on the pattern of loadings we found in the model fitting section? Factor 1 included items like enthusiasm, excitement, newness, and living a life that is worthwhile, which were consistent with previous findings of the "exciting life" factor. Factor 2 items all centered around meaning and/or life goals, and we therefore described it as the "purposeful life" factor. We were especially interested in this second factor, since a meaningful, purposeful life was the variable that the PIL was designed to measure, and which was consistent with the original logotherapy framework on which the measure was based.

Purpose in Life test-Short Form (PIL-SF)

One of the main goals of our work with the PIL was to discover if it was a reliable and valid measurement of meaning and purpose in life, as described earlier. In looking at the EFA results (from our work and the work of others), we found that the first factor seemed to measure excitement, while our second factor seemed to assess a person's perception of their meaning/purpose. Consequently, we decided to develop a shorter PIL form that only evaluated the traditional meaning in life construct as consistent with the original intention of the developers of the measure. We used items 3, 8, and 20 because of previous analyses, but also included item 4 because it overtly measures a person's meaning. As described above, we had some trouble with

item 4 because it was considered a "bad" item in that particular analysis by splitting across our two factors. However, this item has clear face validity, as it appears to measure what we are aiming to measure (i.e., the item is transparent in measuring meaning by asking about a person's meaningful life). In other studies, item 4 has been grouped with items 3, 8, and 20, and adding this item was advantageous for reliability; scales with more items tend to be more reliable and show higher Cronbach's alpha scores. We developed the PIL-SF as a way to assess only meaning and purpose.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is very similar to EFA in terminology, but the goal of a CFA is to determine if a theorized factor model replicates in a new set of data. Since we had previously sorted our coins into coherent piles (e.g., exciting life and purposeful life factors), a CFA would analyze if those piles can be reproduced in a new set of data, rather than sorting the coins again. This point seems like a small distinction, but it is reasonably important: we were *confirming* a model of the data instead of *exploring* the possible models of the data. We already know the number of factors to expect and the placement of items in those factors, as we expected the four items of the PIL-SF to group together. This starting point means that we focused on two broad questions in CFA: 1) Can I find the same simple structure? and 2) Is my solution adequate and interpretable? To answer these questions, we used IBM SPSS's AMOS structural equation modeling (SEM) software, which allowed us to create a graphical representation of our model and designate how the structure should look. This program is fairly user friendly, but other options such as Mplus, Lisrel, and EQS will also perform these analyses. We have included a

drawing of our PIL-SF model in Figure 2 (Caption: A visual depiction of the confirmatory factor analysis model of the PIL-SF.). The squares on the diagram represent measured items that have actual scores in our dataset. Here, we have used items 3, 4, 8, and 20 as a way to estimate our latent factor of a meaningful/purposeful life. Latent factors are not measured directly (i.e., there is not one specific number for the variable in the data collected) but by way of using these items. A latent factor is the underlying construct found in our EFA and is represented by a circle on the diagram. We believe that the latent factor (meaningful/purposeful life) causes the answers on our items, so the arrows go from latent variable to measured variable (cause → effect). For example, if a participant believed they lead a meaningful life, they were likely to mark high scores on the questionnaire. Lastly, each measured variable is connected to an error variance, which is calculated to help us determine how well we have measured our model. These error variances are represented by a circle because they are estimated based on our data and model picture, not a response that we collected from participants.

Simple structure

Simple structure in CFA is a little different than EFA, in that we only allow items to load onto one factor at a time. In essence, we forced the items to show simple structure because of our previous results that indicated that they should only be connected to one factor. We used one factor, so we were concerned that the items were still strongly connected to that latent variable. Here we checked the loadings of the items on the factor, which are shown in Figure 2 next to each → arrow. These strong, positive loadings were all above our 0.300 cut-off described earlier. This result indicated that we still were able to show simple structure with our replication of the

factor.

Adequate solution

Model fit in CFA is assessed in the same fashion as EFA with the fit indices we described earlier. Researchers will also often report chi-square and degrees of freedom (*df*), but these values can be biased by large sample sizes – a catch-22 since large sample sizes are recommended for factor analysis. Our analysis of the four items for the meaning/purposeful life factor was excellent with very low residual statistics: RMSEA (0.00) and SRMR (0.00), and very high goodness of fit statistics: CFI (1.00), TLI (1.01), and NFI (1.00) for both datasets we examined of the PIL. Cronbach's alpha was calculated to assess internal consistency reliability, and we found a good score (0.86). We took these results to mean that the PIL-SF was a replicable, reliable scale to measure perceived purpose and meaning in life. To end this section, we will discuss the potential problems with CFA, as well as some extensions our research group has worked on with scale development.

Problems with CFA

As with any statistical procedure, CFA has a few quirky problems that can occur. The most common issue in SEM is when a model fails to converge. While this sounds like an accident waiting to happen, in reality it means that something is wrong with the model picture we were trying to examine. Sometimes, even with a good theoretical approach, our ideas about models just do not pan out. This problem can be hard to detect because we are invested in our

model, but presents us with a good opportunity to step back and determine if we are missing something about the constructs we are trying to measure. Another converging issue occurs with identification of a model. We were required to use more known pieces of information than unknown numbers to estimate, unless we wanted to be a popular gossip magazine. For example, we have one unknown factor in our model (the circle used to denote the meaning/purposeful life factor) and four known variables by using the four items from the PIL-SF (the squares in our model). We checked for identification by looking at the degrees of freedom in our model (*df*), which should be greater than zero. Lastly, another regular problem we might have encountered is Heywood cases. Heywood cases occur when an error variance is estimated to be a negative number, which should not be mathematically possible – variance is calculated by using squared scores. This problem is often tied to non-normal data, small sample sizes, or even outliers, which is why we highlighted the importance of data screening at the beginning of this article.

Multigroup analyses and other complex designs

We have presented an overview of one of the simplest types of CFAs with only one latent variable, but there are many extensions to these analyses. Some of the most commonly used designs involve second-order CFAs where multiple latent variables are stacked into a hierarchical design. For example, researchers who study intelligence believe that a generalized intelligence factor is linked to verbal and performance factors, which are then measured by parts of a standardized intelligence test. Multigroup modeling allows us to determine if our scale gives us the same results when we assess people of varying demographics (i.e., sex, race/ethnicity). This facet of scale measurement can be useful in determining if and where group differences

exist, which leads to critical decisions for cut-off scores on scales that assess constructs like psychological distress. Other potential uses for CFA include multitrait-multimethod (MTMM) designs that try to discriminate between multiple factors and their measurement methods, and multiple indicator multiple causes (MIMIC) designs that analyze how continuous variables (such as age) affect the performance of a scale. For the interested reader, Barbara Byrne has written fantastic SEM books that discuss these designs within the different programs available.

Conclusion

This article has shown the development of a short form version of the Purpose in Life test by walking through the steps of both exploratory and confirmatory factor analysis. Although many of our results were very strong, we still dealt with hiccups in the research along the way – large sample sizes take a long time to assess, data entry and screening can be tedious, and even our first try with the meaning/purposeful life factor indicated one item was potentially "bad." However, our replication of these studies and results with confirmatory analysis were key. Such replication studies are critical in any scientific endeavor. When results appear to be reliable and valid, we can use them to explore other fascinating questions, such as the relationship between meaning in life, alcohol use, and depression, as well as how meaning in life relates to self-efficacy, life satisfaction, and psychological distress following the Gulf Oil Spill.

Exercises and Discussion Questions

1. Why would a researcher be interested in using factor analyses when developing scales?

2. Explain how the rules of simple structure apply to each type of factor analysis. Why might those rules be important for understanding factors?
3. Why would researchers wish to replicate their study if their first exploratory analyses were what they expected to find?
4. If researchers expect to find three underlying factors, what other options might they explore to determine the "real" number of factors to analyze in an exploratory factor analysis?
5. Why are exploratory and confirmatory factor analyses separate steps in the scale development process?
6. What problems might occur with confirmatory factor analysis? Be sure to also consider the problems with planning for the research study.
7. Why is the consideration of fit indices for an adequate solution important if the factor analysis already shows simple structure?

Further Readings

Hicks, J. A., & Routledge, C. (Eds.). (2013). *The experience of meaning in life: Classical perspectives, emerging themes, and controversies.* Dordrecht: Springer.

Markman, K. D., Proulx, T., & Lindberg, M. J. (Eds.). (2013). *The psychology of meaning.* Washington, DC: American Psychological Association.

Wong, P. T. P. (Ed.). (2012). *The human quest for meaning: Theories, research, and applications* (2nd ed.). New York: Routledge/Taylor & Francis.

References

Crumbaugh, J. C., & Maholick, L. T. (1964/1969). *Manual of instructions for the Purpose in Life test*. Abilene, TX: Viktor Frankl Institute of Logotherapy.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Taylor and Francis.

Frankl, V. E. (1959/2006). *Man's search for meaning*. Boston: Beacon Press.

Kaiser, H.F. (1970). A second generation little jiffy. *Psychometrika, 35*, 401-415.

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavior Research Methods, 38,* 88-91.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *The Annual Review of Psychology*, *59,* 537-563. DOI: 10.1146/annurev.psych.59.103006.093735

Steger, M. F. (2012). Experiencing meaning in life: Optimal functioning at the nexus of spirituality, psychopathology, and well-being. In P. T. P. Wong (Eds.), *The human quest for meaning (2nd Ed)* pp. 165-184. New York: Routledge.

Tabachnick, B. G., & Fidell, L. S. (2012). *Using multivariate statistics* (6th ed.). Boston: Pearson.