JMP057: Exploratory Factor Analysis of Trust in Online Sellers

Exploratory Factor Analysis (EFA), Bartlett's Test, KMO Test

Produced by

Ross Metusalem, JMP Academic Team
Ross.Metusalem@jmp.com

Muralidhara A, JMP Academic Team
muralidhara.a@jmp.com

Exploratory Factor Analysis of Trust in Online Sellers

Exploratory Factor Analysis (EFA), Bartlett's Test, KMO Test

Key Ideas

This case study demonstrates the use of exploratory factor analysis – sometimes simply called factor analysis or common factor analysis – to explore the perceptions of online shoppers. The goal is to find a meaningful interpretation of observed variables in terms of unobserved or latent factors.

Background

Online shopping has influenced the world of marketing significantly. Various online retailers across the globe sell a multitude of products and services, and consumers often prefer the convenience of shopping from home. While it provides an opportunity to shop 24/7, online shopping also exposes customers to increased risk. They can fall victim to fraud through fake websites set up to steal personal information or through the theft of their information from a seller's database. Even without fraud, the relatively lower overhead of setting up an online store gives easy access to the marketplace to sellers who lack experience or skill, creating the risk of incorrect, incomplete, or lost orders. Therefore, to promote sales, online sellers have a strong incentive to encourage customers to view them as trustworthy.

The Task

Anna is a market researcher for an online retailer who wanted to identify the underlying factors driving customers' trust in online sellers. She adopted a primary method of data collection and prepared a set of 20 statements related to features of the online shopping experience that might influence customers' trust in the seller. In a questionnaire, respondents were asked to rate on a scale of 0 to 5 how strongly each feature would increase their sense of trust in the seller, with 0 being "not at all" and 5 being "very much." The 20 questions and their corresponding short abbreviations are provided in Exhibit 1.

Anna used exploratory factor analysis (EFA) to analyze the trust ratings data. With EFA, Anna can summarize the information from the 20 questionnaire items using a smaller number of latent constructs or factors. She then can apply her domain expertise to interpret the latent factors in terms of potential drivers of customer trust.

Exhibit 1 The questionnaire

#	Statements	Abbreviation
1	The website allows purchases without requiring an account (i.e., “guest” purchases)	Guest_purchase
2	The website displays testimonials from satisfied customers	Testimonials
3	The payment page prominently displays a security icon (e.g., a lock)	Security_icon
4	The website is easy to navigate and use	Navigation
5	The website presents a link to its return policy during the purchase	Return_policy
6	The website allows you to disable cross-website tracking	Cross_tracking
7	The website clearly states its policies regarding protection or confidentiality of personal information	Protection
8	The website is endorsed by a well-known celebrity or organization	Celebrity
9	The website clearly indicates whenever providing personal information is optional	Personal_info
10	The website accepts secure payment services (e.g., PayPal)	Secure_payment
11	The website has a modern design and appearance	Modern_design
12	The website advertises in well-known media outlets	Advertisement
13	The website sign-in or purchase processes use a CAPTCHA system (e.g., identifying objects in blurry photos)	Captcha
14	The website sign-in process offers multifactor authentication	Multifactor_authen
15	The online seller also operates physical stores	Physical_store
16	When providing an email address, you have a clear option for opting out of marketing emails	Opt_out
17	The website offers a “chat now” button to connect with support	Chat_now
18	The contact information for the seller includes a physical address	Contact_info
19	The website displays only the last four digits of your credit card number	Credit_card
20	The website offers 24-hour customer support	Customer_support

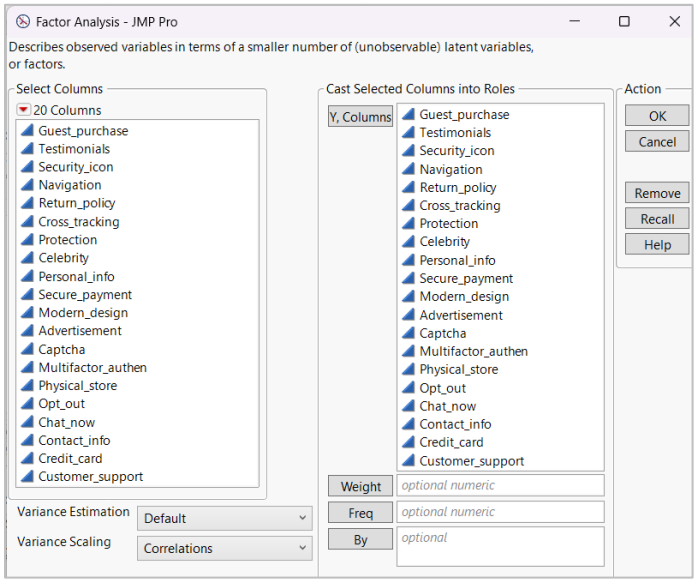
Anna collected data from 445 respondents, which can be found in consumer-fa.jmp. The Column Properties of each column include a note containing the full text of the corresponding questionnaire item. Hover your pointer over the column header to quickly view the full text.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) can be defined as the orderly simplification of interrelated measures. Traditionally, it has been used to explore the possible latent factor structure of a set of observed variables without imposing a preconceived structure on the outcome. The main difference between principal component analysis (PCA) and EFA is that the PCA analyzes the total variance, whereas EFA considers only common or shared variance. EFA assumes that there is one or more unobservable constructs that give rise to the data that we observe.

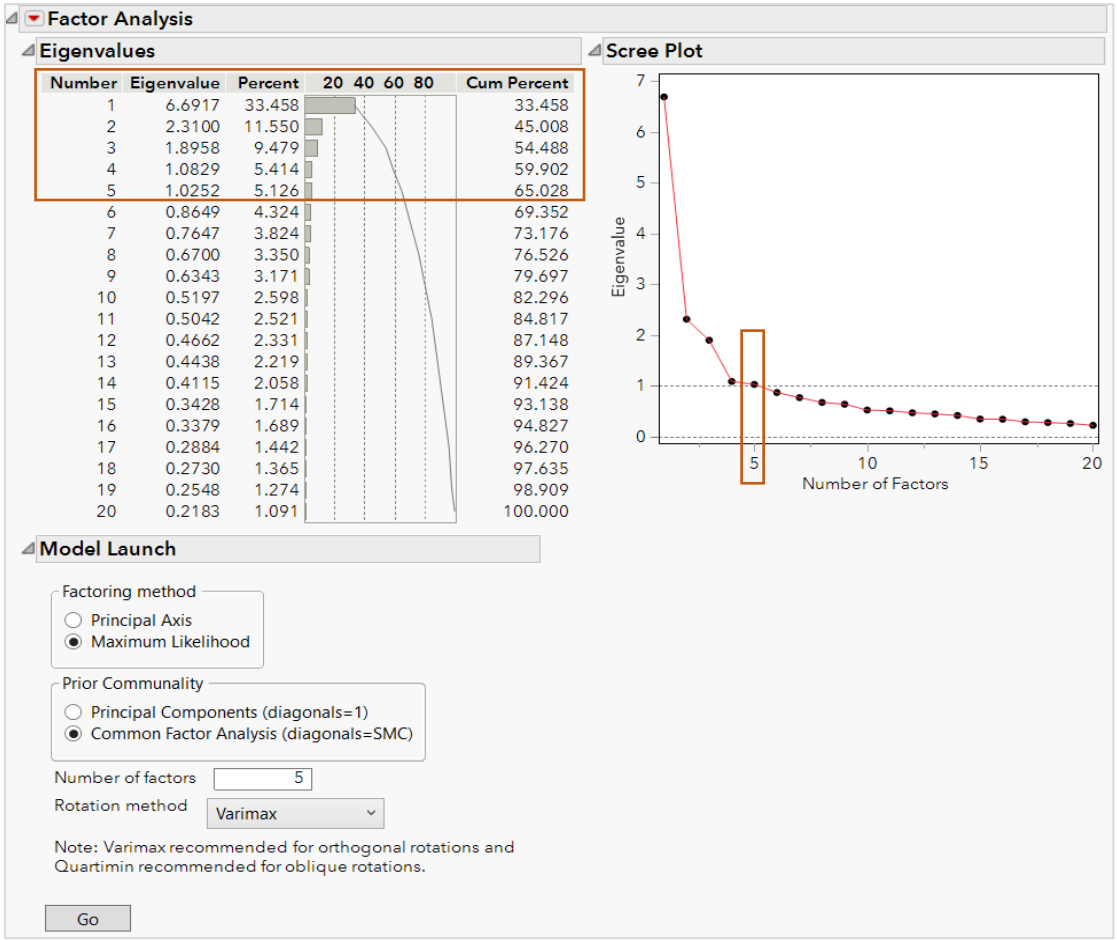
Exhibit 2 depicts launching the Factor Analysis platform in JMP. After running the EFA, the initial output will have an eigenvalues table, a scree plot, and a model launch section as shown in Exhibit 3.

Exhibit 2 Factor analysis dialog



(Analyze ➔ Multivariate Methods ➔ Factor Analysis. Select all the columns and click Y, Columns. ➔ Click OK.)

Exhibit 3 Initial report for choosing number of factors



Determining if the Data are Suitable for Analysis

Before we proceed with EFA, we need to test how well-suited the data are for EFA. There are two prerequisite tests of suitability: Bartlett's test of sphericity and the Keiser-Meyer-Olkin test.

Bartlett's test of sphericity tests the null hypothesis that the correlation matrix is an identity matrix, which would indicate that the variables are uncorrelated and, therefore, unsuitable for extraction of latent factors. If the test result is statistically significant (a p-value less than 0.05), we can conclude that the variables are correlated to some degree and are suitable for EFA.

The Kaiser-Meyer-Olkin (KMO) test measures sampling adequacy for each variable in the model and for the complete model. The test statistic, the measure of sampling adequacy (MSA), is a measure of the proportion of variance that might be common variance, potentially due to latent factors. The higher the MSA value is for a specific variable, the more suited that variable is to EFA. Variables with MSA values less than 0.5 may require remedial action, either by removing the variable(s) from the analysis or by including other variables correlated with those having low MSA values.

In Exhibit 4, we see that the p-value for Bartlett's test is less than 0.05, and the MSA values are greater than 0.5 for all but one variable (Captcha). These two tests confirm the suitability of the data for EFA, though in practice we may want to remediate the low MSA for the variable Captcha. For expository purposes, we will leave this variable in the analysis and proceed.

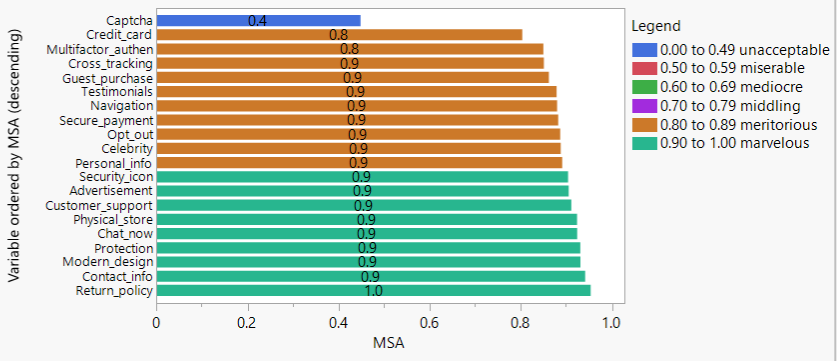
Exhibit 4 Bartlett's and KMO tests outputs

Bartlett's Test of Sphericity

ChiSquare	DF	Prob>ChiSq
4005.452	190	<.0001*

Kaiser-Meyer-Olkin Test

	MSA (Measure of Sampling Adequacy)
Guest_purchase	0.862
Testimonials	0.879
Security_icon	0.904
Navigation	0.880
Return_policy	0.954
Cross_tracking	0.851
Protection	0.931
Celebrity	0.888
Personal_info	0.891
Secure_payment	0.882
Modern_design	0.931
Advertisement	0.905
Captcha	0.448
Multifactor_authen	0.850
Physical_store	0.924
Opt_out	0.887
Chat_now	0.924
Contact_info	0.942
Credit_card	0.803
Customer_support	0.911
Overall MSA	0.903



Under the red triangle next to Factor Analysis, select Bartlett's Test of Sphericity and Keiser-Meyer-Olkin (KMO) test.

Choosing the Number of Factors

As part of the initial output, JMP produces eigenvalues and a scree plot, as shown in Exhibit 3. They are the result of an initial principal components analysis of the data and are intended to help in choosing the number of latent factors to extract. The eigenvalues table includes the percent of the total variance captured by each principal component, a bar chart illustrating these percent values, and the cumulative percent contributed by each successive principal component. We see that the first principal component accounts for 33.45% of the variation, the second accounts for 11.55%, and so forth. The first five principal components account for 65.03% of the total variation, and the contributions from the remaining principal

components are negligible. The number of eigenvalues that are greater than or equal to 1.0 can be taken as a guideline for the number of factors to extract. Based on this criterion, we would extract five factors.

The scree plot shows eigenvalues versus the number of factors. It can be used as an additional guideline to determine the number of factors to extract, with the point at which the plotted line becomes approximately level (the *elbow*) can also be used to determine the sufficient number of factors. The elbow on the scree plot indicates that a model with four or five factors would be appropriate. We choose to fit a model with five factors.

Model Launch

The model launch provides two factoring methods for estimating the parameters of the model, namely principal axis and maximum likelihood. The principal axis method performs eigenvalue decomposition on a reduced correlation or covariance matrix, where the diagonal of the matrix is replaced by an estimate of the communality of the variables. This is a computationally efficient method, but it does not allow for hypothesis testing. Maximum likelihood enables testing of hypotheses about the number of common factors as well as obtaining of model fit statistics. JMP selects this method by default.

The model launch also has two prior communality options for estimating the proportion of variance contributed by common factors for each variable, namely principal components and common factor analysis. The common factor analysis option, which sets the diagonal entries to squared multiple correlations, is used here. These values reflect the proportions of each variable's variance explained by all other variables in the analysis.

JMP considers the number of eigenvalues that exceed 1 as the default number of factors to extract. As previously mentioned, the value here is 5.

Factor rotation is used to support interpretability of the extracted factors. Rotations are applied to the factors extracted from the data. The Factor Analysis platform provides a variety of rotation methods that encompass both orthogonal (varimax, equamax, orthomax, parsimax, etc.) and oblique (oblimin, biquartimin, covarimin, promax, etc.) rotations. The varimax method (used here) maximizes the sum of the variances of the squared loadings of a factor on all variables. It is a common rotation method and results in each variable having either a small or large loading on each factor.

EFA Results

The EFA report in Exhibit 5 shows the final communality estimates, significance tests, measures of fit, rotated factor loadings, and a factor loading plot. We will explore each to interpret our results effectively.

The final communality for each variable is the sum of squared factor loadings for that variable. (We discuss loadings further below.) A larger value for a variable indicates stronger influence of one or more factors on that variable. Here, Captcha has a very low final communality value of 0.02, which is unsurprising given its low MSA score from the KMO test. It appears that responses to the Captcha item are not being driven by any of the underlying factors we have uncovered.

The output also provides two chi-squared significance tests. For the first, the null hypothesis is that none of the common factors explain the intercorrelations among the variables. This is Bartlett's test for sphericity, which we ran separately at a previous analysis step. Because the p-value is less than 0.05, we conclude that there is at least one latent factor that explains the intercorrelations among the variables.

The second test is for the null hypothesis that N factors are sufficient to explain the intercorrelations among the variables. Here, N=5. Rejection of the null hypothesis indicates that more factors might be required to explain the intercorrelations among the variables. Since the p-value is greater than 0.05, we fail to reject the null hypothesis and continue the analysis using five factors.

Exhibit 5 Partial EFA output

Final Communality Estimates

Guest_purchase	0.71678
Testimonials	0.48021
Security_icon	0.71931
Navigation	0.82023
Return_policy	0.50701
Cross_tracking	0.57091
Protection	0.69922
Celebrity	0.41727
Personal_info	0.49260
Secure_payment	0.50729
Modern_design	0.69906
Advertisement	0.58377
Captcha	0.02138
Multifactor_authen	0.50142
Physical_store	0.21086
Opt_out	0.45637
Chat_now	0.74854
Contact_info	0.62625
Credit_card	0.15928
Customer_support	0.71201

Factor Analysis on Correlations with 5 Factors: Maximum Likelihood / Varimax

Final Communality Estimates

Variance Explained by Each Factor

Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	190	4005.452	<.0001*

Test	DF	Criterion	ChiSquare	Prob>ChiSq
H0: 5 factors are sufficient. HA: more factors are needed.	100	0.270	116.819	0.1200

Measures of Fit

Measures of Fit	Fit Index
Chi-Square without Bartlett's Correction	119.741
AIC	-80.259
BIC	-490.066
Tucker and Lewis's Index	0.992
Root Mean Square Error of Approximation	0.021

In the Model Launch, set the Factoring Method to Maximum Likelihood, the Prior Communality to Common Factor Analysis, the Number of Factors to 5, and the Rotation Method to Varimax. ➔ Click Go to run the analysis.

The Measures of Fit table shows various goodness-of-fit measures including Tucker and Lewis's Index (TLI) and root mean square error of approximation (RMSEA). A TLI value greater than 0.9 is acceptable, and a RMSEA value close to 0 indicates a good model fit. Our TLI is 0.992 and RMSEA is 0.021, so we continue with interpreting our results.

Exhibit 6 shows a factor loading matrix and plots. Factor loadings measure the correlation between a latent factor and observed variable and can range from -1 to 1, with values close to those extremes indicating that the factor strongly influences the variable in either the negative or positive direction. Loadings close to 0 indicate that the factor has little-to-no influence on the variable.

The rotated factor loading matrix is ordered so that variables associated with the same factor appear next to each other. Use the Suppress Absolute Loading Values Less Than and Dim Text settings to better visualize which variables might be considered to load heavily on a given factor. We see that, for example, Guest_purchase, Cross_tracking, Opt_out, and Personal_info are strongly influenced by Factor 1. Additionally, the loadings for Captcha are low across all five factors. We may have expected this result, given the low MSA value for Captcha when we ran the KMO test.

Use the factor loading plot to visually identify which variables have the largest loading on the factors. Each plot depicts loadings for each variable on any pair of factors. For example, the top plot shows the loadings on Factor 1 vs Factor 2. We can hover on the dots to more easily view the name of the corresponding variable. We see, for example, that Guest_purchase has the strongest Factor 1 loading while Advertisement has the strongest Factor 2 loading. As we might now expect, we find Captcha near the center of each plot, indicating that it does not load strongly on any factor.

Note that we observe only positive loadings here, which is expected, because each statement described an online shopping feature believed to increase customer trust on the 0-to-5 scale. A negative loading for a statement would have indicated that the statement described an online shopping feature that would decrease trust. The survey did not include any such statements.

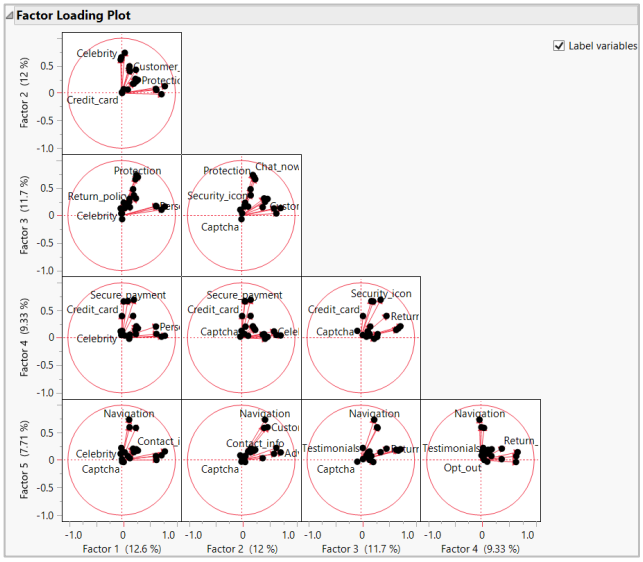
Exhibit 6 Rotated factor loading matrix and factor loading plot

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Guest_purchase	0.804467				
Cross_tracking	0.742711				
Opt_out	0.648466				
Personal_info	0.638635				
Advertisement		0.734721			
Testimonials		0.653300			
Celebrity		0.610628			
Physical_store		0.399471			
Chat_now			0.737336		
Protection	0.303403		0.698990		
Contact_info			0.659487		
Return_policy			0.478236	0.397976	
Security_icon			0.368357	0.692025	
Secure_payment				0.667688	
Multifactor_authen				0.662978	
Credit_card				0.396214	
Captcha					
Navigation		0.444006			0.735313
Customer_support		0.490363	0.302166		0.598691
Modern_design		0.426496	0.312464		0.586498

Suppress Absolute Loading Value Less Than 0.3

Dim Text 0



Factor Labeling and Using Factor Scores

Our goal was to uncover latent drivers of consumer trust in online sellers. To this end, we apply domain knowledge to interpret each factor by categorizing the variables that have large loadings on the same factor. The interpretation of factors is highly subjective, and arriving at a final interpretation typically requires team consensus. Exhibit 7 summarizes the factors with their corresponding variables and a possible conceptual label given in quotation marks. The discovery of these five factors represents the primary finding of this research study. These five factors can be used to guide the design of online shopping experiences that engender greater customer trust.

Exhibit 7 Questionnaire items grouped into factors based on loadings

Factor 1 “Privacy”	**Factor 2** “Reputation”	**Factor 3** “Transparency”	**Factor 4** “Security”	**Factor 5** “Customer Experience”
Guest_purchase	Advertisement	Chat_now	Security_icon	Navigation
Cross_tracking	Testimonials	Protection	Secure_payment	Customer_support
Opt_out	Celebrity	Contact_info	Multifactor_authen	Modern_design
Personal_info	Physical_store	Return_policy	Credit_card	

Finally, in EFA it is possible to calculate factor scores for each row in the data table (in this case, for each survey respondent). The factor scores represent linear combinations of the observed variables and can be useful in dimensionality reduction before further graphing or analysis. For example, a high Factor 1 (“Privacy”) score for a particular respondent might indicate that this respondent bases trust in online sellers strongly on protection of customer privacy.

The option to save the factor scores to the data table is found under the red triangle, though we will stop short of doing so here because quantifying trust-related factors in individual customers was not an aim of this research.

Summary

Statistical Insights

This case study covered the following statistical analyses procedures:

- Bartlett's test of sphericity
- Keiser-Meyer-Olkin (KMO) test
- Interpreting eigenvalues and scree plots
- Exploratory factor analysis
- Interpreting factor loadings

Managerial Implications

Anna drew the following conclusions from the analysis:

- Shoppers trust in online sellers may be composed of five latent factors: privacy of personal information, seller reputation, seller transparency, website security, and a good customer experience.
- The data suggest that the use of a Captcha system by an online seller does not seem to be related to any of these underlying factors.

JMP Features and Hints

This case used the Factor Analysis platform in JMP. It also leveraged Bartlett's test of sphericity and the Keiser-Meyer-Olkin (KMO) test for validating the data for structural detection before proceeding with EFA. We also saw how threshold and dimming settings for the rotated factor loading table can help visually distinguish strong loadings from weak ones.

Exercises

A local news organization in a mid-sized metropolitan area logs the amount of time each subscriber spends viewing content in each section of its website (see below). The company believes that one or more underlying latent factors related to news consumption habits or goals might drive the viewing times across the various website sections. It wants to use EFA to uncover these factors in order to better understand and serve its subscriber base.

The data set news-website-viewing-times.jmp contains data from 1,200 randomly selected subscribers to the news website. The values represent the average number of hours per month each user spent in each of seven website sections in the past 12 months:

- Local News: Current events in the local metropolitan area
- World News: Current events around the world
- Business: Economic and business events and forecasts
- Sci & Tech: Breakthroughs in science, technology, and healthcare
- Sports: Local and national sports coverage
- Culture: Local events, arts, food, and entertainment
- Opinion: Opinion pieces on a range of topics written by members of the editorial board, prominent public figures, and local residents

Enter the seven viewing time variables into the Factor Analysis platform. Follow the prompts below to perform and interpret an exploratory factor analysis on these data.

1. Run Bartlett's test of sphericity. What do the results indicate about the suitability of the data for factor analysis? How do you know?
2. Run the Kaiser-Meyer-Olkin test. According to the results, are there any variables that are not good candidates for inclusion in the analysis? How do you know?
3. According to the information in the eigenvalues table and scree plot, how many factors should you extract in the analysis? Explain your reasoning.

Run an EFA with two factors. Use the principal axis factoring method, the common factor analysis prior communality method, and the varimax rotation method.

4. Interpret the results of the two chi-square significance tests. What do they tell us about how well two latent factors can capture the variation in our seven observed variables?
5. According to the information in the measures of fit table, is our model fit good enough to move on to interpreting of the factors? How do you know?
6. Which viewing time variable is most strongly correlated with Factor 1? How do you know?
7. Interpret the two factors. What underlying factors related to news consumption habits or goals might this analysis have revealed? How do you come to that conclusion? Give each factor a one- or two-word name if that helps your explanation.