Statistical Discovery.™ From SAS.

Version 11

Consumer Research

"The real voyage of discovery consists not in seeking new landscapes, but in having new eyes."

Marcel Proust

JMP, A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513

11.1

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2014. *JMP® 11 Consumer Research*. Cary, NC: SAS Institute Inc.

JMP® 11 Consumer Research

ISBN 978-1-61290-661-4

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, September 2013

2nd printing, February 2014

Technology License Notices

- Scintilla - Copyright © 1998-2012 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

- Telerik RadControls: Copyright © 2002-2012, Telerik. Usage of the included Telerik RadControls outside of JMP is not permitted.

- ZLIB Compression Library - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.

- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.

- Packages - Copyright © 2009-2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS

OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software - Copyright © 1995-2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
  - Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- bzip2, the associated library "libbzip2", and all documentation, are Copyright © 1996-2010, Julian R Seward. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

Get the Most from JMP®

Whether you are a first-time or a long-time user, there is always something to learn about JMP.

Visit JMP.com to find the following:

- live and recorded webcasts about how to get started with JMP
- video demos and webcasts of new features and advanced techniques
- details on registering for JMP training
- schedules for seminars being held in your area
- success stories showing how others use JMP
- a blog with tips, tricks, and stories from JMP staff
- a forum to discuss JMP with other users

http://www.jmp.com/getstarted/

Contents

Consumer Research

Chapter 1

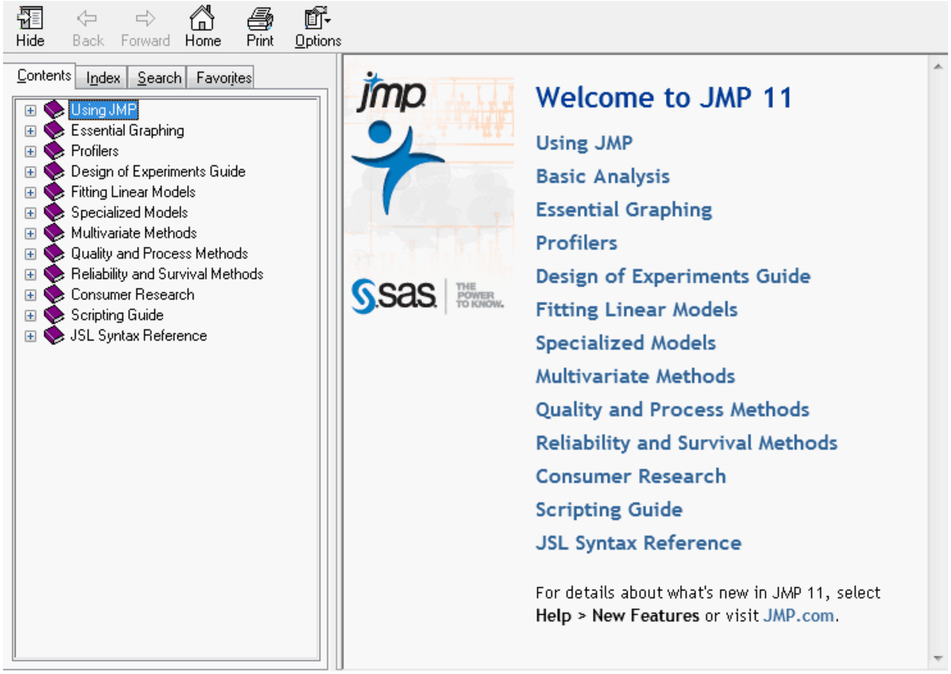
Learn about JMP

Documentation and Additional Resources

This chapter includes the following information:

- book conventions
- JMP documentation
- JMP Help
- additional resources, such as the following:
  - other JMP documentation
  - tutorials
  - indexes
  - Web resources

Figure 1.1 The JMP Help Home Window on Windows



Using JMP
Essential Graphing
Profilers
Design of Experiments Guide
Fitting Linear Models
Specialized Models
Multivariate Methods
Quality and Process Methods
Reliability and Survival Methods
Consumer Research
Scripting Guide
JSL Syntax Reference

Welcome to JMP 11

Using JMP
Basic Analysis
Essential Graphing
Profilers
Design of Experiments Guide
Fitting Linear Models
Specialized Models
Multivariate Methods
Quality and Process Methods
Reliability and Survival Methods
Consumer Research
Scripting Guide
JSL Syntax Reference

For details about what's new in JMP 11, select **Help > New Features** or visit JMP.com.

Contents

Formatting Conventions

The following conventions help you relate written material to information that you see on your screen.

- Sample data table names, column names, pathnames, filenames, file extensions, and folders appear in Helvetica font.
- Code appears in `Lucida Sans Typewriter` font.
- Code output appears in *`Lucida Sans Typewriter`* italic font and is indented farther than the preceding code.
- **Helvetica bold** formatting indicates items that you select to complete a task:
  - buttons
  - check boxes
  - commands
  - list names that are selectable
  - menus
  - options
  - tab names
  - text boxes
- The following items appear in italics:
  - words or phrases that are important or have definitions specific to JMP
  - book titles
  - variables
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit http://www.jmp.com/software/pro/.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Documentation

JMP offers documentation in various formats, from print books and Portable Document Format (PDF) to electronic books (e-books).

- Open the PDF versions from the **Help > Books** menu or from the JMP online Help footers.

- All books are also combined into one PDF file, called *JMP Documentation Library*, for convenient searching. Open the *JMP Documentation Library* PDF file from the **Help > Books** menu.
- e-books are available at Amazon, Safari Books Online, and in the Apple iBookstore.
- You can also purchase printed documentation on the SAS website:

  http://support.sas.com/documentation/onlinedoc/jmp/index.html

JMP Documentation Library

The following table describes the purpose and content of each book in the JMP library.

Document Title	Document Purpose	Document Content
Discovering JMP	If you are not familiar with JMP, start here.	Introduces you to JMP and gets you started creating and analyzing data.
Using JMP	Learn about JMP data tables and how to perform basic operations.	Covers general JMP concepts and features that span across all of JMP, including importing data, modifying columns properties, sorting data, and connecting to SAS.
Basic Analysis	Perform basic analysis using this document.	Describes these Analyze menu platforms: • Distribution • Fit Y by X • Matched Pairs • Tabulate How to approximate sampling distributions using bootstrapping is also included.

Document Title	Document Purpose	Document Content
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Overlay Plot • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Treemap • Scatterplot Matrix • Ternary Plot • Chart Also covers how to create background and custom maps.
Profilers	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the **DOE** menu.

Document Title	Document Purpose	Document Content
Fitting Linear Models	Learn about Fit Model platform and many of its personalities.	Describes these personalities, all available within the Analyze menu Fit Model platform: • Standard Least Squares • Stepwise • Generalized Regression • Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model
Specialized Models	Learn about additional modeling techniques.	Describes these Analyze > Modeling menu platforms: • Partition • Neural • Model Comparison • Nonlinear • Gaussian Process • Time Series • Response Screening The Screening platform in the Analyze > Modeling menu is described in *Design of Experiments Guide*.
Multivariate Methods	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Cluster • Principal Components • Discriminant • Partial Least Squares

Document Title	Document Purpose	Document Content
Quality and Process Methods	Read about tools for evaluating and improving processes.	Describes these Analyze > Quality and Process menu platforms: • Control Chart Builder and individual control charts • Measurement Systems Analysis • Variability / Attribute Gauge Charts • Capability • Pareto Plot • Diagram
Reliability and Survival Methods	Learn to evaluate and improve reliability in a product or system and analyze survival data for people and products.	Describes these Analyze > Reliability and Survival menu platforms: • Life Distribution • Fit Life by X • Recurrence Analysis • Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Survival • Fit Parametric Survival • Fit Proportional Hazards
Consumer Research	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Factor Analysis • Choice • Uplift • Item Analysis

Document Title	Document Purpose	Document Content
Scripting Guide	Learn about taking advantage of the powerful JMP Scripting Language (JSL).	Covers a variety of topics, such as writing and debugging scripts, manipulating data tables, constructing display boxes, and creating JMP applications.
JSL Syntax Reference	Read about many JSL functions on functions and their arguments, and messages that you send to objects and display boxes.	Includes syntax, examples, and notes for JSL commands.

Note: The **Books** menu also contains two reference cards that can be printed: The *Menu Card* describes JMP menus, and the *Quick Reference* describes JMP keyboard shortcuts.

JMP Help

JMP Help is an abbreviated version of the documentation library that provides targeted information. You can open JMP Help in several ways:

- On Windows, press the F1 key to open the Help system window.
- Get help on a specific part of a data table or report window. Select the Help tool ? from the **Tools** menu and then click anywhere in a data table or report window to see the Help for that area.
- Within a JMP window, click the **Help** button.
- Search and view JMP Help on Windows using the **Help > Help Contents**, **Search Help**, and **Help Index** options. On Mac, select **Help > JMP Help**.
- Search the Help at http://jmp.com/support/help/ (English only).

Additional Resources for Learning JMP

In addition to JMP documentation and JMP Help, you can also learn about JMP using the following resources:

- Tutorials (see "Tutorials" on page 19)
- Sample data (see "Sample Data Tables" on page 19)
- Indexes (see "Learn about Statistical and JSL Terms" on page 19)

- Tip of the Day (see "Learn JMP Tips and Tricks" on page 20)
- Web resources (see "JMP User Community" on page 20)
- JMPer Cable technical publication (see "JMPer Cable" on page 20)
- Books about JMP (see "JMP Books by Users" on page 21)
- JMP Starter (see "The JMP Starter Window" on page 21)

Tutorials

You can access JMP tutorials by selecting **Help > Tutorials**. The first item on the **Tutorials** menu is **Tutorials Directory.** This opens a new window with all the tutorials grouped by category.

If you are not familiar with JMP, then start with the **Beginners Tutorial**. It steps you through the JMP interface and explains the basics of using JMP.

The rest of the tutorials help you with specific aspects of JMP, such as creating a pie chart, using Graph Builder, and so on.

Sample Data Tables

All of the examples in the JMP documentation suite use sample data. Select **Help > Sample Data** to do the following actions:

- Open the sample data directory.
- Open an alphabetized list of all sample data tables.
- Find a sample data table within a category.

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\<version_number>\Samples\Data

On Macintosh: \Library\Application Support\JMP\<version_number>\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory.

Learn about Statistical and JSL Terms

The **Help** menu contains the following indexes:

Statistics Index Provides definitions of statistical terms.

Scripting Index Lets you search for information about JSL functions, objects, and display boxes. You can also edit and run sample scripts from the Scripting Index.

Learn JMP Tips and Tricks

When you first start JMP, you see the Tip of the Day window. This window provides tips for using JMP.

To turn off the Tip of the Day, clear the **Show tips at startup** check box. To view it again, select **Help > Tip of the Day**. Or, you can turn it off using the Preferences window. See the *Using JMP* book for details.

Tooltips

JMP provides descriptive tooltips when you place your cursor over items, such as the following:

- Menu or toolbar options
- Labels in graphs
- Text results in the report window (move your cursor in a circle to reveal)
- Files or windows in the Home Window
- Code in the Script Editor

Tip: You can hide tooltips in the JMP Preferences. Select **File > Preferences > General** (or **JMP > Preferences > General** on Macintosh) and then deselect **Show menu tips**.

JMP User Community

The JMP User Community provides a range of options to help you learn more about JMP and connect with other JMP users. The learning library of one-page guides, tutorials, and demos is a good place to start. And you can continue your education by registering for a variety of JMP training courses.

Other resources include a discussion forum, sample data and script file exchange, webcasts, and social networking groups.

To access JMP resources on the website, select **Help > JMP User Community**.

JMPer Cable

The JMPer Cable is a yearly technical publication targeted to users of JMP. The JMPer Cable is available on the JMP website:

http://www.jmp.com/about/newsletters/jmpercable/

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website:

http://www.jmp.com/support/books.shtml

The JMP Starter Window

The JMP Starter window is a good place to begin if you are not familiar with JMP or data analysis. Options are categorized and described, and you launch them by clicking a button. The JMP Starter window covers many of the options found in the **Analyze**, **Graph**, **Tables**, and **File** menus.

- To open the JMP Starter window, select **View** (**Window** on the Macintosh) **> JMP Starter**.
- To display the JMP Starter automatically when you open JMP on Windows, select **File > Preferences > General**, and then select **JMP Starter** from the Initial JMP Window list. On Macintosh, select **JMP > Preferences > Initial JMP Starter Window**.

Chapter **2**

Introduction to Consumer Research

Overview of Customer and Behavioral Research Methods

You already collect information about how customers use a product or service or how satisfied they are with your offerings. The resulting insight lets you create better products and services, happier customers, and more revenue for your organization.

JMP now includes a full suite of tools for performing customer and consumer research. In the past, you might have had to use one product for consumer research work and JMP for design of experiments. Now you can do both types of analyses using a single product, for a more efficient use of your most precious resource: your time. Tools for performing these statistical analyses are now located in one convenient place: the Consumer Research menu. Use the following platforms to analyze your data:

- The Categorical platform supports survey analysis with questions in multiple formats, allowing for both detailed and compact reporting. You can also analyze multiple response questions, where your survey includes questions for which respondents can choose more than one answer. You can output the results in crosstab report tables, use share and frequency charts, view mean scores across responses, and perform tests and comparisons. And when you are finished, you can easily output the completed analysis tables. For more information, see Chapter 3, "Categorical Response Analysis".
- The Factor Analysis platform enables you to discover simple arrangements in the pattern of relationships among variables. It seeks to discover if the observed variables can be explained in terms of a much smaller number of variables or factors. By using factor analysis, you can determine the number of factors that influence a set of measured, observed variables, and the strength of the relationship between each factor and each variable. For more information, see Chapter 4, "Factor Analysis".
- The Choice platform is designed for use in market research experiments, where the ultimate goal is to discover the preference structure of consumers. Then, this information is used to design products or services that have the attributes most desired by consumers. For more information, see Chapter 5, "Choice Models".
- The Uplift platform enables you to maximize the impact of your marketing budget by sending offers only to individuals who are likely to respond favorably, even when you have large data sets and many possible behavioral or demographic predictors. You can use uplift models to make such predictions. This method has been developed to help optimize marketing decisions, define personalized medicine protocols, or, more generally, to identify characteristics of individuals who are likely to respond to some action. For more information, see Chapter 6, "Uplift Models".

- The Item Analysis platform provides a method of scoring tests. Based on Item Response Theory (IRT), the platform helps analyze the design, analysis, and scoring of tests, questionnaires, and other tools that measure abilities, attitudes, and other variables. Although classical test theory methods have been widely used for a century, IRT provides a better and more scientifically based scoring procedure. For more information, see Chapter 7, “Item Analysis”.

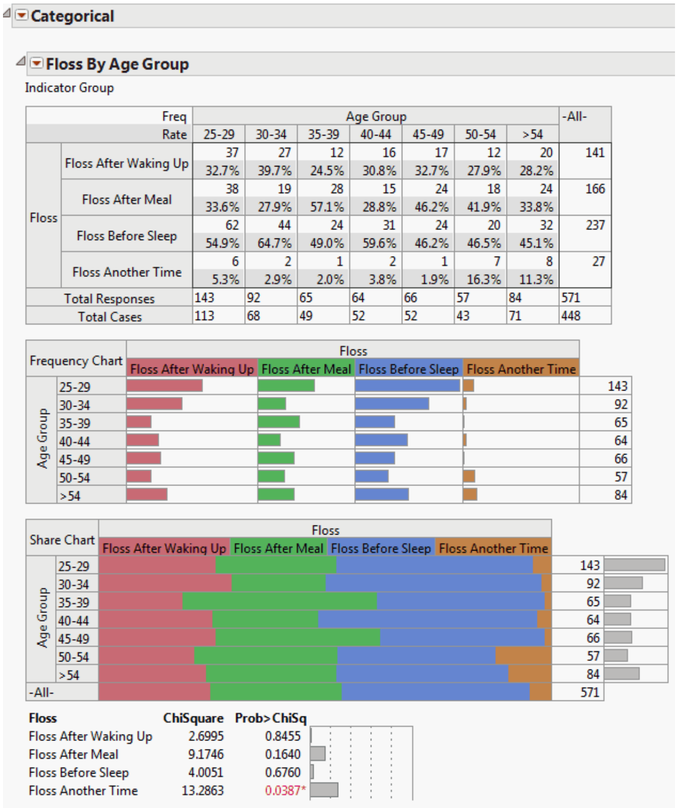
Chapter 3

Categorical Response Analysis

Analyzing Survey and Other Counting Data

The Categorical platform tabulates and summarizes categorical response data, including multiple response data, and calculates test statistics. The strength of the Categorical platform is that it can handle responses in a wide variety of formats without needing to reshape the data. It is designed to handle survey and other categorical response data, such as defect records, side effects, and so on.

Figure 3.1 Categorical Analysis Example



Categorical

Floss By Age Group

Indicator Group

	Freq Rate	Age Group 25-29	 30-34	 35-39	 40-44	 45-49	 50-54	 >54	-All-
Floss	Floss After Waking Up	37 32.7%	27 39.7%	12 24.5%	16 30.8%	17 32.7%	12 27.9%	20 28.2%	141
	Floss After Meal	38 33.6%	19 27.9%	28 57.1%	15 28.8%	24 46.2%	18 41.9%	24 33.8%	166
	Floss Before Sleep	62 54.9%	44 64.7%	24 49.0%	31 59.6%	24 46.2%	20 46.5%	32 45.1%	237
	Floss Another Time	6 5.3%	2 2.9%	1 2.0%	2 3.8%	1 1.9%	7 16.3%	8 11.3%	27
Total Responses		143	92	65	64	66	57	84	571
Total Cases		113	68	49	52	52	43	71	448

Floss	ChiSquare	Prob>ChiSq
Floss After Waking Up	2.6995	0.8455
Floss After Meal	9.1746	0.1640
Floss Before Sleep	4.0051	0.6760
Floss Another Time	13.2863	0.0387*

Contents

Categorical Platform Overview

The Categorical platform can produce results from a rich variety of organizations of data, as reflected in the tabbed panels that enable you to specify the analyses that you want. The Categorical platform has capabilities similar to other platforms. The choice of platform depends on your focus, the shape of your data, and the desired level of detail. The strength of the Categorical platform is that it can handle responses in a wide variety of formats without needing to reshape the data. Table 3.1 shows several of JMP's analysis platforms and their strengths.

Table 3.1 Comparing JMP's Categorical Analyses

Platform	Specialty
Distribution	Separate, ungrouped categorical responses.
Fit Y By X: Contingency	Two-way situations, including chi-square tests, correspondence analysis, agreement.
Pareto Plot	Graphical analysis of multiple-response data, especially multiple-response defect data, with more rate tests than Fit Y By X.
Variability Chart: Attribute	Attribute gauge studies, with more detail on rater agreement.
Fit Model	Logistic categorical responses and generalized linear models.
Partition, Neural Net	Specific categorical response models.

Example of the Categorical Platform

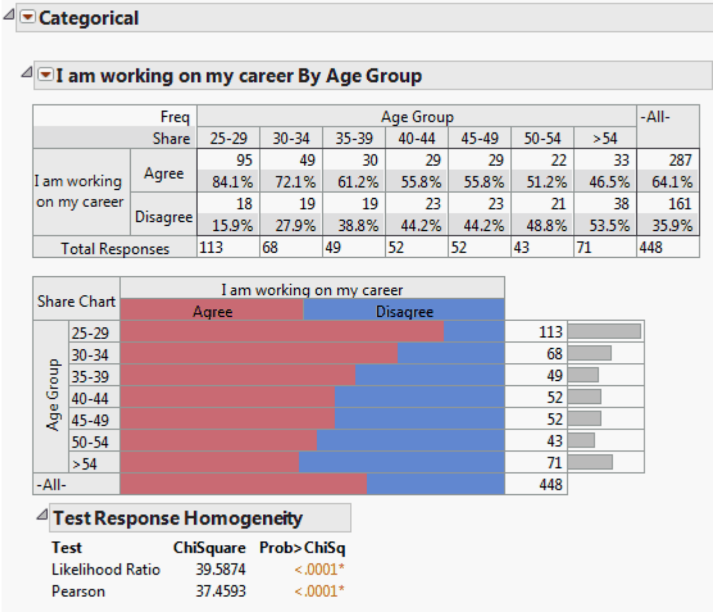
This example uses the Consumer Preferences.jmp sample data table, which contains survey data on people's attitudes and opinions, and some questions concerning oral hygiene (source: Rob Reul, Isometric Solutions).

1. Open the Consumer Preferences.jmp sample data table.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select I am working on my career and click **Responses** on the Simple tab.
4. Select Age Group and click **X, Grouping Category**.
5. Click **OK**.
6. Select **Crosstab Transposed** from the Categorical red triangle menu.

7. Select **Test Response Homogeneity** from the Categorical red triangle menu.

Figure 3.2 details the responses indicating that a respondent is currently working on his or her career and the age group. From the analysis, you can determine that of the 448 respondents, 64.1% indicated that they were working on their career. Of those responding positively, the highest majority working on their career were in the age group 25-29 at 84.1%. The highest majority of those responding oppositely were in the age group > 54 at 53.5%.

Figure 3.2 Survey Results by Age Group



Categorical

I am working on my career By Age Group

	Freq Share	Age Group 25-29	 30-34	 35-39	 40-44	 45-49	 50-54	 >54	-All-
I am working on my career	Agree	95 84.1%	49 72.1%	30 61.2%	29 55.8%	29 55.8%	22 51.2%	33 46.5%	287 64.1%
	Disagree	18 15.9%	19 27.9%	19 38.8%	23 44.2%	23 44.2%	21 48.8%	38 53.5%	161 35.9%
Total Responses		113	68	49	52	52	43	71	448

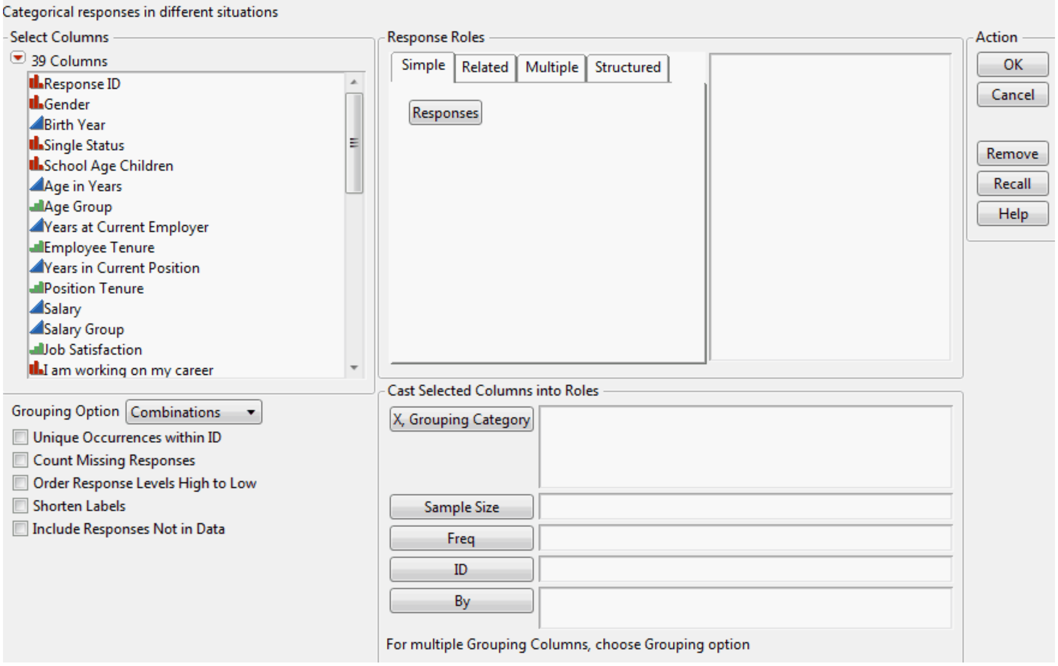
Test Response Homogeneity

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	39.5874	<.0001*
Pearson	37.4593	<.0001*

Launch the Categorical Platform

Launch the Categorical Platform by selecting **Analyze** > **Consumer Research** > **Categorical**.

Figure 3.3 Categorical Platform Launch Window



Categorical responses in different situations

Select Columns

39 Columns

- Response ID
- Gender
- Birth Year
- Single Status
- School Age Children
- Age in Years
- Age Group
- Years at Current Employer
- Employee Tenure
- Years in Current Position
- Position Tenure
- Salary
- Salary Group
- Job Satisfaction
- I am working on my career

Grouping Option: Combinations

□ Unique Occurrences within ID
□ Count Missing Responses
□ Order Response Levels High to Low
□ Shorten Labels
□ Include Responses Not in Data

For multiple Grouping Columns, choose Grouping option

The launch window includes tabs for a variety of response roles (Simple, Related, and Multiple) and a Structured tab where you can create your own structured responses. The following sections describe the different response types and effects.

Response Roles

Use the response roles buttons within the tabs to choose selected columns as responses with specified roles. You can also drag column names to the response list. The response roles are summarized in Table 3.2.

Simple Tab

The default tab, Simple, contains a single button, Responses. This is appropriate for all basic analyses that do not have a special structure. You can drag column names from the Select Columns list to the Response list, or you can select columns and then click **Responses**. If a column has a Multiple Response column property, JMP automatically changes the handling of the column to recognize this property.

Related Tab

The Related tab contains a set of response columns that all have the same type of categories in them:

Aligned Responses Performs the analysis like the default analysis, but shows the results more compactly by aligning the analyses side-by-side into one larger table.

Repeated Measures Indicates that the columns reflect responses made by the same individual at different times, and you are interested in the changes between the times.

Rater Agreement Is useful when each column is a rating for the same question, but by different individuals (raters) and you want to study how much the raters agree on their responses.

Multiple Tab

The Multiple tab is for multiple responses; when a question can involve checking off more than one choice. There are a variety of ways of storing multiple response data, so there are several buttons in the tab to accommodate the various means:

Multiple Response Means that you have several columns acting like fill-in-the-blank columns to specify the multiple responses.

Multiple Response by ID Indicates that you have several rows in a table corresponding to the multiple responses in one column, and the individuals are identified by an ID column.

Multiple Delimited Signifies that you have one column that has several responses in it separated by a comma.

Indicator Group Denotes that there is a column for each possible response, and each column is an indicator (for example, it has only two values, like 0 or 1).

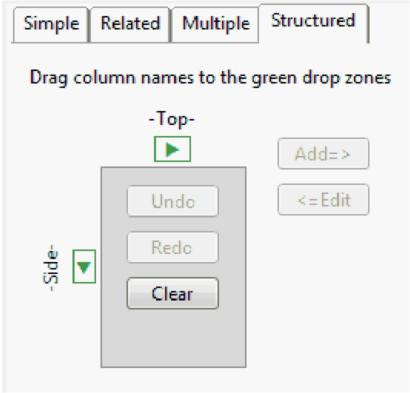
Response Frequencies Also has a column for each possible response, but has frequency counts instead of an indicator.

Free Text Is used for comment fields where the analysis counts the frequency of each word used. Free Text gives word counts in both word order and frequency order, and the rate of non-empty text. For more information about Free Text, refer to "Free Text Report Options" on page 54.

Structured Tab

The Structured tab enables you to construct complex tables of descriptive statistics by dragging column names into green icon drop zones to create side-by-side and nested results. You can nest a variable within or beside another variable according to the structure that you want for the top and side of the table. Continue to drag columns, either beside or nested within another column, to specify the structure. For more information about structured reports, refer to "Structured Report Options" on page 55.

Figure 3.4 Structured Tab



To create a structured table:

1. Drag a column name to the green drop zone at the **Top** or **Side** of the table.
2. You can then add more variables to the side or under another variable by dragging a column name to the appropriate drop zone.

   To remove a selection, click the selection and then select **Undo**.

   To add a selection back that you just removed, select **Redo**.

   To clear all of the selections, select **Clear**.
3. When you are finished creating the table, click **Add=>** to add the variables to the response list.

   To make a revision once you have added your selection to the response list, click the selection and then select **<=Edit**.
4. Complete the remainder of the launch window as necessary and click **OK**.

   The Categorical report window appears.
5. Should you want to make a change to the table, select **Relaunch Dialog** from the Categorical red triangle menu. The launch window reappears where you can make edits to your selections.

A few guidelines with Structured effects:

- Structured always assumes that the innermost terms on the Side are responses, and that all other terms are sample grouping factors.
- You can analyze multiple response terms in the form of delimited multiple response columns, but you must indicate that it is a multiple response by having a Multiple Response column property. Use the **Col Info** window to add this property, as needed.

Table 3.2 Response Roles

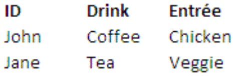
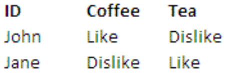
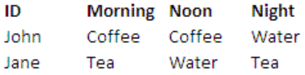
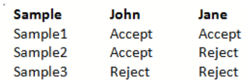
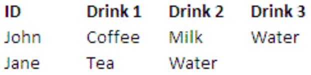
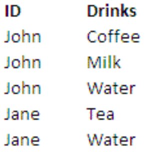
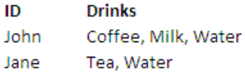
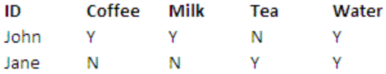
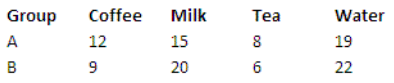
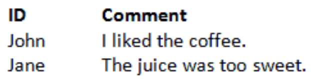
Response Role	Description	Example Data
Simple Tab		
Responses	Separate responses are in each column, resulting in a separate analysis for each column.	 **ID** **Drink** **Entrée** John Coffee Chicken Jane Tea Veggie
Related Tab		
Aligned Responses	Responses share common categories across columns, resulting in better-organized reports.	 **ID** **Coffee** **Tea** John Like Dislike Jane Dislike Like
Repeated Measures	Aligned responses from an individual across different times or situations.	 **ID** **Morning** **Noon** **Night** John Coffee Coffee Water Jane Tea Water Tea
Rater Agreement	Aligned responses from different raters evaluating the same unit, to study agreement across raters.	 **Sample** **John** **Jane** Sample1 Accept Accept Sample2 Accept Reject Sample3 Reject Reject
Multiple Tab		
Multiple Response	Aligned responses, where multiple responses are entered across several columns, but treated as one grouped response.	 **ID** **Drink 1** **Drink 2** **Drink 3** John Coffee Milk Water Jane Tea Water
Multiple Response by ID	Multiple responses across rows that have the same ID values.	 **ID** **Drinks** John Coffee John Milk John Water Jane Tea Jane Water
Multiple Delimited	Several responses in a single cell, separated by commas.	 **ID** **Drinks** John Coffee, Milk, Water Jane Tea, Water
Indicator Group	Binary responses across columns, like selected or deselected, yes or no, but all in a related group.	 **ID** **Coffee** **Milk** **Tea** **Water** John Y Y N Y Jane N N Y Y

Table 3.2 Response Roles *(Continued)*

Response Role	Description	Example Data
Response Frequencies	Columns containing frequency counts for each response level, all in a related group.	 Group Coffee Milk Tea Water A 12 15 8 19 B 9 20 6 22
Free Text	Counts the frequency of each word used in a comment field.	 ID Comment John I liked the coffee. Jane The juice was too sweet.
Structured Tab		
	Drag variables to green drop zones to create your own structured table.	

Cast Selected Columns into Roles

The lower right panel of the Launch window has the following options:

X, Grouping Category Defines sample groups to break the counts into. By default, it tabulates each combination of X values, but uses the Grouping Option below for other combinations.

Sample Size Defines the number of individual units in the group for which that frequency is applicable to, for multiple response roles with summarized data. For example, a Freq column might indicate 50 defects, where the sample size variable would reflect the defects for a batch of 100 units.

Freq Specifies the column containing frequency counts for each row for presummarized data.

ID Only required and used when Multiple Response by ID is selected.

By Identifies a variable to produce a separate analysis for each value that appears in the column.

Other Launch Window Options

Several launch options are presented in the lower left panel of the window that can be specified before the analysis. The options can also be selected later from the Categorical red triangle menu, and have the effect of rerunning the platform with the new setting. The default settings for some of the launch options can be changed in the Categorical red triangle menu. For more information, refer to “Set Preferences” on page 53.

Grouping Option Specifies whether you want to use the X columns individually or in a combination. Use this option only to specify more than one X (Grouping) column, and denote whether you want to treat the Xs one at a time, or in a fully nested grouping, or both. For example, if the X Columns are Region and Age Group, you can get separate tables for Response by Region and Response by Age Group (each individually) or get a nested table with each age group within each region (combinations), or both.

Combinations gives frequency results for combinations of the X variables.

Each Individually gives frequency results for each X variable individually.

Both gives frequency results for combinations of the X variables, and individually.

Unique Occurrences within ID Allows duplicate response levels within a subject to be counted only once. An ID variable must be specified.

The following options can be specified on the launch window as well as from the Categorical report red triangle menu. They are also available as Preference settings. For more information, refer to "Statistical Options" on page 39 and "Set Preferences" on page 53.

Count Missing Responses Changes the behavior to tabulate missing values as categories, while still excluding them from statistical comparisons. When you have missing values, this specifies whether you want to see them tabulated beside the nonmissing data, or just excluded. Missing values can be either standard (numeric NAN or character empty) or a code declared as missing with the column property Missing Value Codes.

Order Response Levels High to Low Changes the response order but keeps the X order from low to high. The default ordering is low to high. You can control the ordering with a column property (**Value Ordering**), but if you always want to see the high values first, then select this option. Often, ordered categories are ratings, and you want to see the positive ratings first.

Shorten Labels Shortens labels by removing common prefixes and suffixes. Sometimes surveys code a lengthy label that contain a common prefix or suffix. For example, "Occurred 5 to 10 times in the last year" might be a level, but the phrase "in the last year" is repeated for each value label, and you do not need to see it repeated in the report. This option trims the common prefixes and suffixes. It also changes multiple blanks into single blanks. The option only applies to value labels, not column names.

Include Responses Not in Data Includes a count for values that were not in the data. Sometimes when you conduct a survey and give choices, one of the choices is not selected. If you still want to see the choices that are not in the data, use this option. It determines the missing categories from the value labels in the column.

Supercategories

When ratings are involved in a data set (for example, a five point scale), you might want to know the percent of the responses in the top two or other subset of ratings. Such a group of ratings can be defined in the data through the column property, Supercategories.

The term *Supercategories* refers to the extra slots in a table to aggregate over groups of categories. The Supercategories property supports four keywords: Group, Mean, Std Dev, and All. Mean and Std Dev calculate statistics for value scores, and All aggregates across all levels. For example, a “Top Two” Group supercategory could aggregate the two top categories in a response, as specified. Although we support Mean, Std Dev, and All, we do not recommend using them because they are available as built-in statistics as well as supercategories.

Create Supercategories by selecting a column and then **Column Info > Column Properties > Supercatagories**. On the column properties window, select a column, enter a **Supercategory Name** and click **Add**. You can hide the results by selecting **Hide** from the red triangle menu. The option suppresses the category and only reports the group.

Supercategories can be scripted in an expression, such as:

```
{Group( "Top Two", {8, 9} ), All, Mean, StdDev}
```

Supercategories can also be specified in a Categorical launch command, where the properties are listed inside parentheses after the column name:

```
Categorical(Supercategories(Y({Group("Top Two",{4,5}),All})),…
```

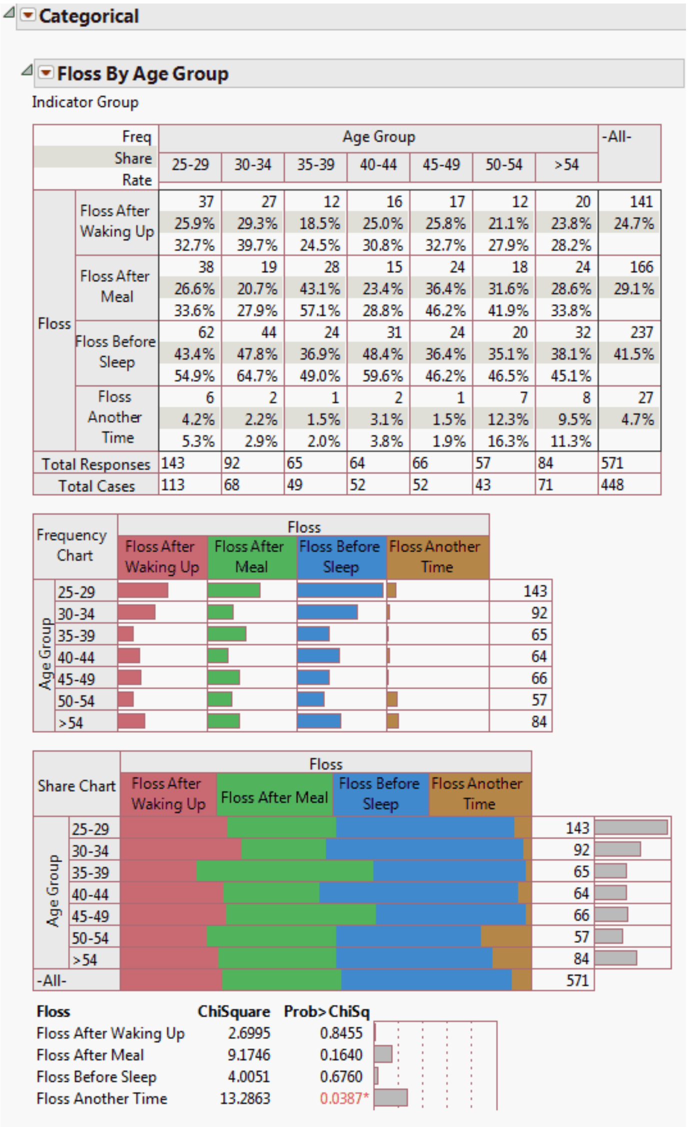
Supercategory support has been added for all response effects except Repeated Measures and Rater Agreement. Some response effects do not support Mean and Std Dev slots, because they do not have a natural score.

The Categorical Report

The Categorical platform produces a report with several tables and bar charts depending on your selections. You might or might not see all of the following options depending on which response type and options you selected. Frequencies, Share of Responses, and Rate Per Case appear in a single table by default. A Share Chart also appears by default. You can select to view a Frequency Chart or Transposed Frequency Chart.

You can select to view or hide each option (**Frequencies**, **Share of Responses**, **Rate Per Case**, **Share Chart**, **Frequency Chart**, or **Transposed Freq Chart**) from the Categorical red triangle menu. Data from Consumer Preferences.jmp is displayed in Figure 3.5.

Figure 3.5 The Categorical Report



Categorical

Floss By Age Group

Indicator Group

Freq Share Rate		Age Group							-All-
		25-29	30-34	35-39	40-44	45-49	50-54	>54	
Floss	Floss After Waking Up	37 25.9% 32.7%	27 29.3% 39.7%	12 18.5% 24.5%	16 25.0% 30.8%	17 25.8% 32.7%	12 21.1% 27.9%	20 23.8% 28.2%	141 24.7%
	Floss After Meal	38 26.6% 33.6%	19 20.7% 27.9%	28 43.1% 57.1%	15 23.4% 28.8%	24 36.4% 46.2%	18 31.6% 41.9%	24 28.6% 33.8%	166 29.1%
	Floss Before Sleep	62 43.4% 54.9%	44 47.8% 64.7%	24 36.9% 49.0%	31 48.4% 59.6%	24 36.4% 46.2%	20 35.1% 46.5%	32 38.1% 45.1%	237 41.5%
	Floss Another Time	6 4.2% 5.3%	2 2.2% 2.9%	1 1.5% 2.0%	2 3.1% 3.8%	1 1.5% 1.9%	7 12.3% 16.3%	8 9.5% 11.3%	27 4.7%
Total Responses		143	92	65	64	66	57	84	571
Total Cases		113	68	49	52	52	43	71	448

Floss	ChiSquare	Prob>ChiSq
Floss After Waking Up	2.6995	0.8455
Floss After Meal	9.1746	0.1640
Floss Before Sleep	4.0051	0.6760
Floss Another Time	13.2863	0.0387*

The topmost item in the table is a Frequency count (Freq), showing the frequency counts for each category with the total frequency (Total Responses) and total units (Total Cases) at the bottom of the table.

In this example, the number of responses and cases for each age group by the 7 segments are displayed. The last two rows show the total number of responses (571) and cases (448).

The Share of Responses (Share) is determined by dividing each count by the total number of responses. The number represents the percent of the response among all the responses in the sample (frequency divided by response total). This is either a column percentage or row percentage depending on whether your table has the responses on top or down the side (transposed).

For example, examine the second row of the table for Floss After Waking Up. The 37 responses who floss when they wake up were 25.9% of all responses (37/143*100).

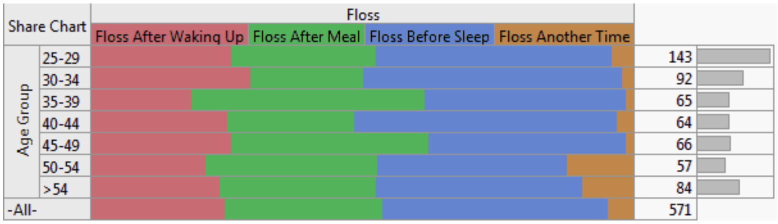
The Rate Per Case (Rate) divides each count in the frequency table by the total number of cases. If you have multiple responses per case (subject), there are two types of percentages; the rate per case is frequency as a percent of total cases, whereas the share of responses is the frequency as a percent of the total responses. Rate is available only for multiple responses.

For example, in the third row of the table (Floss After Waking Up), the 37 respondents are from 113 cases, making the rate per respondent 32.7%.

Share Chart

The Share Chart presents a divided bar chart. The bar length is proportional to the percentage of responses for each type. The bar chart on the right shows the number of responses.

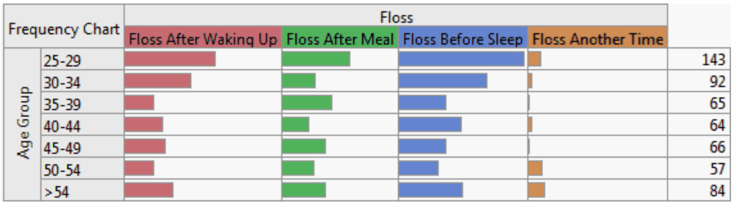
Figure 3.6 Share Chart



Frequency Chart

The Frequency Chart shows response frequencies. The bars reflect the frequency count on the same scale and the number of responses are displayed to the right. To view the frequency chart, select **Frequency Chart** from the Categorical red triangle menu.

Figure 3.7 Frequency Chart



The Transposed Freq Chart option produces a transposed version of the Frequency Chart. Marginal totals are given for each response, as opposed to each *X* variable.

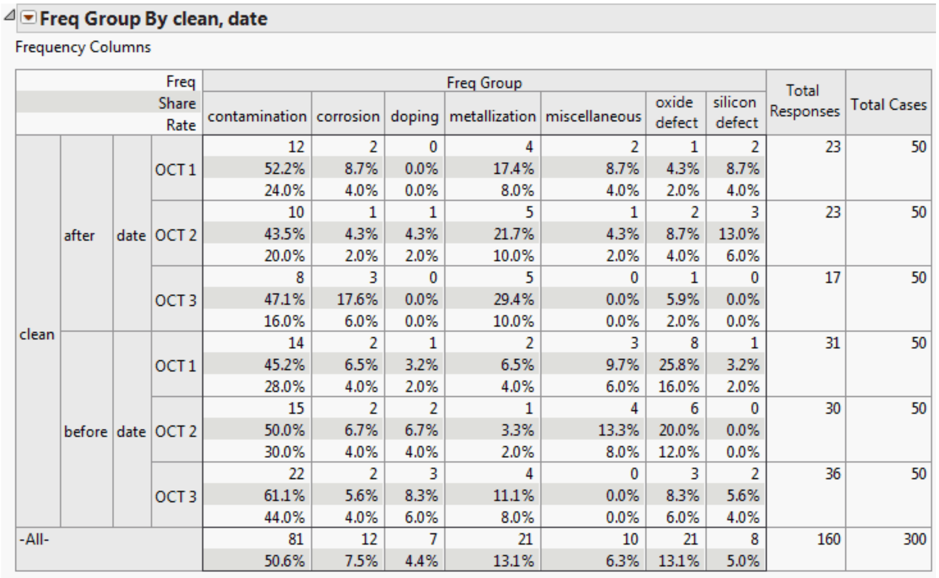
Categorical Platform Options

The Categorical red triangle menu provides commands that customize the appearance of the report and provide the means to test and compare your results. The following options appear in the menu depending on response roles and options selected. You might or might not view all of the options depending on your selections.

Report Options

The default report format is the **Crosstab** format, which gathers all three statistics for each sample and response together. The Crosstab format displays the responses on the top and the sample categories down the side, with multiple table elements together in each cell of the cross tabulation.

Figure 3.8 Crosstab Format



Freq Group By clean, date

Frequency Columns

			Freq	Freq Group							Total Responses	Total Cases
			Share Rate	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect		
clean	after	date	OCT 1	12 52.2% 24.0%	2 8.7% 4.0%	0 0.0% 0.0%	4 17.4% 8.0%	2 8.7% 4.0%	1 4.3% 2.0%	2 8.7% 4.0%	23	50
			OCT 2	10 43.5% 20.0%	1 4.3% 2.0%	1 4.3% 2.0%	5 21.7% 10.0%	1 4.3% 2.0%	2 8.7% 4.0%	3 13.0% 6.0%	23	50
			OCT 3	8 47.1% 16.0%	3 17.6% 6.0%	0 0.0% 0.0%	5 29.4% 10.0%	0 0.0% 0.0%	1 5.9% 2.0%	0 0.0% 0.0%	17	50
	before	date	OCT 1	14 45.2% 28.0%	2 6.5% 4.0%	1 3.2% 2.0%	2 6.5% 4.0%	3 9.7% 6.0%	8 25.8% 16.0%	1 3.2% 2.0%	31	50
			OCT 2	15 50.0% 30.0%	2 6.7% 4.0%	2 6.7% 4.0%	1 3.3% 2.0%	4 13.3% 8.0%	6 20.0% 12.0%	0 0.0% 0.0%	30	50
			OCT 3	22 61.1% 44.0%	2 5.6% 4.0%	3 8.3% 6.0%	4 11.1% 8.0%	0 0.0% 0.0%	3 8.3% 6.0%	2 5.6% 4.0%	36	50
-All-				81 50.6%	12 7.5%	7 4.4%	21 13.1%	10 6.3%	21 13.1%	8 5.0%	160	300

The Crosstab format has a transposed version, **Crosstab Transposed**, which is useful when there are a lot of response categories but not a lot of samples. Crosstab Transposed displays the responses down the side and the sample categories across the top, with multiple table elements together in each cell.

The Structured analysis always uses the Crosstab Transposed form, but in a more complex arrangement. The Free Text analysis has its own specialized reports. For more information about Free Text, refer to "Free Text Report Options" on page 54.

The **Legend** displays or hides the legend for the response column on the Share Chart.

Statistical Options

The main question of interest in any table is whether shares or rates vary from sample group to sample group, and specifically which sample groups are significantly different from which other sample groups.

There are two families of tests and comparisons, which correspond to single category responses and multiple responses:

- With single responses, a given response is just one response category, and the question is whether the share of responses is different across sample categories.
- With multiple responses, each individual can select several categories, and the question is whether the rate is different across sample categories.

Single responses are tested with a chi-square test of homogeneity. However, there are two types of this test: the Likelihood Ratio Chi-square and the Pearson Chi-square. It is a matter of personal preference and training which one you prefer. An option, Chi-square Test Choices, on the Categorical red triangle menu, enables you to show one or the other, or both. For more information, refer to "Test Options" on page 52.

For multiple responses, each response is treated in a separate account, with a probability of the count for each subject as a Poisson distribution (allowing for multiples of the same category). Each response is tested to determine whether the parameters are the same across sample categories.

The following options appear in the Categorical red triangle menu depending on context:

Table 3.3 Categorical Platform Commands

	Command	Supported Response Contexts	Question	Details
These require multiple response data	Test Response Homogeneity	• Responses • Aligned Responses • Repeated Measures • Response Frequencies (if no Sample Size) • Structured	Are the probabilities across the response categories the same across sample categories?	Marginal Homogeneity (Independence) Test, both Pearson and Chi-square likelihood ratio chi-square. For more information, refer to "Test Response Homogeneity" on page 43.
	Test Each Response	• Multiple Response • Multiple Response by ID (with Sample Size) • Multiple Delimited • Response Frequencies with Sample Size • Structured	For each response category, are the rates the same across sample categories?	Poisson regression on sample for each defect frequency. For more information, refer to "Test Each Response" on page 44.

Table 3.3 Categorical Platform Commands *(Continued)*

Command	Supported Response Contexts	Question	Details
Agreement Statistic	Rater Agreement	How closely do raters agree, and is the lack of agreement symmetrical?	Kappa for agreement, Bowker and McNemar for symmetry. For more information, refer to "Rater Agreement" on page 47.
Transition Report	Repeated Measures	How have the categories changed across time?	Transition counts and rates matrices. For more information, refer to "Repeated Measures" on page 48.
Cell Chisq	Responses	How do I further analyze the results to obtain more information?	For more information, refer to "Cell Chisq" on page 45.
Compare Each Sample	• Responses • Aligned Responses • Repeated Measures • Response Frequencies (if no Sample Size) • Structured	Do levels of the response category differ significantly?	For more information, refer to "Compare Each Sample" on page 49.
Compare Each Cell	• Single and Multiple Responses • Structured	Do pairs of levels within the two response categories differ significantly?	For more information, refer to "Compare Each Cell" on page 50.

Table 3.3 Categorical Platform Commands *(Continued)*

Command	Supported Response Contexts	Question	Details
Test Options	• ChiSquare Test Choices • Show Warnings • Order by Significance • Hide Nonsignificant	How do I further analyze the results to obtain more information?	For more information, refer to "Test Options" on page 52.

There are a series of options that add more detail for each sample group. The options that appear depend on your selections and the details of your analysis:

Total Responses Shows the sum of the frequency counts for each sample group.

Total Cases For multiple response columns, shows the number of cases (subjects), which are different from the number of responses.

Total Cases Responding For multiple response columns used in Structured tables, counts each person who responded at least once. People who did not respond at all are not included.

Mean Score Calculates the response means, using the numeric categories, or value scores. This is enabled for columns that use numeric codes, or for categories that have a **Value Scores** property. To make the Mean Score interpretable, you can assign specific value scores in the Column Info window with the Value Scores column property. For more information and an example, refer to "Mean Score Example" on page 62.

Mean Score Comparison Compares (using a t-test) the means across sample groups, showing which groups are significantly different. For more information about the letter codes, refer to "Comparisons with Letters" on page 51.

Std Dev Score Calculates the standard deviation of the value scores.

Order by Mean Score Orders the mean score calculations. The option only appears when there are no X columns in the analysis.

Save Tables Saves the report to a new data table. For more information, refer to "Save Tables" on page 52.

Filter Filters data to specific groups or ranges. Opens the Local Data Filter panel allowing you to identify varying subsets of data. For more information, refer to *Using JMP*.

Contents Summary Collects all of the tests and mean scores into a summary at the top of the report with links to the associated item.

Format Elements Enables you to specify formats for Frequencies, Shares and Rates, and how zeros are displayed. By default, Frequencies are Fixed Dec with 7 Width and 0 Decimals and Shares and Rates are Percent with 6 Width and 1 Decimal.

Arrange in Rows Arranges the reports across the page as opposed to down. Enter the number of reports that you want to view across the window.

Set Preferences Enables you to set preferences for future launches and sessions. For more information, refer to "Set Preferences" on page 53.

Category Options Contains options (Grouping Option, Count Missing Response, Order Response Levels High to Low, Shorten Labels, and Include Responses Not in Data) that are also presented on the launch window that could be specified before the analysis. The options can also be selected here and have the effect of rerunning the platform with the new option setting. For more information, refer to "Other Launch Window Options" on page 33.

Force Crosstab Shading Forces shading on crosstab reports even if the preference is set to no shading.

Relaunch Dialog Enables you to return to the launch window and edit the specifications for a structured table. For more information, refer to "Structured Tab" on page 30.

Script Contains options that are available to all platforms. See *Using JMP*.

Test Response Homogeneity

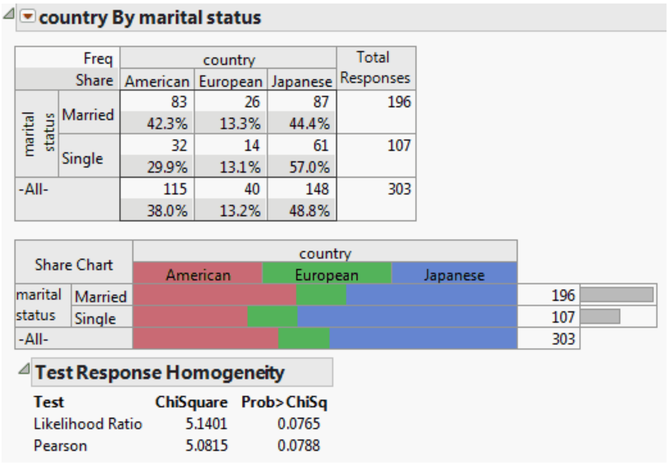
Test Response Homogeneity is the standard chi-square test (for single responses) across all sample categories. There is typically one categorical response variable and one categorical sample variable. Multiple sample variables are treated as a single variable.

The test is the chi-square test for marginal homogeneity of response patterns, testing that the response probabilities are the same across samples. This is equivalent to a test for independence when the sample category is like a response. There are two versions of this test, the Pearson form and the Likelihood Ratio form, both with chi-square statistics. The Test Options menu (ChiSquare Test Choices) is used to show or hide the Likelihood Ratio or Pearson tests. If Show Warnings is turned on, the report displays if the frequencies are too low to make good tests.

As an example:

1. Open the Car Poll.jmp sample data table.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select country and click **Responses** on the Simple tab.
4. Select marital status and click **X, Grouping Category**.
5. Click **OK**.
6. Select **Test Response Homogeneity** from the Categorical red triangle menu.

Figure 3.9 Test Response Homogeneity



country By marital status

Freq Share		American	European	Japanese	Total Responses
marital status	Married	83 42.3%	26 13.3%	87 44.4%	196
	Single	32 29.9%	14 13.1%	61 57.0%	107
-All-		115 38.0%	40 13.2%	148 48.8%	303

Test Response Homogeneity

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.1401	0.0765
Pearson	5.0815	0.0788

The Share Chart indicates that the married group is more likely to buy American cars, and the single group is more likely to buy Japanese cars, but the statistical test only shows a significance of 0.08. Therefore, the difference in response probabilities across marital status is not statistically significant at an alpha level of 0.05.

Test Each Response

Test Each Response is the standard chi-square test for multiple responses, with one test statistic for each response category. When there are multiple responses, each response category can be modeled separately. The question is whether the response rates are the same across samples. For each response category, we assume that the frequency count has a random Poisson distribution. The rate test is obtained using a Poisson regression (through generalized linear models) of the frequency per unit modeled by the sample categorical variable. The result is a likelihood ratio chi-square test of whether the rates are different across samples.

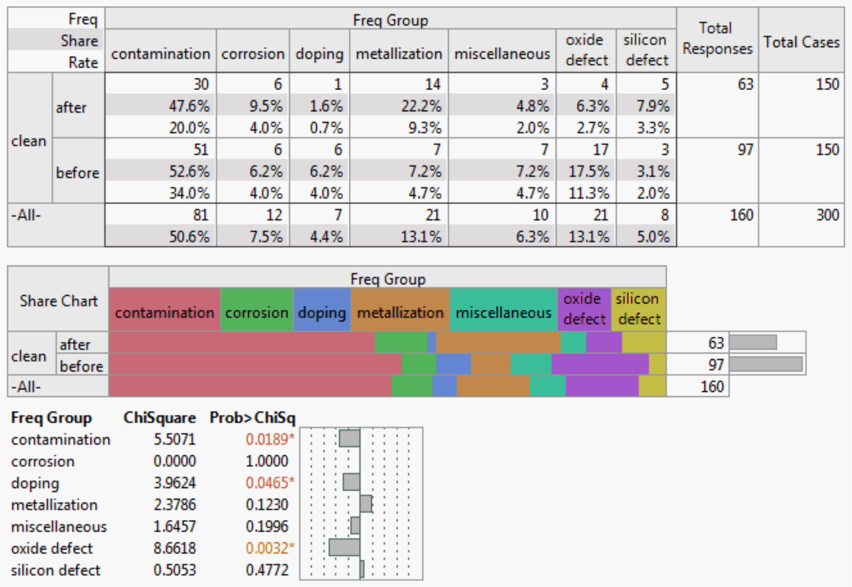
This test can also be done by the Pareto platform, as well as in the Generalized Linear Model personality of the Fit Model platform.

As an example:

1. Open the Failure3Freq.jmp sample data table in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select all of the defect columns and click **Response Frequencies** on the Multiple tab.
4. Select clean and click **X, Grouping Category** to compare the samples across the clean treatment variable.
5. Select SampleSize and click **Sample Size**.
6. Click **OK**.

7. Select **Test Each Response** from the Categorical red triangle menu.

Figure 3.10 Test Each Response



Freq Share Rate		Freq Group							Total Responses	Total Cases
		contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect		
clean	after	30 47.6% 20.0%	6 9.5% 4.0%	1 1.6% 0.7%	14 22.2% 9.3%	3 4.8% 2.0%	4 6.3% 2.7%	5 7.9% 3.3%	63	150
	before	51 52.6% 34.0%	6 6.2% 4.0%	6 6.2% 4.0%	7 7.2% 4.7%	7 7.2% 4.7%	17 17.5% 11.3%	3 3.1% 2.0%	97	150
-All-		81 50.6%	12 7.5%	7 4.4%	21 13.1%	10 6.3%	21 13.1%	8 5.0%	160	300

Freq Group	ChiSquare	Prob>ChiSq
contamination	5.5071	0.0189*
corrosion	0.0000	1.0000
doping	3.9624	0.0465*
metallization	2.3786	0.1230
miscellaneous	1.6457	0.1996
oxide defect	8.6618	0.0032*
silicon defect	0.5053	0.4772

For which defects are the rates significantly different across the clean treatments? The *p*-values show that oxide defect is the most significantly different, followed by contamination, then doping. The other defects are not significantly different with this amount of data.

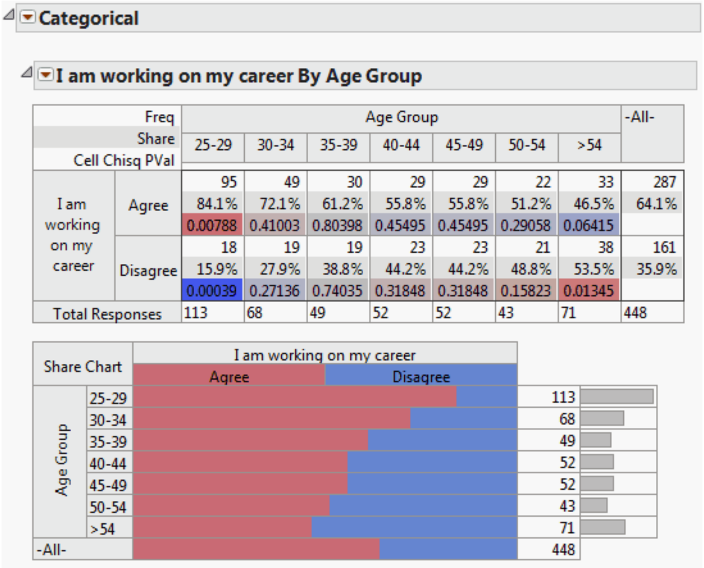
Cell Chisq

For single responses, Cell Chisq displays the cell-by-cell composition of the Pearson chi-square overall, and also shows which cells have relatively more (red) or less (blue) than expected if they were the same across sample categories. The value shown is the p-value for the chi-square. The color is bright when they are significant, and grayer when less significant, denoting visually where the significant differences are.

As an example:

1. Open the Consumer Preferences.jmp sample data table.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select I am working on my career and click **Responses** on the Simple tab.
4. Select Age Group and click **X, Grouping Category**.
5. Click **OK**.
6. Select **Crosstab Transposed** from the red triangle menu.
7. Select **Cell Chisq** from the red triangle menu.

Figure 3.11 Cell Chisq



Categorical

I am working on my career By Age Group

	Freq Share Cell Chisq PVal	Age Group 25-29	 30-34	 35-39	 40-44	 45-49	 50-54	 >54	-All-
I am working on my career	Agree	95 84.1% 0.00788	49 72.1% 0.41003	30 61.2% 0.80398	29 55.8% 0.45495	29 55.8% 0.45495	22 51.2% 0.29058	33 46.5% 0.06415	287 64.1%
	Disagree	18 15.9% 0.00039	19 27.9% 0.27136	19 38.8% 0.74035	23 44.2% 0.31848	23 44.2% 0.31848	21 48.8% 0.15823	38 53.5% 0.01345	161 35.9%
Total Responses		113	68	49	52	52	43	71	448

Relative Risk

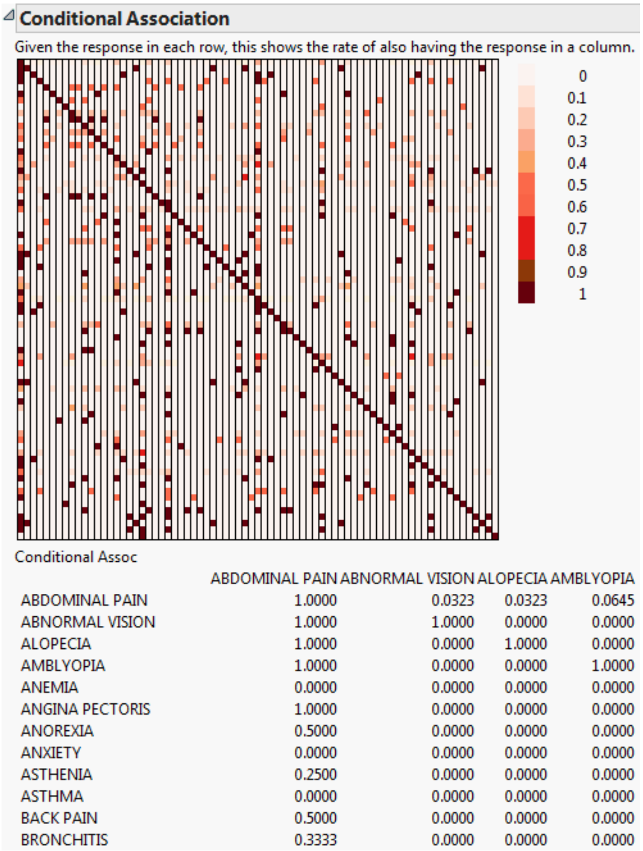
The Relative Risk option is used to compute relative risks for different responses. The risk of responses is computed for each level of the X, Grouping variable. The risks are compared to get a relative risk. This option is available only when the **Unique occurrences within ID** box is checked on the Categorical launch window.

A common application of this analysis is when the responses represent adverse events (side effects), and the X variable represents a treatment (drug versus placebo). The risk for getting each side effect is computed for both the drug and placebo. The relative risk is the ratio of the two risks.

Conditional Association

The **Conditional Association** option is used to compute the conditional probability of one response given a different response. A table and color map of the conditional probabilities are given. This option is available only when the **Unique occurrences within ID** box is checked on the Categorical launch window. A common application of this analysis is when the responses represent adverse events (side effects) from a drug. The computations represent the conditional probability of one side effect given the presence of another side effect. For AdverseR.jmp, given the response in each row, Figure 3.12 shows the rate of also having the response in a column. Figure 3.12 only displays a few variables in the table due to size constraints.

Figure 3.12 Conditional Association



Conditional Association

Given the response in each row, this shows the rate of also having the response in a column.

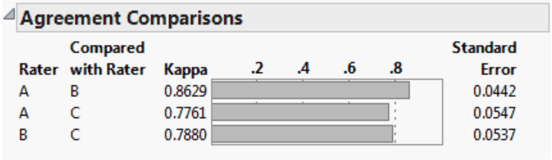
Conditional Assoc

	ABDOMINAL PAIN	ABNORMAL VISION	ALOPECIA	AMBLYOPIA
ABDOMINAL PAIN	1.0000	0.0323	0.0323	0.0645
ABNORMAL VISION	1.0000	1.0000	0.0000	0.0000
ALOPECIA	1.0000	0.0000	1.0000	0.0000
AMBLYOPIA	1.0000	0.0000	0.0000	1.0000
ANEMIA	0.0000	0.0000	0.0000	0.0000
ANGINA PECTORIS	1.0000	0.0000	0.0000	0.0000
ANOREXIA	0.5000	0.0000	0.0000	0.0000
ANXIETY	0.0000	0.0000	0.0000	0.0000
ASTHENIA	0.2500	0.0000	0.0000	0.0000
ASTHMA	0.0000	0.0000	0.0000	0.0000
BACK PAIN	0.5000	0.0000	0.0000	0.0000
BRONCHITIS	0.3333	0.0000	0.0000	0.0000

Rater Agreement

The Rater Agreement analysis answers the questions of how closely raters agree with one another and if the lack of agreement is symmetrical. For example, open Attribute Gauge.jmp. The Attribute Chart script runs the Variability Chart platform, which has a test for agreement among raters.

Figure 3.13 Agreement Comparisons

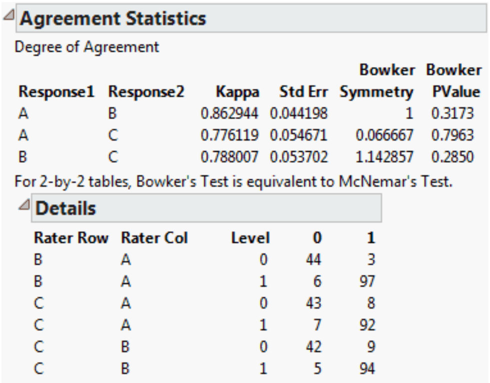


Agreement Comparisons

Rater	Compared with Rater	Kappa	Standard Error
A	B	0.8629	0.0442
A	C	0.7761	0.0547
B	C	0.7880	0.0537

Launch the Categorical platform and designate the three raters (A, B, and C) as **Rater Agreement** responses on the Related tab on the launch window. In the resulting report, you have a similar test for agreement that is augmented by a symmetry test that the lack of agreement is symmetric.

Figure 3.14 Agreement Statistics



Agreement Statistics

Degree of Agreement

				Bowker	Bowker
Response1	Response2	Kappa	Std Err	Symmetry	PValue
A	B	0.862944	0.044198	1	0.3173
A	C	0.776119	0.054671	0.066667	0.7963
B	C	0.788007	0.053702	1.142857	0.2850

For 2-by-2 tables, Bowker's Test is equivalent to McNemar's Test.

Details

Rater Row	Rater Col	Level	0	1
B	A	0	44	3
B	A	1	6	97
C	A	0	43	8
C	A	1	7	92
C	B	0	42	9
C	B	1	5	94

Repeated Measures

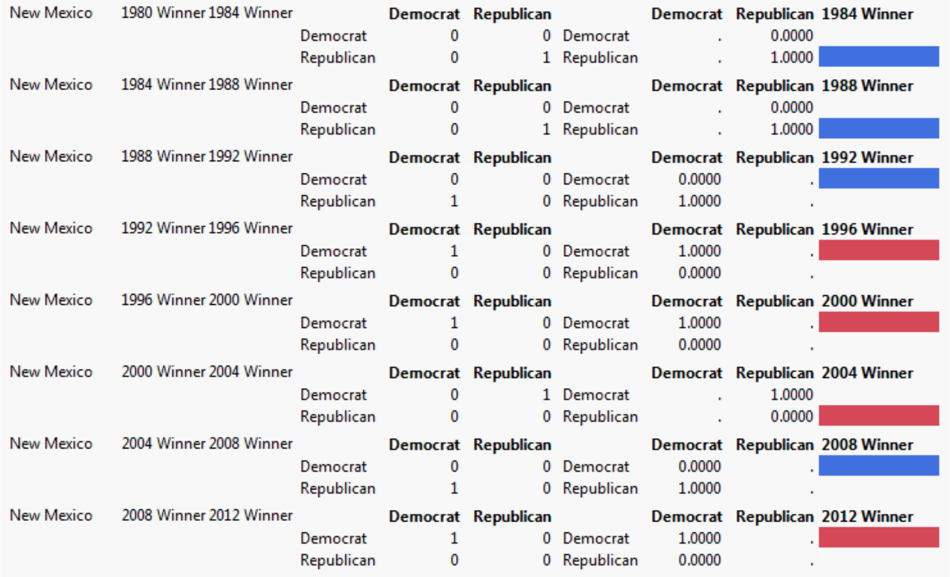
Repeated Measures declares that the columns reflect responses made by the same individual at different times, and you are interested in the changes between the times. Individual reports are displayed for each item, with a transition report at the end demonstrating the transition counts and rate matrices.

As an example:

1. Open the Presidential Elections.jmp sample data table.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select 1980 Winner through 2012 Winner and click **Repeated Measures** on the Related tab.
4. Select State and click **X, Grouping Category**.
5. Click **OK**.

Scroll through the responses to see how each State has voted over the years. Note that New Mexico has varied between Democratic and Republic over the years.

Figure 3.15 Repeated Measures



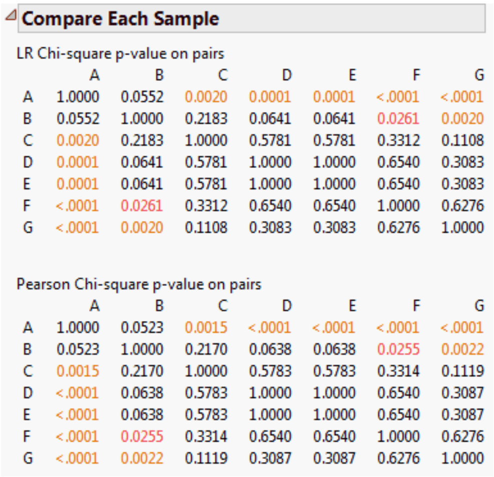
New Mexico	1980 Winner	1984 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1984 Winner**
			Democrat	0	0	Democrat	.	0.0000	
			Republican	0	1	Republican	.	1.0000	🟦
New Mexico	1984 Winner	1988 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1988 Winner**
			Democrat	0	0	Democrat	.	0.0000	
			Republican	0	1	Republican	.	1.0000	🟦
New Mexico	1988 Winner	1992 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1992 Winner**
			Democrat	0	0	Democrat	0.0000	.	🟦
			Republican	1	0	Republican	1.0000	.	
New Mexico	1992 Winner	1996 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1996 Winner**
			Democrat	1	0	Democrat	1.0000	.	🟥
			Republican	0	0	Republican	0.0000	.	
New Mexico	1996 Winner	2000 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2000 Winner**
			Democrat	1	0	Democrat	1.0000	.	🟥
			Republican	0	0	Republican	0.0000	.	
New Mexico	2000 Winner	2004 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2004 Winner**
			Democrat	0	1	Democrat	.	1.0000	
			Republican	0	0	Republican	.	0.0000	🟥
New Mexico	2004 Winner	2008 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2008 Winner**
			Democrat	0	0	Democrat	0.0000	.	🟦
			Republican	1	0	Republican	1.0000	.	
New Mexico	2008 Winner	2012 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2012 Winner**
			Democrat	1	0	Democrat	1.0000	.	🟥
			Republican	0	0	Republican	0.0000	.	

Compare Each Sample

For a given response, Compare Each Sample tests whether the response probability for each of its levels differs from the response probabilities for its other levels. In simple situations, the Compare Each Sample report consists of symmetric matrices of *p*-values, as shown in Figure 3.16.

In addition, a new row or column, entitled Compare, appears in the Crosstabs table. The Compare row is placed at the bottom of the table, or the Compare column is placed at the far right. (Whether a row or column is appended depends on whether Crosstab or Crosstab Transposed is specified.) The Compare row or column contains letter codes showing which sample categories differ significantly. For more information about the letter codes, refer to "Comparisons with Letters" on page 51.

Figure 3.16 Compare Each Sample



Compare Each Sample

LR Chi-square p-value on pairs

	A	B	C	D	E	F	G
A	1.0000	0.0552	0.0020	0.0001	0.0001	<.0001	<.0001
B	0.0552	1.0000	0.2183	0.0641	0.0641	0.0261	0.0020
C	0.0020	0.2183	1.0000	0.5781	0.5781	0.3312	0.1108
D	0.0001	0.0641	0.5781	1.0000	1.0000	0.6540	0.3083
E	0.0001	0.0641	0.5781	1.0000	1.0000	0.6540	0.3083
F	<.0001	0.0261	0.3312	0.6540	0.6540	1.0000	0.6276
G	<.0001	0.0020	0.1108	0.3083	0.3083	0.6276	1.0000

Pearson Chi-square p-value on pairs

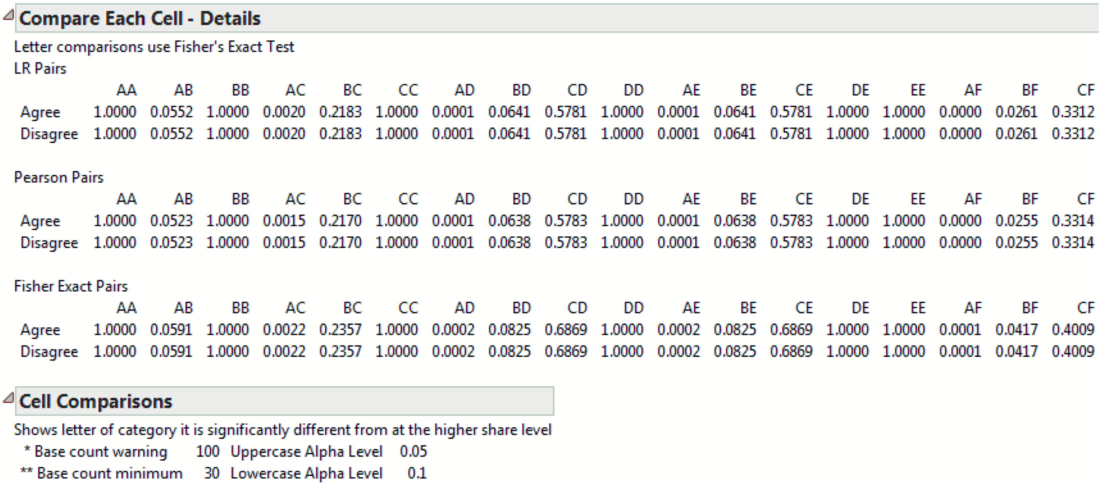
	A	B	C	D	E	F	G
A	1.0000	0.0523	0.0015	<.0001	<.0001	<.0001	<.0001
B	0.0523	1.0000	0.2170	0.0638	0.0638	0.0255	0.0022
C	0.0015	0.2170	1.0000	0.5783	0.5783	0.3314	0.1119
D	<.0001	0.0638	0.5783	1.0000	1.0000	0.6540	0.3087
E	<.0001	0.0638	0.5783	1.0000	1.0000	0.6540	0.3087
F	<.0001	0.0255	0.3314	0.6540	0.6540	1.0000	0.6276
G	<.0001	0.0022	0.1119	0.3087	0.3087	0.6276	1.0000

Compare Each Cell

For a given response and a given X variable, Compare Each Cell tests, for each level of the X variable, whether the response probabilities differ across the levels of the response. In other words, Compare Each Cell tests response probabilities across the cells in a given row of the Crosstabs table. The Compare Each Cell report gives *p*-values in a tabular format. The letters across the top indicate the response levels tested for the given level of the X variable. An example is shown in Figure 3.17.

In addition, when a cell differs significantly from other cells, a letter code is inserted into the appropriate cell in the Crosstabs table. For details on the letter codes and on their placement in cells, refer to “Comparisons with Letters” on page 51.

Figure 3.17 Compare Each Cell (cut off after column CF)



Compare Each Cell - Details

Letter comparisons use Fisher's Exact Test

LR Pairs

	AA	AB	BB	AC	BC	CC	AD	BD	CD	DD	AE	BE	CE	DE	EE	AF	BF	CF
Agree	1.0000	0.0552	1.0000	0.0020	0.2183	1.0000	0.0001	0.0641	0.5781	1.0000	0.0001	0.0641	0.5781	1.0000	1.0000	0.0000	0.0261	0.3312
Disagree	1.0000	0.0552	1.0000	0.0020	0.2183	1.0000	0.0001	0.0641	0.5781	1.0000	0.0001	0.0641	0.5781	1.0000	1.0000	0.0000	0.0261	0.3312

Pearson Pairs

	AA	AB	BB	AC	BC	CC	AD	BD	CD	DD	AE	BE	CE	DE	EE	AF	BF	CF
Agree	1.0000	0.0523	1.0000	0.0015	0.2170	1.0000	0.0001	0.0638	0.5783	1.0000	0.0001	0.0638	0.5783	1.0000	1.0000	0.0000	0.0255	0.3314
Disagree	1.0000	0.0523	1.0000	0.0015	0.2170	1.0000	0.0001	0.0638	0.5783	1.0000	0.0001	0.0638	0.5783	1.0000	1.0000	0.0000	0.0255	0.3314

Fisher Exact Pairs

	AA	AB	BB	AC	BC	CC	AD	BD	CD	DD	AE	BE	CE	DE	EE	AF	BF	CF
Agree	1.0000	0.0591	1.0000	0.0022	0.2357	1.0000	0.0002	0.0825	0.6869	1.0000	0.0002	0.0825	0.6869	1.0000	1.0000	0.0001	0.0417	0.4009
Disagree	1.0000	0.0591	1.0000	0.0022	0.2357	1.0000	0.0002	0.0825	0.6869	1.0000	0.0002	0.0825	0.6869	1.0000	1.0000	0.0001	0.0417	0.4009

Cell Comparisons

Shows letter of category it is significantly different from at the higher share level

* Base count warning	100	Uppercase Alpha Level	0.05
** Base count minimum	30	Lowercase Alpha Level	0.1

Comparisons with Letters

The Compare Each Cell, Compare Each Sample, and Compare Mean Scores commands use a system of letters to identify sample categories. The first sample is “A”, the second “B”. For more than 26 samples, numbers are appended after the letters. The letters are shown in the sample headings when a comparison command is turned on.

If two sample categories are significantly different, the letter of the sample with greater share is placed into the comparison cell of the other category that is significantly different. To find out if a given sample group, for example “B”, is significantly different, then you have to look both in the comparison cell for column “B” for other letters, and also in all the other cells across the sample groups for a “B”.

Lowercase letters are also used for comparisons that are slightly less significant, according to Table 3.4. These comparisons suffer when the count for that sample group (the Base Count) is small, and asterisks start to appear in the comparison cells to warn you.

The comparison features are controlled by four options set in Preferences or through a script. For more information, refer to “Set Preferences” on page 53.

Table 3.4 Letter Comparisons

Uppercase alpha level	0.05	The significance level for which uppercase letters show differences.
Lowercase alpha level	0.10	The significance level for which lowercase letters show differences.
Base Count minimum	≤ 29	The count for a sample that leads to a ** warning.

Table 3.4 Letter Comparisons *(Continued)*

Base Count warning	30 to 99	The count for a sample that leads to a * warning.

Test Options

The Test Options menu on the Categorical red triangle menu has the following options depending on your selections:

ChiSquare Test Choices Single responses are tested with a chi-square test of homogeneity; either the Likelihood Ratio Chi-square or the Pearson Chi-square, or both. Options are: **Both LR and Pearson**, **LR Only**, or **Pearson Only**. You can set an option in Preferences.

Show Warnings Shows warnings for chi-square tests related to small sample sizes.

Order by Significance Reorders the reports so that the most significant reports are at the top. This option only applies to reports with one homogeneity test.

Hide Nonsignificant Suppresses reports that are deemed non-significant. This option only applies to reports with one homogeneity test.

Save Tables

The Save Tables menu on the Categorical red triangle menu has the following options depending on your selections:

Save Frequencies Saves the Frequency report to a new data table, without the marginal totals.

Save Share of Responses Saves the Share of Responses report to a new data table, without the marginal totals.

Save Rate Per Case Saves the Rate Per Case report to a new data table, without the marginal totals.

Save Transposed Frequencies Saves the Transposed Freq Chart report to a new data table, without the marginal totals.

Save Transposed Share of Responses Saves a transposed version of the Share of Responses report to a new data table

Save Transposed Rate Per Case Saves a transposed version of the Rate Per Case report to a new data table.

Save Test Rates Saves the results of the Test Each Response option to a new data table.

Save Test Homogeneity Saves the results of the Test Response Homogeneity option to a new data table.

Save Excel File Creates a Microsoft Excel spreadsheet with the structure of the crosstab-format report. The option maps all of the tables to one sheet, with the response categories as rows, the samples as columns, sharing the headings for samples across multiple tables. When there are multiple elements in each table cell, you have the option to make them multiple or single cells in Microsoft Excel.

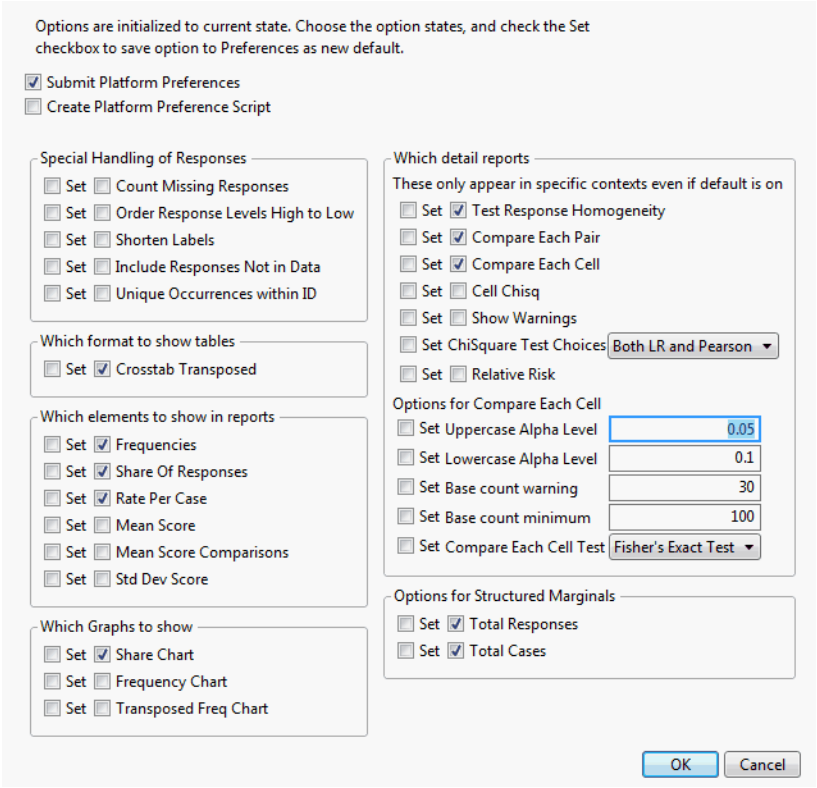
Set Preferences

You can specify settings and set preferences within the Categorical platform. Several options are available on the launch window and can be specified before the analysis. Some of the options can also be selected from the Categorical red triangle menu, and have the effect of rerunning the analysis with the new setting.

The options are initialized to the current state. Select the appropriate options and select either **Submit Platform Preferences** or **Create Platform Preference Script** to submit the options to your preferences as the new default. When the Categorical platform is launched, the preferences associated with the current preference set are enacted.

Preferences can be administered and shared through a script. The best way to share a preference set widely is to create an add-in, so that if the preference settings are reset to the initial state, the add-in could restore the preferred set.

Figure 3.18 Set Preferences Window



Options are initialized to current state. Choose the option states, and check the Set checkbox to save option to Preferences as new default.

☑ Submit Platform Preferences
□ Create Platform Preference Script

Special Handling of Responses
□ Set □ Count Missing Responses
□ Set □ Order Response Levels High to Low
□ Set □ Shorten Labels
□ Set □ Include Responses Not in Data
□ Set □ Unique Occurrences within ID

Which format to show tables
□ Set ☑ Crosstab Transposed

Which elements to show in reports
□ Set ☑ Frequencies
□ Set ☑ Share Of Responses
□ Set ☑ Rate Per Case
□ Set □ Mean Score
□ Set □ Mean Score Comparisons
□ Set □ Std Dev Score

Which Graphs to show
□ Set ☑ Share Chart
□ Set □ Frequency Chart
□ Set □ Transposed Freq Chart

Which detail reports
These only appear in specific contexts even if default is on
□ Set ☑ Test Response Homogeneity
□ Set ☑ Compare Each Pair
□ Set ☑ Compare Each Cell
□ Set □ Cell Chisq
□ Set □ Show Warnings
□ Set ChiSquare Test Choices Both LR and Pearson
□ Set □ Relative Risk

Options for Compare Each Cell
□ Set Uppercase Alpha Level 0.05
□ Set Lowercase Alpha Level 0.1
□ Set Base count warning 30
□ Set Base count minimum 100
□ Set Compare Each Cell Test Fisher's Exact Test

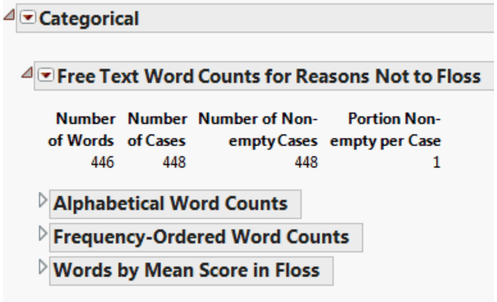
Options for Structured Marginals
□ Set ☑ Total Responses
□ Set ☑ Total Cases

Free Text Report Options

Free Text is used for comment fields where the analysis counts the frequency of each word used. Free Text gives word counts in both word order and frequency order, and the rate of non-empty text. The following example uses the Consumer Preferences.jmp sample data table, which contains survey data relating to oral hygiene preferences. A comment field was included in the survey asking for reasons why the participant did not floss.

1. Open the Consumer Preferences.jmp sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select Reasons Not to Floss and click **Free Text** on the Multiple tab.
4. Click **OK**.
5. Select **Score Words by Column** from the red triangle menu and then select Floss. Click **OK**.

Figure 3.19 Free Text Report Example



Categorical

Free Text Word Counts for Reasons Not to Floss

Number of Words	Number of Cases	Number of Non-empty Cases	Portion Non-empty per Case
446	448	448	1

Alphabetical Word Counts

Frequency-Ordered Word Counts

Words by Mean Score in Floss

Figure 3.19 details the free text word counts the respondents included as reasons why they do not floss. From the analysis, you can determine the number of words, cases, non-empty cases, and portions of non-empty cases. You can also view the word counts alphabetically, in terms of frequency, or by the mean scores. There are more commands to further customize the analysis on the Free Text red triangle menu on the report:

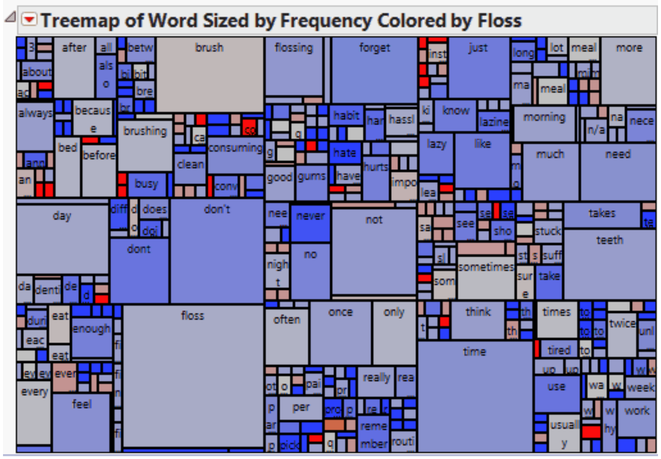
Score Words by Column Calculates for each word the average score for another column for the rows that the word appears. If you save a Microsoft Word table later, it will have these scores in the saved data table, plus a Treemap script (Figure 3.20) to colorize by these scores.

Save Indicators for Most Frequent Words Prompts you for the number of words to make indicators for, and then creates the indicator columns for the most frequent words in the data table indicating if that word appeared.

Save Word Table Creates a new table of all the words, their frequency, and the scores with respect to any of the columns scored.

Remove Removes the table from the report.

Figure 3.20 Treemap Example



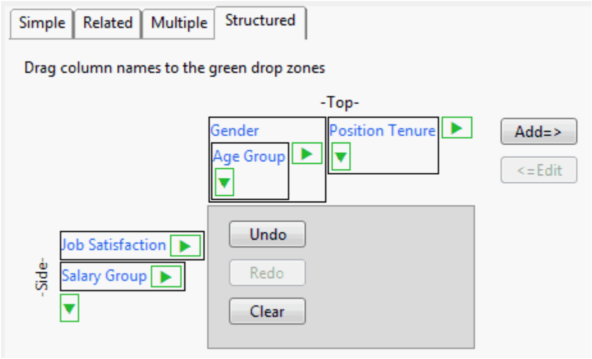
Treemap of Word Sized by Frequency Colored by Floss

Structured Report Options

The Structured tab enables you to construct complex tables of descriptive statistics by dragging column names into green icon drop zones to create side-by-side and nested results. The following example uses the Consumer Preferences.jmp sample data table. From this data, suppose that you wanted to compare job satisfaction and salary against gender by age group and position tenure.

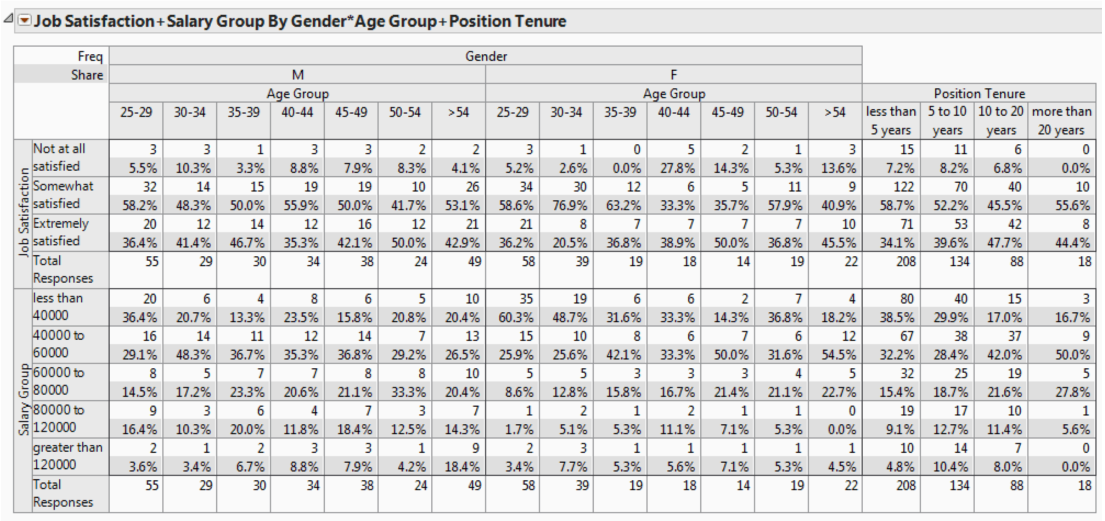
1. Open the Consumer Preferences.jmp sample data table.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select the Structured tab.
4. Drag Gender to the green drop zone at the **Top** of the table on the Structured tab.
5. Drag Age Group to the green drop zone just below Gender.
6. Drag Position Tenure to the green drop zone at the **Top** of the table next to Gender.
7. Drag Job Satisfaction to the green drop zone at the **Side** of the table.
8. Drag Salary Group to the green drop zone at the **Side** of the table under Job Satisfaction.

Figure 3.21 Structured Tab Report Setup



9. Click **Add=>**.
10. Click **OK**.

Figure 3.22 Structured Tab Report Example



Job Satisfaction+Salary Group By Gender*Age Group+Position Tenure

	Freq	Gender																	
	Share	M							F										
		Age Group							Age Group							Position Tenure			
		25-29	30-34	35-39	40-44	45-49	50-54	>54	25-29	30-34	35-39	40-44	45-49	50-54	>54	less than 5 years	5 to 10 years	10 to 20 years	more than 20 years
Job Satisfaction	Not at all satisfied	3 5.5%	3 10.3%	1 3.3%	3 8.8%	3 7.9%	2 8.3%	2 4.1%	3 5.2%	1 2.6%	0 0.0%	5 27.8%	2 14.3%	1 5.3%	3 13.6%	15 7.2%	11 8.2%	6 6.8%	0 0.0%
	Somewhat satisfied	32 58.2%	14 48.3%	15 50.0%	19 55.9%	19 50.0%	10 41.7%	26 53.1%	34 58.6%	30 76.9%	12 63.2%	6 33.3%	5 35.7%	11 57.9%	9 40.9%	122 58.7%	70 52.2%	40 45.5%	10 55.6%
	Extremely satisfied	20 36.4%	12 41.4%	14 46.7%	12 35.3%	16 42.1%	12 50.0%	21 42.9%	21 36.2%	8 20.5%	7 36.8%	7 38.9%	7 50.0%	7 36.8%	10 45.5%	71 34.1%	53 39.6%	42 47.7%	8 44.4%
	Total Responses	55	29	30	34	38	24	49	58	39	19	18	14	19	22	208	134	88	18
Salary Group	less than 40000	20 36.4%	6 20.7%	4 13.3%	8 23.5%	6 15.8%	5 20.8%	10 20.4%	35 60.3%	19 48.7%	6 31.6%	6 33.3%	2 14.3%	7 36.8%	4 18.2%	80 38.5%	40 29.9%	15 17.0%	3 16.7%
	40000 to 60000	16 29.1%	14 48.3%	11 36.7%	12 35.3%	14 36.8%	7 29.2%	13 26.5%	15 25.9%	10 25.6%	8 42.1%	6 33.3%	7 50.0%	6 31.6%	12 54.5%	67 32.2%	38 28.4%	37 42.0%	9 50.0%
	60000 to 80000	8 14.5%	5 17.2%	7 23.3%	7 20.6%	8 21.1%	8 33.3%	10 20.4%	5 8.6%	5 12.8%	3 15.8%	3 16.7%	3 21.4%	4 21.1%	5 22.7%	32 15.4%	25 18.7%	19 21.6%	5 27.8%
	80000 to 120000	9 16.4%	3 10.3%	6 20.0%	4 11.8%	7 18.4%	3 12.5%	7 14.3%	1 1.7%	2 5.1%	1 5.3%	2 11.1%	1 7.1%	1 5.3%	0 0.0%	19 9.1%	17 12.7%	10 11.4%	1 5.6%
	greater than 120000	2 3.6%	1 3.4%	2 6.7%	3 8.8%	3 7.9%	1 4.2%	9 18.4%	2 3.4%	3 7.7%	1 5.3%	1 5.6%	1 7.1%	1 5.3%	1 4.5%	10 4.8%	14 10.4%	7 8.0%	0 0.0%
	Total Responses	55	29	30	34	38	24	49	58	39	19	18	14	19	22	208	134	88	18

Figure 3.22 shows that the majority of both the male and female respondents were somewhat satisfied with their jobs, with the highest percentage of males being in the 25-29 age group, while the females were in the 30-34 age group. Most of those who were somewhat satisfied had been in their current position for less than 5 years.

The following options are available from the structured report's red triangle menu:

Show Letters Forces the table to display the column letter IDs, which usually come out automatically when you do a compare command.

Specify Comparison Groups Enables you to specify groups when the sample groups that you want to test and compare are not the same as the innermost term's structure. To use this option, you must look at the letter IDs, and then enter sets of letter IDs, separated by a slash, representing each group, separating multiple groups from each other by commas. For example, the default grouping might be "A/B/C, E/D/F", but you want to test A with E, B with D and C with F, so you specify the groups as "A/E, B/D, C/F". This determines which letters appear in the comparison fields. In addition, a summary report shows the overall tests for each column group.

Remove Removes the table from the report.

Additional Examples of the Categorical Platform

The following examples come from testing a fabrication line on three different occasions under two different conditions. Each set of operating conditions yielded 50 data points. Inspectors recorded the following types of defects:

- contamination
- corrosion
- doping
- metallization
- miscellaneous
- oxide defect
- silicon defect

Each unit could have several defects or even several defects of the same kind. We illustrate the data in a variety of different examples all within the Categorical platform.

Multiple Response

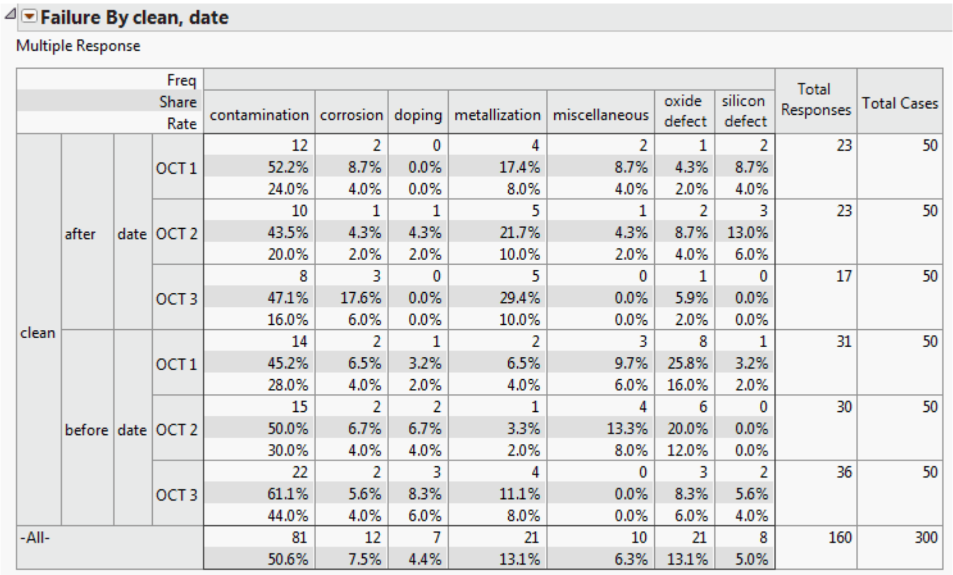
Suppose that the defects for each unit are entered via a web page, but because each unit rarely has more than three defect types, the form has three fields to enter any of the defect types for a unit, as in Failure3MultipleField.jmp.

1. Open the Failure3MultipleField.jmp sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select Failure1, Failure2, and Failure3 and click **Multiple Response** on the Multiple tab.

   These columns contain defect types and are the variables that you want to inspect.
4. Select clean and date and click **X, Grouping Category**.
5. Click **OK**.

Figure 3.23 lists failure types and counts for each failure type from the **Multiple Response** analysis.

Figure 3.23 Multiple Response Failures from Failure3MultipleField.jmp



Failure By clean, date

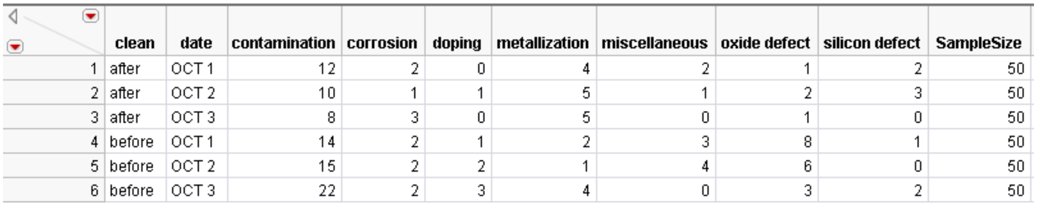
Multiple Response

			Freq Share Rate	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	Total Responses	Total Cases
clean	after	date	OCT 1	12	2	0	4	2	1	2	23	50
				52.2%	8.7%	0.0%	17.4%	8.7%	4.3%	8.7%		
				24.0%	4.0%	0.0%	8.0%	4.0%	2.0%	4.0%		
			OCT 2	10	1	1	5	1	2	3	23	50
				43.5%	4.3%	4.3%	21.7%	4.3%	8.7%	13.0%		
				20.0%	2.0%	2.0%	10.0%	2.0%	4.0%	6.0%		
			OCT 3	8	3	0	5	0	1	0	17	50
				47.1%	17.6%	0.0%	29.4%	0.0%	5.9%	0.0%		
				16.0%	6.0%	0.0%	10.0%	0.0%	2.0%	0.0%		
	before	date	OCT 1	14	2	1	2	3	8	1	31	50
				45.2%	6.5%	3.2%	6.5%	9.7%	25.8%	3.2%		
				28.0%	4.0%	2.0%	4.0%	6.0%	16.0%	2.0%		
			OCT 2	15	2	2	1	4	6	0	30	50
				50.0%	6.7%	6.7%	3.3%	13.3%	20.0%	0.0%		
				30.0%	4.0%	4.0%	2.0%	8.0%	12.0%	0.0%		
			OCT 3	22	2	3	4	0	3	2	36	50
				61.1%	5.6%	8.3%	11.1%	0.0%	8.3%	5.6%		
				44.0%	4.0%	6.0%	8.0%	0.0%	6.0%	4.0%		
-All-				81	12	7	21	10	21	8	160	300
				50.6%	7.5%	4.4%	13.1%	6.3%	13.1%	5.0%		

Response Frequencies

Suppose the data have columns containing frequency counts for each batch and a column showing the total number of units of the batch, as in Failure3Freq.jmp.

Figure 3.24 Failure3Freq.jmp Data Table



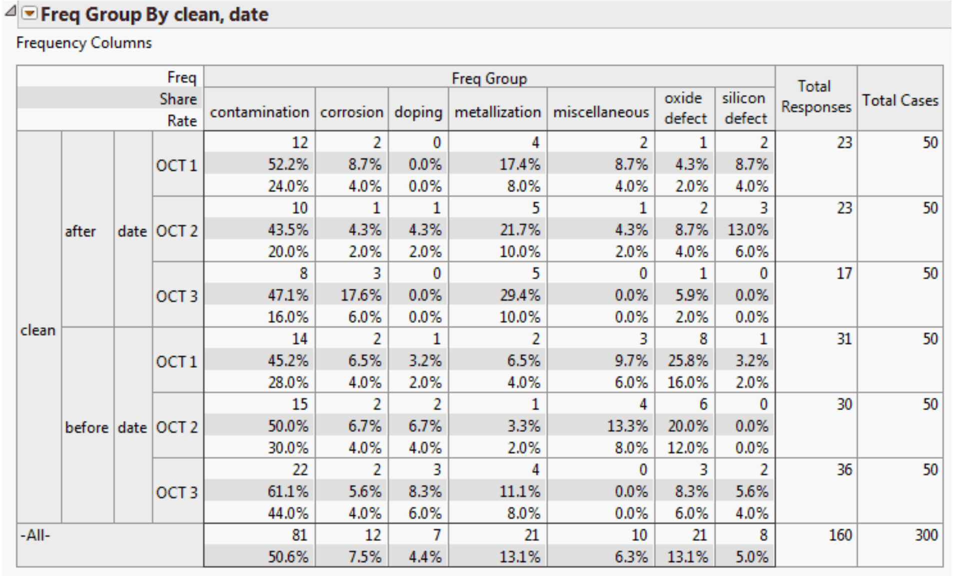
	clean	date	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	SampleSize
1	after	OCT 1	12	2	0	4	2	1	2	50
2	after	OCT 2	10	1	1	5	1	2	3	50
3	after	OCT 3	8	3	0	5	0	1	0	50
4	before	OCT 1	14	2	1	2	3	8	1	50
5	before	OCT 2	15	2	2	1	4	6	0	50
6	before	OCT 3	22	2	3	4	0	3	2	50

1. Open the Failure3Freq sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select the frequency variables (contamination, corrosion, doping, metallization, miscellaneous, oxide defect, silicon defect) and click **Response Frequencies** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category**.

5. Select Sample Size and click **Sample Size**.
6. Click **OK**.

The resulting output in Figure 3.25 shows a frequency count table, with a separate column for each of the seven batches. The last two columns show the total number of defects (**Total Responses**) and cases (**Total Cases**).

Figure 3.25 Defect Rate Output



Freq Group By clean, date

Frequency Columns

			Freq Share Rate	Freq Group							Total Responses	Total Cases
				contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect		
clean	after	date	OCT 1	12	2	0	4	2	1	2	23	50
				52.2%	8.7%	0.0%	17.4%	8.7%	4.3%	8.7%		
				24.0%	4.0%	0.0%	8.0%	4.0%	2.0%	4.0%		
			OCT 2	10	1	1	5	1	2	3	23	50
				43.5%	4.3%	4.3%	21.7%	4.3%	8.7%	13.0%		
				20.0%	2.0%	2.0%	10.0%	2.0%	4.0%	6.0%		
			OCT 3	8	3	0	5	0	1	0	17	50
				47.1%	17.6%	0.0%	29.4%	0.0%	5.9%	0.0%		
				16.0%	6.0%	0.0%	10.0%	0.0%	2.0%	0.0%		
	before	date	OCT 1	14	2	1	2	3	8	1	31	50
				45.2%	6.5%	3.2%	6.5%	9.7%	25.8%	3.2%		
				28.0%	4.0%	2.0%	4.0%	6.0%	16.0%	2.0%		
			OCT 2	15	2	2	1	4	6	0	30	50
				50.0%	6.7%	6.7%	3.3%	13.3%	20.0%	0.0%		
				30.0%	4.0%	4.0%	2.0%	8.0%	12.0%	0.0%		
			OCT 3	22	2	3	4	0	3	2	36	50
				61.1%	5.6%	8.3%	11.1%	0.0%	8.3%	5.6%		
				44.0%	4.0%	6.0%	8.0%	0.0%	6.0%	4.0%		
-All-				81	12	7	21	10	21	8	160	300
				50.6%	7.5%	4.4%	13.1%	6.3%	13.1%	5.0%		

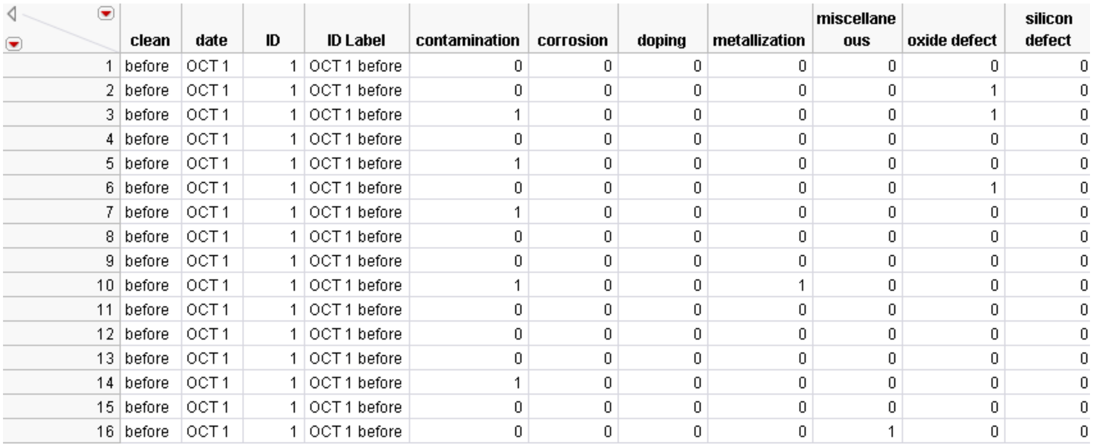
Each Frequency Group contains the following information:

- The total number of defects for each defect type. For example, after cleaning on Oct 1st, there were 12 contamination defects.
- The share of responses. For example, after cleaning on Oct 1st, the 12 contamination defects were (12/23) accounting for 52.2% of all defects.
- The rate per case. For example, after cleaning on Oct 1st, the 12 contamination defects are from 50 units (12/50) making the rate per unit 24%.

Indicator Group

In some cases, the data is not yet summarized, so there are individual records for each unit. We illustrate this situation with the data table, Failures3Indicators.jmp.

Figure 3.26 Failure3Indicators.jmp Data Table



	clean	date	ID	ID Label	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect
1	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
2	before	OCT 1	1	OCT 1 before	0	0	0	0	0	1	0
3	before	OCT 1	1	OCT 1 before	1	0	0	0	0	1	0
4	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
5	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
6	before	OCT 1	1	OCT 1 before	0	0	0	0	0	1	0
7	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
8	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
9	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
10	before	OCT 1	1	OCT 1 before	1	0	0	1	0	0	0
11	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
12	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
13	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
14	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
15	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
16	before	OCT 1	1	OCT 1 before	0	0	0	0	1	0	0

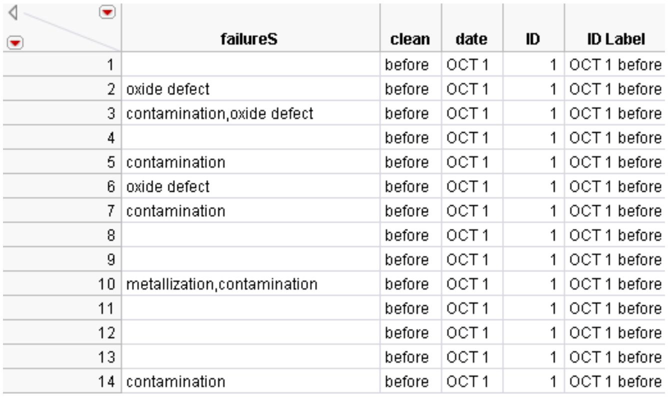
1. Open the Failures3Indicators.jmp sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select the defect columns (contamination, corrosion, doping, metallization, miscellaneous, oxide defect, silicon defect) and click **Indicator Group** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category**.
5. Click **OK**.

When you click **OK**, you get the same output as in the Response Group example (Figure 3.25).

Multiple Delimited

Suppose that an inspector entered the observed defects for each unit. The defects are listed in a single column, delimited by a comma, as in Failures3Delimited.jmp. Note in the partial data table, shown below, that some units did not have any observed defects, so the failureS column is empty.

Figure 3.27 Failure3Delimited.jmp Data Table



	failureS	clean	date	ID	ID Label
1		before	OCT 1	1	OCT 1 before
2	oxide defect	before	OCT 1	1	OCT 1 before
3	contamination,oxide defect	before	OCT 1	1	OCT 1 before
4		before	OCT 1	1	OCT 1 before
5	contamination	before	OCT 1	1	OCT 1 before
6	oxide defect	before	OCT 1	1	OCT 1 before
7	contamination	before	OCT 1	1	OCT 1 before
8		before	OCT 1	1	OCT 1 before
9		before	OCT 1	1	OCT 1 before
10	metallization,contamination	before	OCT 1	1	OCT 1 before
11		before	OCT 1	1	OCT 1 before
12		before	OCT 1	1	OCT 1 before
13		before	OCT 1	1	OCT 1 before
14	contamination	before	OCT 1	1	OCT 1 before

1. Open the Failures3Delimited.jmp sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select failureS and click **Multiple Delimited** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category**.
5. Select ID and click **ID**.
6. Click **OK**.

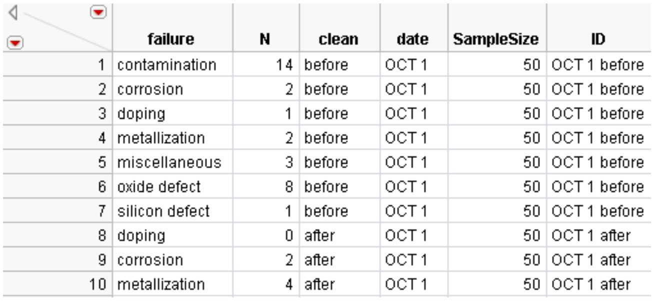
When you click **OK**, you get the same output as in Figure 3.25.

Note: If more than one delimited column is specified, separate analyses are produced for each column.

Multiple Response by ID

Suppose each failure type is a separate record, with an ID column that can be used to link together different defect types for each unit, as in Failure3ID.jmp.

Figure 3.28 Failure3ID.jmp Data Table



	failure	N	clean	date	SampleSize	ID
1	contamination	14	before	OCT 1	50	OCT 1 before
2	corrosion	2	before	OCT 1	50	OCT 1 before
3	doping	1	before	OCT 1	50	OCT 1 before
4	metallization	2	before	OCT 1	50	OCT 1 before
5	miscellaneous	3	before	OCT 1	50	OCT 1 before
6	oxide defect	8	before	OCT 1	50	OCT 1 before
7	silicon defect	1	before	OCT 1	50	OCT 1 before
8	doping	0	after	OCT 1	50	OCT 1 after
9	corrosion	2	after	OCT 1	50	OCT 1 after
10	metallization	4	after	OCT 1	50	OCT 1 after

1. Open the Failure3ID.jmp sample data table, located in the Quality Control folder.
2. Select **Analyze** > **Consumer Research** > **Categorical**.
3. Select failure and click **Multiple Response by ID** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category**.
5. Select SampleSize and click **Sample Size**.
6. Select N and click **Freq**.
7. Select ID and click **ID**.
8. Click **OK**.

When you click **OK**, you get the same output as in Figure 3.25.

Mean Score Example

You can calculate response means in your data using Value Scores. To make the Mean Score interpretable, you can assign specific value scores in the **Column Info** window with the **Value Scores** column property. For more information about column properties, refer to *Using JMP*.

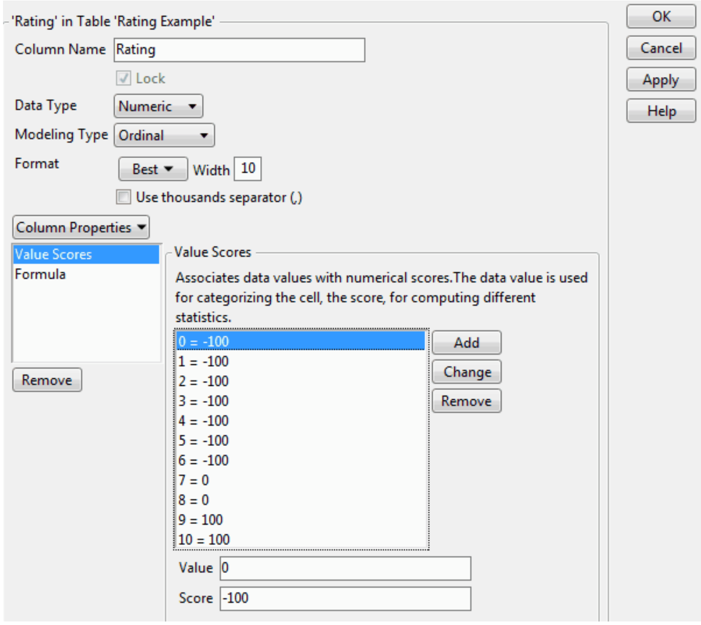
In this example, you can assign Value Scores to calculate the Net Promoter Score (Reichheld, HBR 2003), which summarizes an 11-level rating with a favorability score between -100 and 100. Anything with a value of 6 or below is regarded as a detractor.

1. Run the following script:

```
New Table("Rating Example",
  Add Rows(300),
  New Script("Categorical",Categorical(Responses(:Rating),Mean Score(1))),
  New Column("Rating",Numeric,Ordinal,
    Set Property(
      "Value Scores",
{0=-100,1=-100,2=-100,3=-100,4=-100,5=-100,6=-100,7=0,8=0,9=100,10=100}),
    Formula(Random Category(
        0.05,0,0.05,1,0.05,2,0.05,3,0.05,4,
        0.05,5,0.05,6,0.05,7,0.05,8,0.3,9,0.25,10)),
    Set Selected
 )
);
```

2. A data table with 300 rows of random rating data is created. Value scores were also defined for the Rating column. To view the scores, right-click the Rating column and select **Column Properties > Value Scores**.

Figure 3.29 Column Properties - Value Scores



'Rating' in Table 'Rating Example'

Column Name: Rating

☑ Lock

Data Type: Numeric

Modeling Type: Ordinal

Format: Best Width 10

□ Use thousands separator (,)

Column Properties

- Value Scores
- Formula

Remove

Value Scores

Associates data values with numerical scores.The data value is used for categorizing the cell, the score, for computing different statistics.

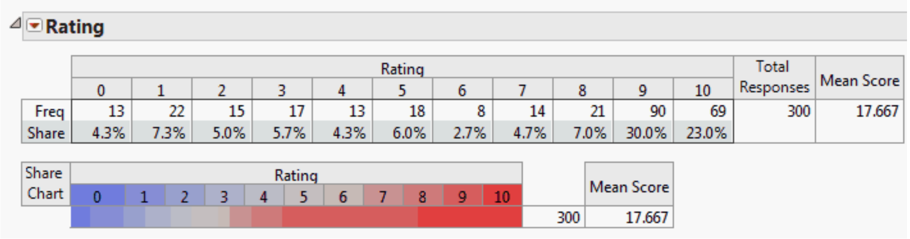
```
0 = -100
1 = -100
2 = -100
3 = -100
4 = -100
5 = -100
6 = -100
7 = 0
8 = 0
9 = 100
10 = 100
```

Value: 0

Score: -100

3. Select **Analyze** > **Consumer Research** > **Categorical**.
4. Select Rating and click **Responses** on the Simple tab.
5. Click **OK**.
6. Select **Mean Score** from the Categorical red triangle menu.

Figure 3.30 Rating Example Report



Rating

	Rating											Total Responses	Mean Score
	0	1	2	3	4	5	6	7	8	9	10		
Freq	13	22	15	17	13	18	8	14	21	90	69	300	17.667
Share	4.3%	7.3%	5.0%	5.7%	4.3%	6.0%	2.7%	4.7%	7.0%	30.0%	23.0%		

Share Chart	Rating		Mean Score
	0 1 2 3 4 5 6 7 8 9 10	300	17.667

Based on the defined value scores, a mean score of 17.667 was determined. Your results might be different as the Rating column values are random.

Chapter 4

Factor Analysis

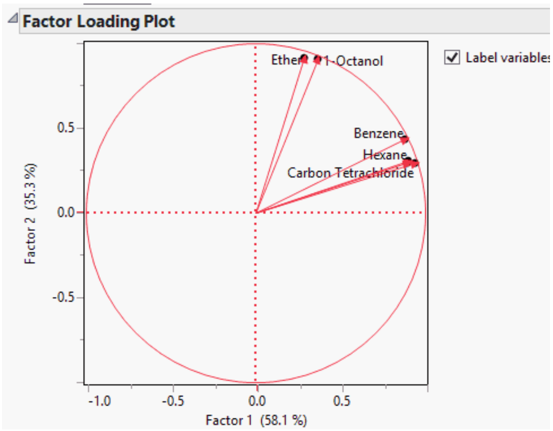
Identify Factors within Variables

Factor analysis (also known as common factor analysis and exploratory factor analysis) seeks to describe a collection of observed variables in terms of a smaller collection of (unobservable) latent variables, or factors. These factors, which are defined as linear combinations of the observed variables, are constructed to explain variation that is *common* to the observed variables. A primary goal of factor analysis is to achieve a meaningful interpretation of the observed variables through the factors. Another goal is to reduce the number of variables.

Factor analysis is used in many areas, and is of particular value in psychology, sociology, and education. In these areas, factor analysis is used to understand how manifest behavior can be interpreted in terms of underlying patterns and structures. For example, measures of participation in outdoor activities, hobbies, exercise, and travel, may all relate to a factor that can be described as "active versus inactive personality type". Factor analysis attempts to explain correlations among the observed variables in terms of the factor. In particular, it allows you to determine how much of the variance in each observable variable is accounted for by the factors you have identified. It also tells you how much of the variance in all the variables is accounted for by each factor.

Use factor analysis when you need to explore or interpret underlying patterns and structure in your data. Also consider using it to summarize the information in your variables using a smaller number of latent variables.

Figure 4.1 Rotated Factor Loading



Contents

Factor Analysis Platform Overview

Factor analysis models a set of observable variables in terms of a smaller number of unobservable factors. These factors account for the correlation or covariance between the observed variables. Once the factors are extracted, you perform factor rotation in order to obtain a meaningful interpretation of the factors.

Consider a situation where you have ten observed variables, X_1, X_2, …, X_{10}. Suppose that you want to model these ten variables in terms of two latent factors, F_1 and F_2. For convenience, it is assumed that the factors are uncorrelated and that each has mean zero and variance one. The model that you want to derive is of the form:

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

It follows that $Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$. The portion of the variance of X_i that is attributable to the factors, the common variance or *communality*, is $\beta_{i1}^2 + \beta_{i2}^2$. The remaining variance, $Var(\varepsilon_i)$, is the specific variance, and is considered to be unique to X_i.

The Factor Analysis platform provides a Scree Plot for the eigenvalues of the correlation or covariance matrix. You can use this as a guide in determining the number of factors to extract. Alternatively, you can accept the platform's suggestion of setting the number of factors equal to the number of eigenvalues that exceed one.

The platform provides two factoring methods for estimating the parameters of this model: Principal Components and Maximum Likelihood.

JMP provides two options for estimating the proportion of variance contributed by common factors for each variable. These Prior Communality options impose assumptions on the diagonal of the correlation (or covariance) matrix. The Principal Components option treats the correlation matrix, which has ones on its diagonal (or the covariance matrix with variances on its diagonal), as the structure to be analyzed. The Common Factor Analysis option sets the diagonal entries to values that reflect the proportion of the variation that is shared with other variables.

To support interpretability of the extracted factors, you rotate the factor structure. The Factor Analysis platform provides a variety of rotation methods that encompass both orthogonal and oblique rotations.

In contrast with factor analysis which looks at common variance, principal component analysis accounts for the total variance of the observed variables. See the Principal Components chapter in the *Multivariate Methods* book.

Example of the Factor Analysis Platform

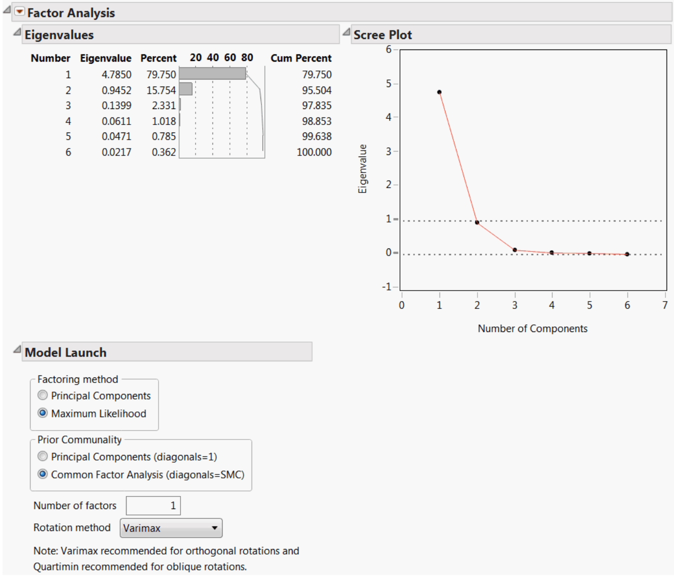
To view an example Factor Analysis report for a data table for two factors:

1. Open the data table Solubility.jmp sample data table.
2. Select **Analyze > Consumer Research > Factor Analysis**.

   The Factor Analysis launch window appears.
3. Select all of the continuous columns and click **Y, Columns**.
4. Keep the default **Estimation Method** and **Variance Scaling**.
5. Click **OK**.

   The initial Factor Analysis report appears.

Figure 4.2 Initial Factor Analysis Report



Factor Analysis

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.7850	79.750		79.750
2	0.9452	15.754		95.504
3	0.1399	2.331		97.835
4	0.0611	1.018		98.853
5	0.0471	0.785		99.638
6	0.0217	0.362		100.000

Scree Plot

Model Launch

Factoring method
○ Principal Components
● Maximum Likelihood

Prior Communality
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors 1

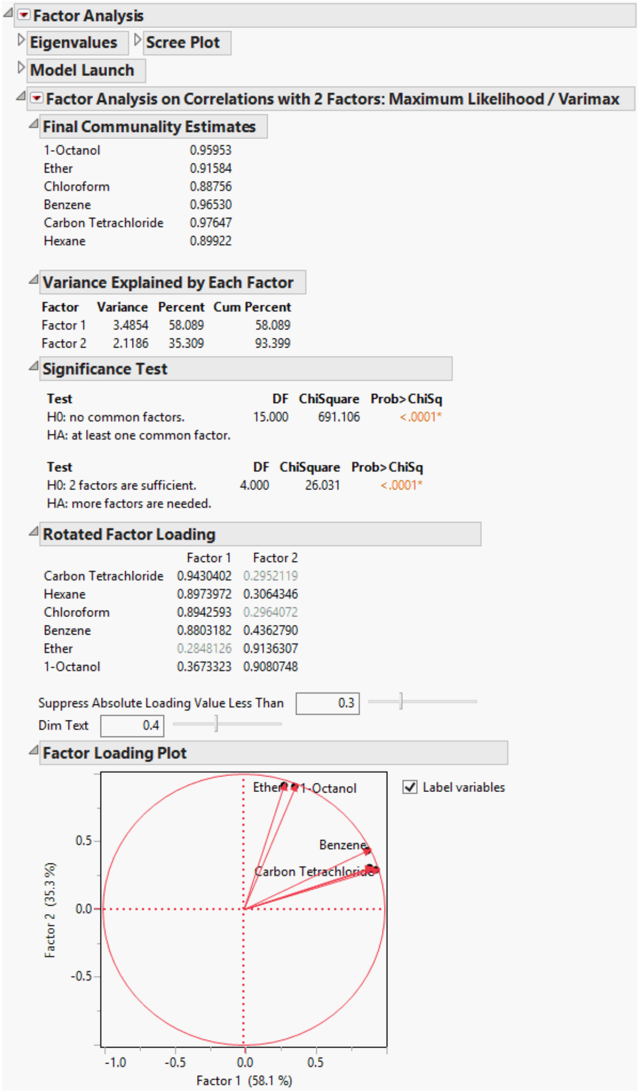
Rotation method Varimax

Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

6. For the Model Launch, select the following options:
   - Factoring Method as **Maximum Likelihood**
   - Prior Communality as **Common Factor Analysis**
   - Number of factors = 2
   - Rotation Method as **Varimax**
7. After all selections are made, click **Go**.

   The Factor Analysis report appears.

Figure 4.3 Example Factor Analysis Report



Factor Analysis on Correlations with 2 Factors: Maximum Likelihood / Varimax

Final Community Estimates

1-Octanol	0.95953
Ether	0.91584
Chloroform	0.88756
Benzene	0.96530
Carbon Tetrachloride	0.97647
Hexane	0.89922

Variance Explained by Each Factor

Factor	Variance	Percent	Cum Percent
Factor 1	3.4854	58.089	58.089
Factor 2	2.1186	35.309	93.399

Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	15.000	691.106	<.0001*

Test	DF	ChiSquare	Prob>ChiSq
H0: 2 factors are sufficient. HA: more factors are needed.	4.000	26.031	<.0001*

Rotated Factor Loading

	Factor 1	Factor 2
Carbon Tetrachloride	0.9430402	0.2952119
Hexane	0.8973972	0.3064346
Chloroform	0.8942593	0.2964072
Benzene	0.8803182	0.4362790
Ether	0.2848126	0.9136307
1-Octanol	0.3673323	0.9080748

Suppress Absolute Loading Value Less Than 0.3
Dim Text 0.4

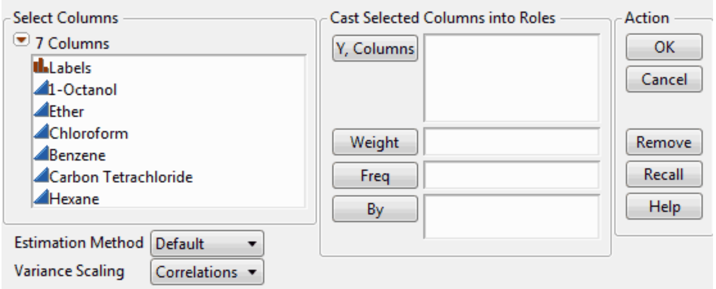
Factor Loading Plot

The report lists the communality estimates, variance, significance tests, rotated factor loadings, and a factor loading plot. Note that in the Factor Loading Plot, Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and Factor 2 relates to the Ether–1-Octanol cluster of variables. See "Factor Analysis Model Fit Options" on page 76 for details of the information shown in the report.

Launch the Factor Analysis Platform

Launch the Factor Analysis platform by selecting **Analyze > Consumer Research > Factor Analysis**. This example uses the Solubility.jmp sample data table.

Figure 4.4 Factor Analysis Launch Window



Estimation Method: Default
Variance Scaling: Correlations

Y, Columns Lists the continuous columns to be analyzed.

Weight Enables you to weight the analysis to account for pre-summarized data.

Freq Identifies a column whose numeric values assign a frequency to each row in the analysis.

By Creates a Factor Analysis report for each value specified by the By column so that you can perform separate analyses for each group.

Estimation Method Lists different methods for fitting the model. For details about the methods, see the Multivariate chapter in the *Multivariate Methods* book.

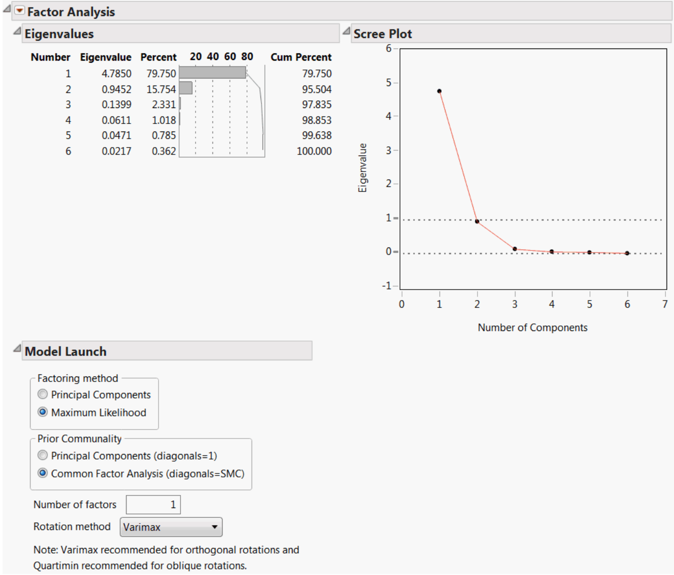
Variance Scaling Lists the scaling methods for performing the factor analysis based on **Correlations** (the same as Principal Components), **Covariances**, or **Unscaled**.

The Factor Analysis Report

The initial Factor Analysis report shows Eigenvalues and the Scree Plot. The Eigenvalues are obtained from a principal components analysis. The Scree Plot graphs these eigenvalues. The number of factors that JMP suggests in the Model Launch equals the number of eigenvalues that exceed 1.0.

Alternatively, you can use the scree plot to guide your initial choice for number of factors. The number of eigenvalues that appear before the scree plot levels out can provide an upper bound on the number of factors.

Figure 4.5 Factor Analysis Report



Factor Analysis

Eigenvalues

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	4.7850	79.750		79.750
2	0.9452	15.754		95.504
3	0.1399	2.331		97.835
4	0.0611	1.018		98.853
5	0.0471	0.785		99.638
6	0.0217	0.362		100.000

Scree Plot

Model Launch

Factoring method
○ Principal Components
● Maximum Likelihood

Prior Communality
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors 1

Rotation method Varimax

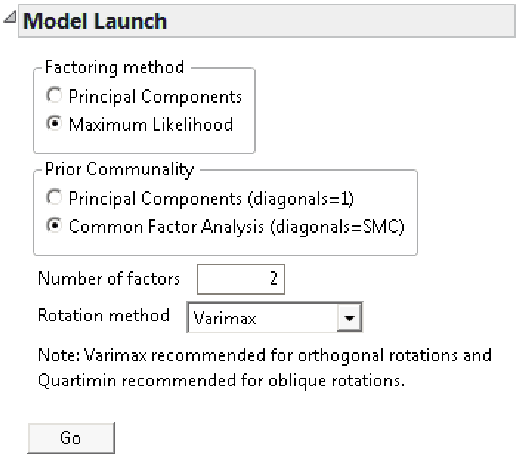
Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

In the example shown in Figure 4.5, the Scree Plot begins to level out after the second eigenvalue. The Eigenvalues table indicates that the first eigenvalue accounts for 79.75% of the variation and the second eigenvalue accounts for 15.75%, for a total of 95.50% of the total variation. The third eigenvalue only explains 2.33% of the variation, and the contributions from the remaining eigenvalues are negligible. Although the **Number of factors** box is initially set to 1, this analysis suggests that extracting 2 factors is appropriate.

Model Launch

To configure the Factor Analysis model, use the Model Launch section at the bottom of the Factor Analysis Report (Figure 4.6).

Figure 4.6 Model Launch



Model Launch

Factoring method
○ Principal Components
● Maximum Likelihood

Prior Communality
○ Principal Components (diagonals=1)
● Common Factor Analysis (diagonals=SMC)

Number of factors: 2

Rotation method: Varimax

Note: Varimax recommended for orthogonal rotations and Quartimin recommended for oblique rotations.

The Model Launch section enables you to configure the following options:

1. **Factoring method -** the method for extracting factors.
   - The **Principal Components** method is a computationally efficient method, but it does not allow for hypothesis testing.
   - The **Maximum Likelihood** method has desirable properties and allows you to test hypotheses about the number of common factors.

Note: The **Maximum Likelihood** method requires a positive definite correlation matrix. If your correlation matrix is not positive definite, select the **Principal Components** method.

2. **Prior Communality -** the method for estimating the proportion of variance contributed by common factors for each variable.
   - **Principal Components (diagonals = 1)** sets all communalities equal to 1, indicating that 100% of each variable's variance is shared with the other variables. Using this option with Factoring Method set to **Principal Components** results in principal component analysis.
   - **Common Factor Analysis (diagonals = SMC)** sets the communalities equal to squared multiple correlation (SMC) coefficients. For a given variable, the SMC is the RSquare for a regression of that variable on all other variables.
3. The **Number of factors** (or principal components) determined by eigenvalues greater than or equal to 1.0 or from the scree plot where the graph begins to level out.

Note: Alternatively, the *Kaiser criterion* retains those factors with eigenvalues greater than 1.0. In our example, only factor 1 would be retained for analysis.

4. The **Rotation method** to align the factor directions with the original variables for ease of interpretation. The default value is **Varimax**. See "Rotation Methods" on page 73 for a description of the available selections.
5. Click **Go** to generate the Factor Analysis report.

   Depending on the selected Variance Scaling, the appropriate factor analysis results appear. See "Factor Analysis Model Fit Options" on page 76 for details about the contents of the report. The Factor Analysis on Correlations and Factor Analysis on Unscales reports show the same information.

Rotation Methods

Rotations align the directions of the factors with the original variables so that the factors are more interpretable. You hope for clusters of variables that are highly correlated to define the rotated factors.

After the initial extraction, the factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are also uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more useful patterns than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable.

For each rotation method, we used the example described in "Example of the Factor Analysis Platform" on page 68 to view the Rotated Factor Loading and Factor Loading Plot.

Orthogonal Rotation Methods

Table 4.1 lists the available orthogonal (that is, uncorrelated) rotation methods.

Table 4.1 Orthogonal Rotation Methods

Method	SAS PROC FACTOR Equivalent
Varimax	ROTATE=ORTHOMAX with GAMMA = 1 **Note:** This is the default selection.
Biquartimax	ROTATE=ORTHOMAX with GAMMA = 0.5
Equamax	ROTATE=ORTHOMAX with GAMMA = number of factors/2
Factorparsimax	ROTATE=ORTHOMAX with GAMMA = number of variables

Table 4.1 Orthogonal Rotation Methods *(Continued)*

Method	SAS PROC FACTOR Equivalent
Orthomax	ROTATE=ORTHOMAX Or ROTATE=ORTHOMAX(*p*), where *p* as the orthomax weight or the GAMMA = value. **Note:** The default *p* value is 1 unless specified otherwise in the GAMMA = option. For additional information about orthomax weight, see the SAS documentation, "Simplicity Functions for Rotations."
Parsimax	ROTATE=ORTHOMAX with GAMMA = $\left[\frac{(nvar(nfact-1))}{(nvar+nfact-2)}\right]$ where *nvar* is the number of variables, and *nfact* is the number of factors.
Quartimax	ROTATE=ORTHOMAX with GAMMA=0

Oblique Rotation Methods

Table 4.2 lists the available oblique (that is, correlated) rotation methods.

Table 4.2 Oblique Rotation Methods

Method	SAS PROC FACTOR Equivalent

Biquartimin	ROTATE=OBLIMIN(.5)
	Or
	ROTATE=OBLIMIN with TAU=.5

Covarimin	ROTATE=OBLIMIN(1)
	Or
	ROTATE=OBLIMIN with TAU=1

Obbiquartimax	ROTATE=OBBIQUARTIMAX
Obequamax	ROTATE=OBEQUAMAX
Obfactorparsimax	ROTATE=OBFACTORPARSIMAX

Table 4.2 Oblique Rotation Methods *(Continued)*

Method	SAS PROC FACTOR Equivalent
Oblimin	ROTATE=OBLIMIN, where the default *p* value is zero, unless specified otherwise in the TAU= option. ROTATE=OBLIMIN(*p*) specifies *p* as the oblimin weight or the TAU= value. **Note:** For additional information about oblimin weight, see the SAS documentation, "Simplicity Functions for Rotations."
Obparsimax	ROTATE=OBPARSIMAX
Obquartimax	ROTATE=OBQUARTIMAX
Obvarimax	ROTATE=OBVARIMAX
Quartimin	ROTATE=OBLIMIN(0) or ROTATE=OBLIMIN with TAU=0
Promax	ROTATE=PROMAX

Factor Analysis Platform Options

The Factor Analysis platform red triangle menu enables you to select to view or hide the following report elements:

Eigenvalues A table that indicates the total number of factors extracted based on the eigenvalues (that is, the amount of variance contributed by each factor). The table includes the percent of the total variance contributed by that factor, a bar chart illustrating the percent contribution, and the cumulative percent contributed by each successive factor. The number of eigenvalues greater than or equal to 1.0 can be taken as the number of sufficient factors for analysis.

Scree Plot A plot of the eigenvalues to the number of components (or factors). The plot can be used to determine the number of factors that contribute to the maximum amount of variance. The point at where the graph levels out is can be taken as the number of sufficient factors for analysis.

Script Lists the Script menu options for the platform. See the JMP Platforms chapter in the *Using JMP* book for details.

See Figure 4.2 on page 68 for an example.

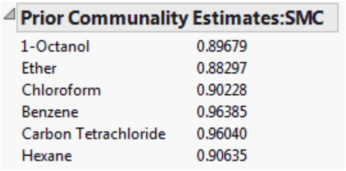
Factor Analysis Model Fit Options

After submitting the Model Launch, the model results appear. The following options are available from the Factor Analysis report’s red triangle menu.

Prior Communality An initial estimate of the communality for each variable. For a given variable, this estimate is the squared multiple correlation coefficient (SMC), or RSquare, for a regression of that variable on all other variables.

Note: The Prior Communality Estimates table only appears if the **Common Factor Analysis (diagonals = SMC)** option is selected.

Figure 4.7 Prior Communality Estimates



Prior Communality Estimates:SMC

1-Octanol	0.89679
Ether	0.88297
Chloroform	0.90228
Benzene	0.96385
Carbon Tetrachloride	0.96040
Hexane	0.90635

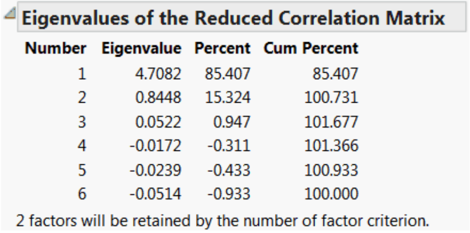
Eigenvalues Shows the eigenvalues of the reduced correlation matrix and the percent of the common variance for which they account. The reduced correlation matrix is the correlation matrix with its diagonal entries replaced by the communality estimates. The eigenvalues indicate the common variance explained by the factors. The Cum Percent can exceed 100% because the reduced correlation matrix is not necessarily positive definite and can have negative eigenvalues.

Note that the table indicates the number of factors retained for analysis.

The Eigenvalues option is only available when the Prior Communality option is set to Common Factor Analysis (diagonals = SMC). The communality estimates are the SMC (square multiple correlation) values.

Figure 4.8 indicates that the first two factors account for 100.731% of the common variance. This pattern suggests that you may not need more that two factors to model your data.

Figure 4.8 Eigenvalues of the Reduced Correlation Matrix



Eigenvalues of the Reduced Correlation Matrix

Number	Eigenvalue	Percent	Cum Percent
1	4.7082	85.407	85.407
2	0.8448	15.324	100.731
3	0.0522	0.947	101.677
4	-0.0172	-0.311	101.366
5	-0.0239	-0.433	100.933
6	-0.0514	-0.933	100.000

2 factors will be retained by the number of factor criterion.

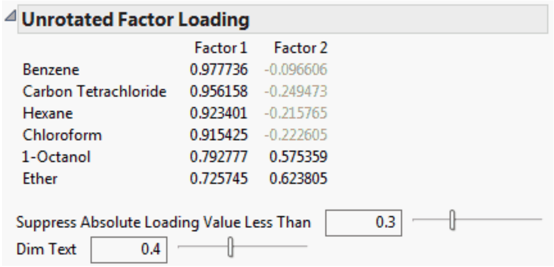
Unrotated Factor Loading Shows the factor loading matrix before rotation. Factor loadings measure the influence of a common factor on a variable. Because the unrotated factors are orthogonal, the factor loading matrix is the matrix of correlations between the variables and the factors. The closer the absolute value of a loading is to 1, the stronger the effect of the factor on the variable.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less than the entered suppressed value entered.

Note: The **Suppress Absolute Loading Values Less Than** value and **Dim Text** value are the same values used in the Rotated Factor Loading table. Changes to one loading table's settings changes the settings in the other loading table.

Figure 4.9 Unrotated Factor Loading



Unrotated Factor Loading

	Factor 1	Factor 2
Benzene	0.977736	-0.096606
Carbon Tetrachloride	0.956158	-0.249473
Hexane	0.923401	-0.215765
Chloroform	0.915425	-0.222605
1-Octanol	0.792777	0.575359
Ether	0.725745	0.623805

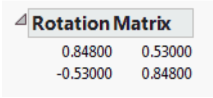
Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

Note: The Unrotated Factor Loading matrix is re-ordered so that variables associated with the same factor appear next to each other.

Rotation Matrix Shows the calculations used for rotating the factor loading plot and the factor loading matrix.

Figure 4.10 Rotation Matrix

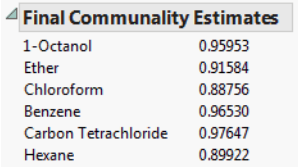


Rotation Matrix

0.84800	0.53000
-0.53000	0.84800

Final Communality Estimates Estimates of the communalities after the factor model has been fit. When the factors are orthogonal, the final communality estimate for a variable equals the sum of the squared loadings for that variable.

Figure 4.11 Final Community Estimates

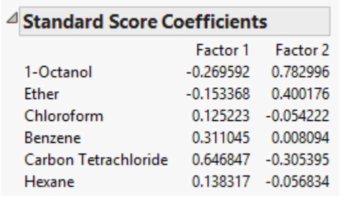


Final Communality Estimates

1-Octanol	0.95953
Ether	0.91584
Chloroform	0.88756
Benzene	0.96530
Carbon Tetrachloride	0.97647
Hexane	0.89922

Standard Score Coefficients Lists the multipliers used to convert factor values when saving rotated components as factors to the source data table.

Figure 4.12 Standard Score Coefficients

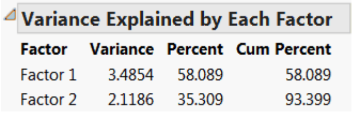


Standard Score Coefficients

	Factor 1	Factor 2
1-Octanol	-0.269592	0.782996
Ether	-0.153368	0.400176
Chloroform	0.125223	-0.054222
Benzene	0.311045	0.008094
Carbon Tetrachloride	0.646847	-0.305395
Hexane	0.138317	-0.056834

Variance Explained by Each Factor Gives the variance, percent, and cumulative percent, of common variance explained by each rotated factor.

Figure 4.13 Variance Explained by Each Factor



Variance Explained by Each Factor

Factor	Variance	Percent	Cum Percent
Factor 1	3.4854	58.089	58.089
Factor 2	2.1186	35.309	93.399

Significance Test If you select **Maximum Likelihood** as the factoring method, the results of two Chi-square tests are provided.

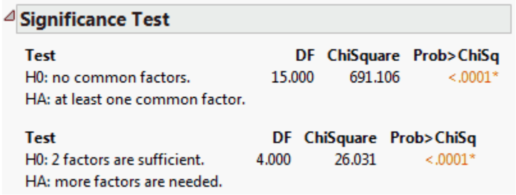
The first test is for H_0: No common factors. This null hypothesis indicates that none of the common factors are sufficient to explain the intercorrelations among the variables.

The second test is for H_0: N factors are sufficient, where N is the specified number of factors. Rejection of this null hypothesis indicates that more factors may be required to explain the intercorrelations among the variables.

The tests in Figure 4.14 indicate that the common factors already included in the model explain some of the intercorrelations, but that more factors are needed.

Note: The Significance Test table only appears if the **Maximum Likelihood** factoring method option is selected.

Figure 4.14 Significance Test



Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	15.000	691.106	<.0001*

Test	DF	ChiSquare	Prob>ChiSq
H0: 2 factors are sufficient. HA: more factors are needed.	4.000	26.031	<.0001*

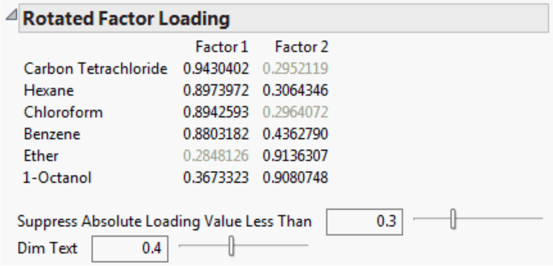
Rotated Factor Loading Shows the factor loading matrix after rotation. If the rotation is orthogonal, these values are the correlations between the variables and the rotated factors.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less than the entered suppressed value entered.

Note: The **Suppress Absolute Loading Values Less Than** value and **Dim Text** value are the same values used in the Unrotated Factor Loading table. Changes to one loading table's settings changes the settings in the other loading table.

Figure 4.15 Rotated Factor Loading



Rotated Factor Loading

	Factor 1	Factor 2
Carbon Tetrachloride	0.9430402	0.2952119
Hexane	0.8973972	0.3064346
Chloroform	0.8942593	0.2964072
Benzene	0.8803182	0.4362790
Ether	0.2848126	0.9136307
1-Octanol	0.3673323	0.9080748

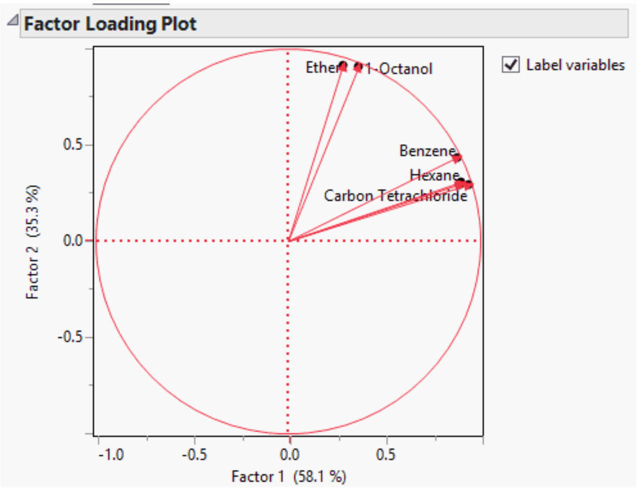
Suppress Absolute Loading Value Less Than 0.3

Dim Text 0.4

Note: The Rotated Factor Loading matrix is re-ordered so that variables associated with the same factor appear next to each other.

Factor Loading Plot The plot of the rotated loading factors.

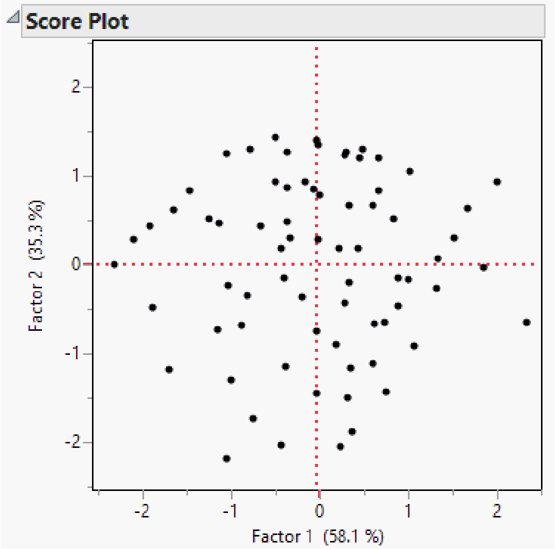
Figure 4.16 Factor Loading Plot



Note that in the Factor Loading Plot, Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and Factor 2 relates to the Ether–1-Octanol cluster of variables. See the matrix of “Rotated Factor Loading” on page 79 for details.

Score Plot The Score Plot graphs each factor’s calculated values in relation to the other adjusting each value for the mean and standard deviation.

Figure 4.17 Score Plot



Score Plot with Imputation Imputes any missing values and creates a score plot. This option is available only if there are missing values.

Save Rotated Components Saves the rotated components to the data table, with a formula for computing the components. The formula cannot evaluate rows with missing values.

Save Rotated Components with Imputation Imputes missing values, and saves the rotated components to the data table. The column contains a formula for doing the imputation, and computing the rotated components. This option appears after the Factor Analysis option is used, and if there are missing values.

Remove Fit Removes the fit model results from the Factor Analysis Fit Model report. This option enables you to change the Model Launch configuration for a new report.

Chapter 5

Choice Models

Fit Models for Choice Experiments

The Choice platform is designed for use in market research experiments, where the ultimate goal is to discover the preference structure of consumers. Then, this information is used to design products or services that have the attributes most desired by consumers.

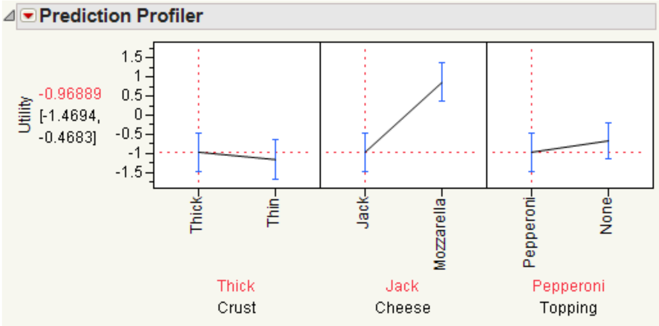
Features provided in the Choice platform include:

- Ability to use information about consumer traits as well as product attributes.
- Integration of data from one, two, or three sources.
- Ability to use the integrated profiler to understand, visualize, and optimize the response (utility) surface.
- Provides subject-level scores for segmenting or clustering your data.
- Uses a special default bias-corrected maximum likelihood estimator described by Firth (1993). This method has been shown to produce better estimates and tests than MLEs without bias correction. In addition, bias-corrected MLEs ameliorate separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

The Choice platform is not appropriate to use for fitting models that involve:

- Ranking or scoring.
- Nested hierarchical choices. (PROC MDC in SAS/ETS can be used for such analysis.)

Figure 5.1 Choice Platform Example - Prediction Profiler



Contents

Choice Modeling Platform Overview

Choice modeling, pioneered by McFadden (1974), is a powerful analytic method used to estimate the probability of individuals making a particular choice from presented alternatives. Choice modeling is also called conjoint modeling, discrete choice analysis, and conditional logistic regression.

The Choice Modeling platform uses a form of conditional logistic regression. Unlike simple logistic regression, choice modeling uses a linear model to model choices based on response attributes and not solely upon subject characteristics. For example, in logistic regression, the response might be whether you buy brand A or brand B as a function of ten factors or characteristics that describe you such as your age, gender, income, education, and so on. However, in choice modeling, you might be choosing between two cars that are a compound of ten attributes such as price, passenger load, number of cup holders, color, GPS device, gas mileage, anti-theft system, removable-seats, number of safety features, and insurance cost.

When engineers design a product, they routinely make hundreds or thousands of small design decisions. Most of these decisions are not tested by prospective customers. Consequently, these products are not optimally designed. However, if customer testing is not too costly and test subjects (prospective customers) are readily available, it is worthwhile to test more of these decisions via consumer choice experiments.

Modeling costs have recently decreased with improved product and process development techniques and methodologies. Prototyping, including pure digital prototyping, is becoming less expensive, so it is possible to evaluate the attributes and consequences of more alternatives. Another important advancement is the use of the Internet to deliver choice experiments to a wide audience. You can now inform your customers that they can have input into the design of the next product edition by completing a web survey.

Choice modeling can be added to Six Sigma programs to improve consumer products. Six Sigma aims at making products better by improving the manufacturing process and ensuring greater performance and durability. But, Six Sigma programs have not addressed one very important aspect of product improvement—making the products that people actually want. Six Sigma programs often consider the Voice of the Customer and can use customer satisfaction surveys. However, while these surveys can disclose what is wrong with the product, they fail to identify consumer preferences with regard to specific product attributes. Choice experiments provide a tool that enables companies to gain insight for actual customer preferences. Choice modeling analysis can reveal such preferences.

Market research experiments have a long history of success, but performing these experiments has been expensive, and research has previously focused on price elasticity and competitive situations. It is by using these same techniques for product design engineering where choice modeling can have the most impact.

Example of the Choice Platform

Suppose that you are supplying pizza for an airline. You want to find pizza attributes that are optimal for the flying population. So, you have a group of frequent flyers complete a choice survey. In order to weigh the importance of each attribute and to determine whether there are any interactions between the different attributes, you give them a series of choices that require them to state their preference between each pair of choices. One pair of choices might be between two types of pizza that they like, or between two types of pizza that they do not like. Hence, the choice might not always be easy.

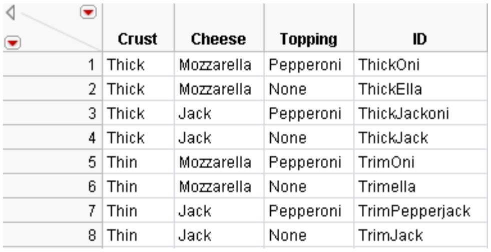
This example examines pizza choices where three attributes, each with two levels, are presented to the subjects:

- crust (thick or thin)
- cheese (mozzarella or Monterey Jack)
- topping (pepperoni or none)

Suppose a subject likes thin crust with mozzarella cheese and no topping, but the choices given to the subject are either a thick crust with mozzarella cheese and pepperoni topping, or a thin crust with Monterey Jack cheese and no topping. Because neither of these pizzas is ideal, the subject has to weigh which of the attributes are more important.

The profile data table lists all the pizza choice combinations that you want to present to the subjects. Each choice combination is given an ID. The profile data table is shown in Figure 5.2.

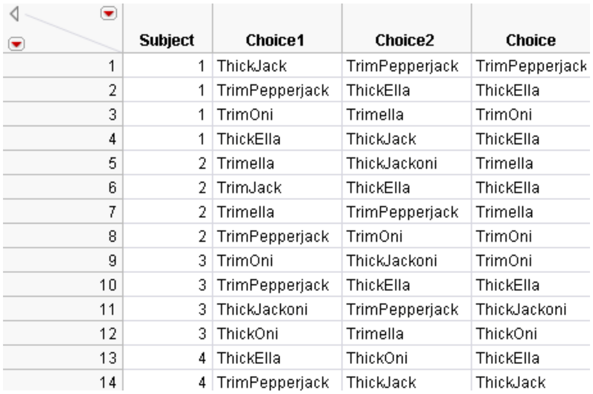
Figure 5.2 Pizza Profile Data Table



	Crust	Cheese	Topping	ID
1	Thick	Mozzarella	Pepperoni	ThickOni
2	Thick	Mozzarella	None	ThickElla
3	Thick	Jack	Pepperoni	ThickJackoni
4	Thick	Jack	None	ThickJack
5	Thin	Mozzarella	Pepperoni	TrimOni
6	Thin	Mozzarella	None	Trimella
7	Thin	Jack	Pepperoni	TrimPepperjack
8	Thin	Jack	None	TrimJack

For the actual survey or experiment, each subject is given four trials, where each trial consists of stating his or her preference between two choice profiles (Choice1 and Choice2). The choice profiles given for each trial are referred to as a choice set. One subject's choice trials can be different from another subject's trials. Refer to the DOE Choice Design platform for generating optimal choice designs. Twelve runs from the first three subjects are shown in Figure 5.3.

Figure 5.3 Pizza Response Data Table Segment

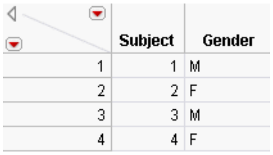


	Subject	Choice1	Choice2	Choice
1	1	ThickJack	TrimPepperjack	TrimPepperjack
2	1	TrimPepperjack	ThickElla	ThickElla
3	1	TrimOni	Trimella	TrimOni
4	1	ThickElla	ThickJack	ThickElla
5	2	Trimella	ThickJackoni	Trimella
6	2	TrimJack	ThickElla	ThickElla
7	2	Trimella	TrimPepperjack	Trimella
8	2	TrimPepperjack	TrimOni	TrimOni
9	3	TrimOni	ThickJackoni	TrimOni
10	3	TrimPepperjack	ThickElla	ThickElla
11	3	ThickJackoni	TrimPepperjack	ThickJackoni
12	3	ThickOni	Trimella	ThickOni
13	4	ThickElla	ThickOni	ThickElla
14	4	TrimPepperjack	ThickJack	ThickJack

Notice that each choice value refers to an ID value in the Profile data table that has the attribute information.

If data about the subject are to be used, a separate Subject data table is needed. This table includes a subject ID column and characteristics of the subject. In the pizza example, the only characteristic or attribute about the Subject is Gender. Subject data for the first four subjects are shown in Figure 5.4. Notice that the response choices and choice sets in the response data table use the ID names given in the profile data set. Similarly, the subject identifications in the response data table match those in the subject data table.

Figure 5.4 Pizza Subject Data Table Segment



	Subject	Gender
1	1	M
2	2	F
3	3	M
4	4	F

Launch the Choice Platform

The Choice platform is unique because it is designed to use data from one, two or three different data tables.

Profile Data Describe the attributes associated with each choice. Each choice can comprise many different attributes, and each attribute is listed as a column in the data table. There is a row for each possible choice, and each possible choice contains a unique ID.

Response Data Contain the experimental results and have the choice set **ID**s for each trial as well as the actual choice selected by the subject. Each subject usually has several trials, or *choice sets*, to cover several choice possibilities. There can be more than one row of data for each subject. For example, an experiment might have 100 subjects with each subject making 12 choice decisions, resulting in 1200 rows in this data table. The Response data are linked to the Profile data through the choice set columns and the actual choice response column. Choice set refers to the set of alternatives from which the subject makes a choice. Grouping variables are sometimes used to align choice indices when more than one group is contained within the data.

Subject Data Are optional, depending on whether subject effects are to be modeled. This source contains one or more attributes or characteristics of each subject and a subject identifier. The Subject data table contains the same number of rows as subjects and has an identifier column that matches a similar column in the Response data table. You can also put Subject data in the Response data table, but it is still specified as a subject table.

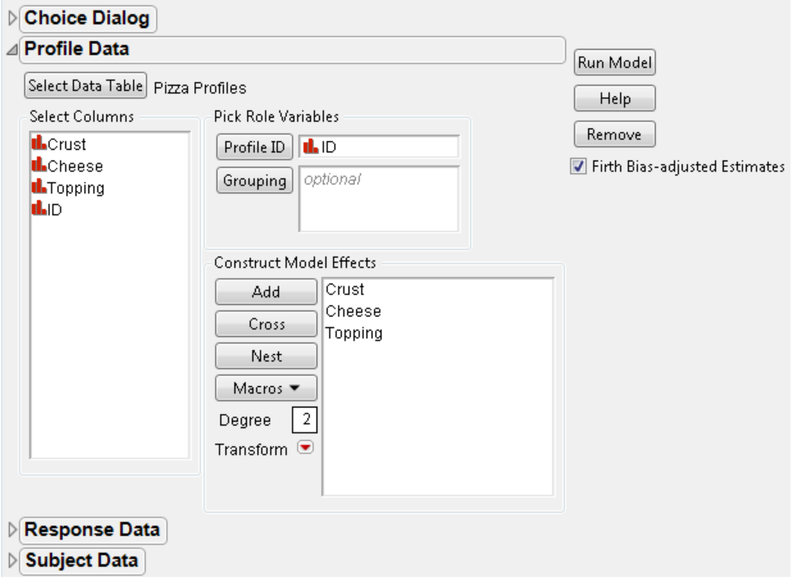
If all your data are contained in one table, you can use the Choice platform, but additional effort is necessary. See the section "One-Table Analysis" on page 103.

Because the Choice platform can use several data tables, no initial assumption is made about using the current data table—as is the case with other JMP platforms. You must select the data table for each of the three choice data sources. You are prompted to select the profile data set and the response data set. If you want to model subject attributes, then a subject data set must also be selected. You can expand or collapse each section of the Choice dialog box, as needed.

To illustrate the Choice platform, three data sets from the pizza example are used and are found in the sample data directory. The first data set is entered into the Profile Data section of the Choice Dialog Box, as shown in Figure 5.5.

1. Select **Analyze** > **Consumer Research** > **Choice** to open the launch dialog box. You see three separate sections for each of the data sources.
2. Select **Select Data Table** under Profile Data. A new dialog box appears, which prompts you to specify the data table for the profile data. You can select from any of the data sets already open in the current JMP session, or you can select **Other**. Selecting **Other** enables you to open a file that is not currently open.
3. Select Pizza Profiles.jmp. The columns from this table now populate the field under **Select Columns** in the Choice Dialog box.
4. Select ID for **Profile ID** under **Pick Role Variables** and **Add** Crust, Cheese, and Topping under **Construct Model Effects.** If the **Profile ID** column does not uniquely identify each row in the profile data table, you need to add **Grouping** columns until the combination of **Grouping** and **Profile ID** columns uniquely identify the row, or profile. For example, if Profile ID = 1 for Survey = A, and a different Profile ID = 1 for Survey = B, then Survey would be used as a **Grouping** column. In this simple experiment, all eight combinations of the three two-level factors were used.

Figure 5.5 Profile Data Set Dialog Box



Select Data Table: Pizza Profiles

Select Columns: Crust, Cheese, Topping, ID

Pick Role Variables: Profile ID: ID; Grouping: optional

Construct Model Effects: Crust, Cheese, Topping

Degree: 2

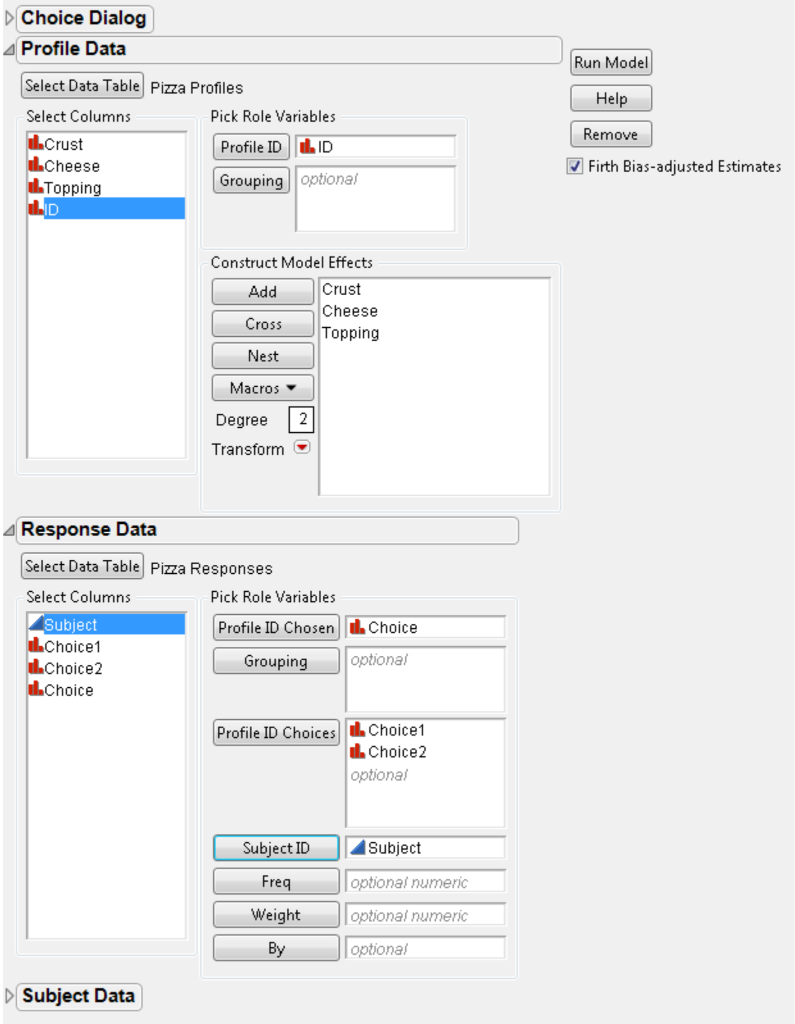
☑ Firth Bias-adjusted Estimates

The second data set is the Response Data containing the experimental results. For the pizza example, Choice1 and Choice2 are the profile ID choices given to a subject on each of four trials. The Choice column contains the chosen preference between Choice1 and Choice2.

5. Open the Response Data section of the dialog box. Click **Select Data Table**. When the Response Data Table dialog box appears, select Pizza Responses.jmp.
6. Select Choice for the **Profile ID Chosen**, and Choice1 and Choice2 for the **Profile ID Choices**.
7. Select Subject for **Subject ID** to identify individual subjects for later analysis. If you are not interested in assessing subjects at this time, it is not necessary to enter the subject data set into the model.
8. **Freq** and **Weight** are used to weight the analysis. For example, if you have summarized a set that had the same trial **Freq** is used for the count in that row of the response data table.

   The completed dialog box, without the subject data set, is shown in Figure 5.6.

Figure 5.6 Response Data Set Dialog Box



Profile Data

Select Data Table: Pizza Profiles

Select Columns: Crust, Cheese, Topping, ID

Pick Role Variables

Profile ID: ID

Grouping: optional

Construct Model Effects: Crust, Cheese, Topping

Degree: 2

☑ Firth Bias-adjusted Estimates

Response Data

Select Data Table: Pizza Responses

Select Columns: Subject, Choice1, Choice2, Choice

Pick Role Variables

Profile ID Chosen: Choice

Grouping: optional

Profile ID Choices: Choice1, Choice2, optional

Subject ID: Subject

Freq: optional numeric

Weight: optional numeric

By: optional

If you are scripting the Choice platform, you can also set the acceptable criterion for convergence when estimating the parameters by adding this command to the `Choice()` specification:

```
Choice( ..., Convergence Criterion( fraction ), ... )
```

See the JMP Scripting Index in the **Help** menu for an example.

Choice Model Output

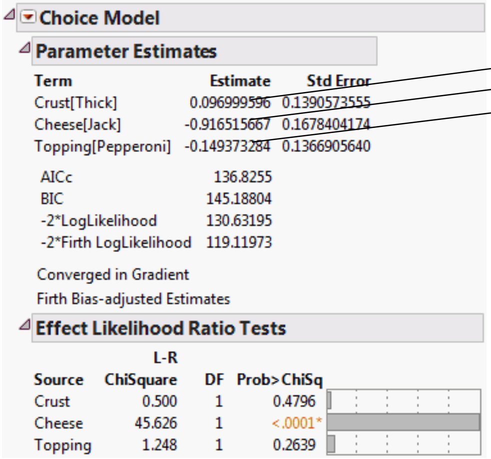
Click **Run Model** to obtain the results. These results are shown in Figure 5.7.

- The resulting parameter estimates are sometimes referred to as *part-worths*. Each part-worth is the coefficient of utility associated with that attribute. By default, these estimates are based on the Firth bias-corrected maximum likelihood estimators, and are, therefore, considered to be more accurate than MLEs without bias correction.
- Comparison criteria are used to help determine the better-fitting model(s) when more than one model is investigated for your data. The model with the lower or lowest criterion value is believed to be the better or best model. Criteria are shown in the Choice Model output and include AICc (corrected Akaike's Information Criterion), BIC (Bayesian Information Criterion), -2*LogLikelihood, and -2*Firth Loglikelihood. The AICc formula is:

  $$\text{AICc} = -2\text{loglikelihood} + 2k + \frac{2k(k+1)}{n-k-1}$$

  where k is the number of estimated parameters in the model and n is the number of observations in the dataset. The BIC formula is: - 2 LogLikelihood + k * ln(n), where k parameters is fitted to data with n observations and LogLikelihood is the maximized log-likelihood. Note that the -2*Firth Loglikelihood result is included only in the report when the Firth Bias-adjusted Estimates check box is checked in the launch window. (See Figure 5.6.) This option is checked by default. The decision to use or not use the Firth Bias-adjusted Estimates does not affect the AICc score or the -2*LogLikelihood results.
- Likelihood ratio tests appear for each effect in the model. These results are obtained by default if the model is fit quickly (less than five seconds). Otherwise, you can select the Choice Model drop down menu and select **Likelihood Ratio Tests**.

Figure 5.7 Choice Model Results with No Subject Data for Pizza Example



Choice Model

Parameter Estimates

Term	Estimate	Std Error
Crust[Thick]	0.096999596	0.1390573555
Cheese[Jack]	-0.916515667	0.1678404174
Topping[Pepperoni]	-0.149373284	0.1366905640

AICc	136.8255
BIC	145.18804
-2*LogLikelihood	130.63195
-2*Firth LogLikelihood	119.11973

Converged in Gradient

Firth Bias-adjusted Estimates

Effect Likelihood Ratio Tests

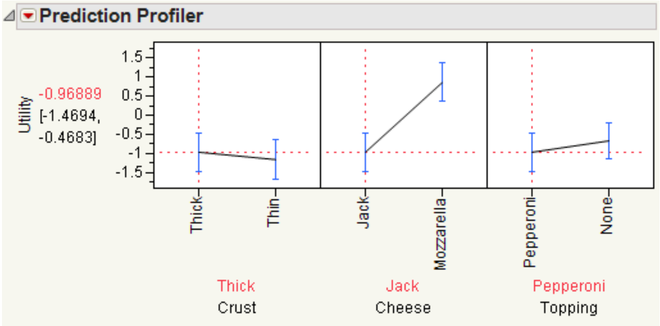
Source	L-R ChiSquare	DF	Prob>ChiSq
Crust	0.500	1	0.4796
Cheese	45.626	1	<.0001*
Topping	1.248	1	0.2639

Part-worths for crust, cheese, and topping

The profiler option is particularly valuable in understanding the model. It shows the value of the linear model, or the utility, as you change each factor, one at a time.

- You specify the profiler option by selecting **Profiler** in the platform menu. The Prediction Profiler is shown in Figure 5.8.
- In this example involving only main effects, the factor showing the greatest difference is Cheese, favoring Mozzarella.
- If there were interactions in the model, more exploration of the profiler would be needed in order to understand the response surface.

Figure 5.8 Prediction Profiler with No Subject Data for Pizza Example

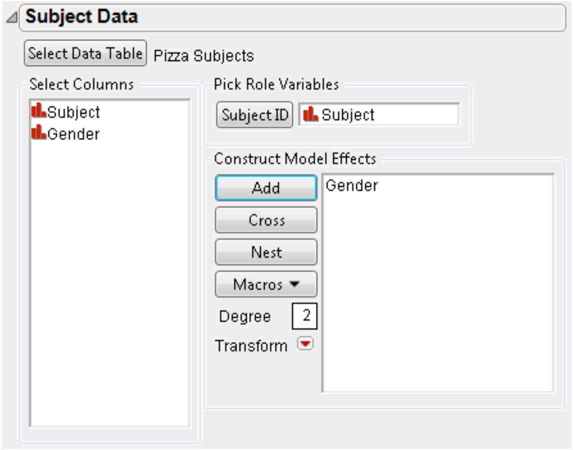


Subject Effects

If you want to include subject effects in the model, you need to open the Subject data table section of the Choice dialog box. Suppose you are interested in Gender main effects.

1. Open the Subject Data section of the launch dialog box. Click **Select Data Table**, and when the Subject Data Table opens, select Pizza Subjects.jmp.
2. Specify Subject as **Subject ID**, and add Gender under **Construct Model Effects**. The Subject data dialog box section for the pizza example is shown in Figure 5.9.
3. Click **Run Model**.

Figure 5.9 Choice Model Subject Data Dialog Box for Pizza Example



Subject Data

Select Data Table: Pizza Subjects

Select Columns: Subject, Gender

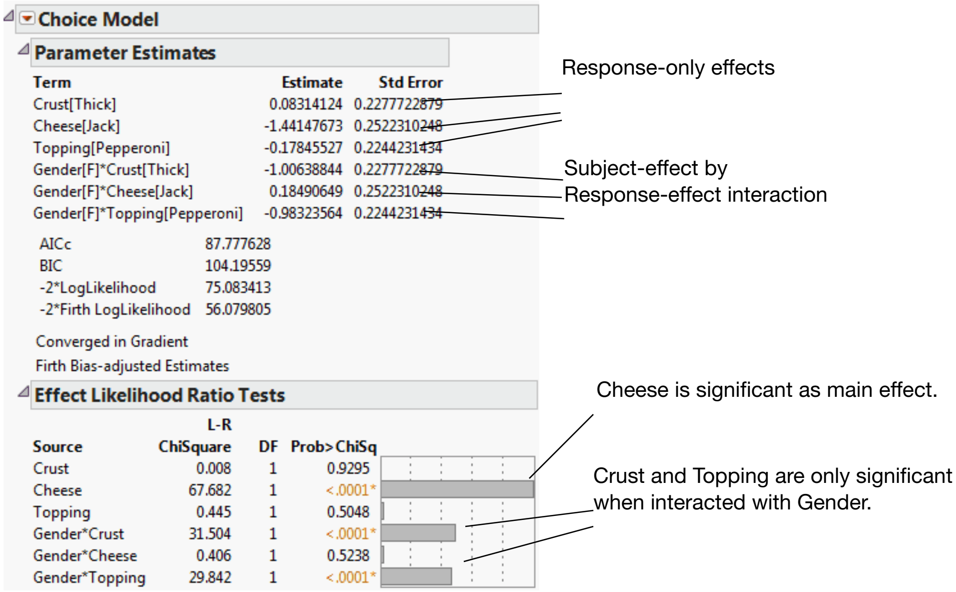
Pick Role Variables — Subject ID: Subject

Construct Model Effects: Gender

Degree: 2

Figure 5.10 shows the parameter estimates and the likelihood ratio tests for the Choice Model with subject effects included. Strong interactions are seen between Gender and Crust and between Gender and Topping. When the Crust and Topping factors are assessed for the entire population, the effects are not significant. However, the effects of Crust and Topping are strong when they are evaluated between Gender groups.

Figure 5.10 Choice Model Results with Subject Effects for Pizza Example



Choice Model

Parameter Estimates

Term	Estimate	Std Error
Crust[Thick]	0.08314124	0.2277722879
Cheese[Jack]	-1.44147673	0.2522310248
Topping[Pepperoni]	-0.17845527	0.2244231434
Gender[F]*Crust[Thick]	-1.00638844	0.2277722879
Gender[F]*Cheese[Jack]	0.18490649	0.2522310248
Gender[F]*Topping[Pepperoni]	-0.98323564	0.2244231434

AICc	87.777628
BIC	104.19559
-2*LogLikelihood	75.083413
-2*Firth LogLikelihood	56.079805

Converged in Gradient

Firth Bias-adjusted Estimates

Effect Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Crust	0.008	1	0.9295
Cheese	67.682	1	<.0001*
Topping	0.445	1	0.5048
Gender*Crust	31.504	1	<.0001*
Gender*Cheese	0.406	1	0.5238
Gender*Topping	29.842	1	<.0001*

Response-only effects

Subject-effect by Response-effect interaction

Cheese is significant as main effect.

Crust and Topping are only significant when interacted with Gender.

The profiler is used to explore the response surface of the model. Select the Choice Model drop-down menu and select the **Profiler**. You can Alt-click any segment of the Profiler graphics to lock that particular factor level. (F is locked by default in this example.) A solid vertical line appears in place of the dotted line. Other factor settings can then be assessed easily for the locked factor level.

- As shown in Figure 5.11, when the female (F) level of **Gender** is locked, the Profiler shows that females prefer pizza with thin crust, mozzarella cheese, and no topping. For example, the Utility measure is higher for Crust equals Thin, meaning that females prefer thin crust.
- Now, switch **Gender** to M to assess male pizza preferences. As shown in Figure 5.12, males prefer thick crust, mozzarella cheese, and pepperoni topping.

Figure 5.11 Prediction Profiler with Subject Data and Female Level Factor Setting

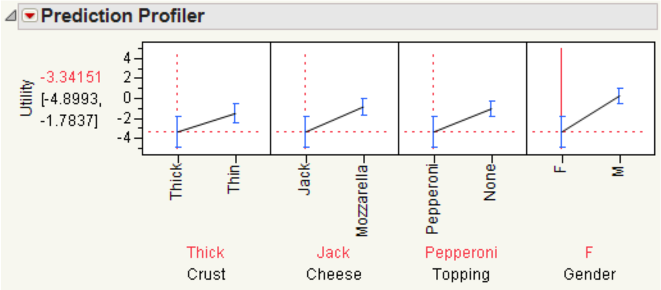
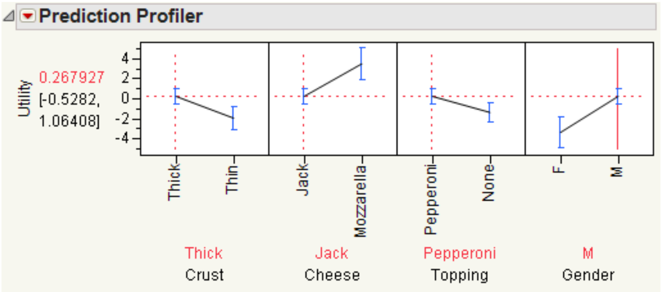


Figure 5.12 Prediction Profiler with Subject Data and Male Level Factor Setting



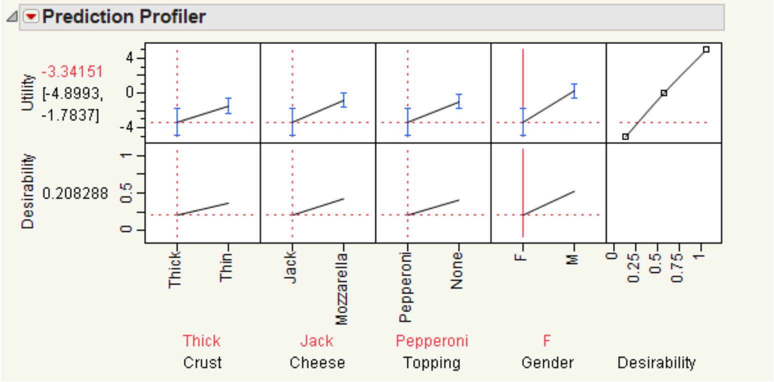
Utility Grid Optimization

The Prediction Profiler enables you to optimize the utility function over a grid of fixed subject factor settings without having to manually manipulate profiler settings:

1. Click the platform drop-down-menu, select **Profiler**, and verify that one of the subject factors is locked. If not, Alt-click within the Profiler plot to lock the desired factor level. Set the Utility function to the maximum values for the other factors by sliding the red dotted vertical line.

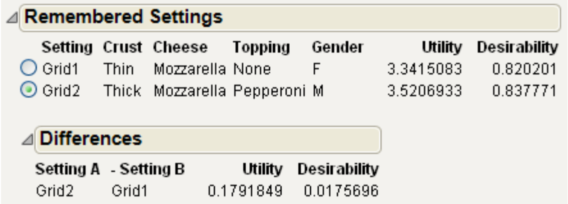
2. Click the red triangle menu of the Prediction Profiler and select **Desirability Functions**. A new row is added to the Prediction Profiler, displaying overall desirability traces and measures. Utility and Desirability Functions are shown together in Figure 5.13.

Figure 5.13 Utility and Desirability Functions



3. From the red triangle menu of the Prediction Profiler, select **Maximize for each Grid Point**. A Remembered Settings table containing the grid settings with the maximum utility and desirability functions, and a table of differences between grids is displayed. See Figure 5.14. As illustrated, this feature can be a very quick and useful tool for selecting the most desirable attribute combinations for a factor.

Figure 5.14 Utility and Desirability Settings


Remembered Settings

Setting	Crust	Cheese	Topping	Gender	Utility	Desirability
○ Grid1	Thin	Mozzarella	None	F	3.3415083	0.820201
● Grid2	Thick	Mozzarella	Pepperoni	M	3.5206933	0.837771

Differences

Setting A	- Setting B	Utility	Desirability
Grid2	Grid1	0.1791849	0.0175696

The grid setting for females shows that the greatest utility and desirability values are obtained when the pizza attributes are thin crust, mozzarella cheese, and no topping. For males, the grid setting shows that the highest utility and desirability values are obtained when the pizza attributes are thick crust, mozzarella cheese, and pepperoni topping.

Choice Platform Options

The Choice Modeling platform has many available options. To access these options, select the platform drop-down menu.

Likelihood Ratio Tests Tests the significance of each effect in the model. These are done by default if the estimate of CPU time is less than five seconds.

Joint Factor Tests Tests each factor in the model by constructing a likelihood ratio test for all the effects involving that factor.

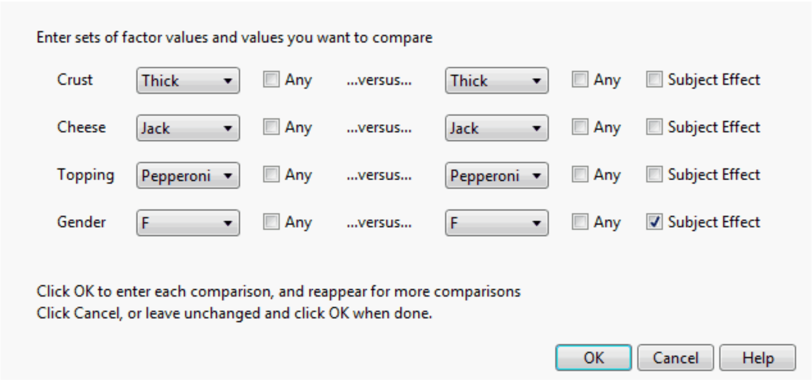
Confidence Intervals Produces a 95% confidence interval for each parameter (by default), using the profile-likelihood method. Shift-click the platform drop-down menu and select **Confidence Intervals** to input alpha values other than 0.05.

Correlation of Estimates Shows the correlations of the parameter estimates.

Effect Marginals Shows the fitted utility values for different levels in the effects, with neutral values used for unrelated factors.

Comparisons Performs comparisons between specific alternative choice profiles. Enables you to select factor values and the values that you want to compare. From here you can compare specific configurations, including comparing all settings on the left or right by selecting the **Any** check boxes. Using **Any** does not compare all combinations across features, but rather all combinations of comparisons, one feature at a time, using the left settings as the settings for the other factors.

Figure 5.15 Comparisons Example



Enter sets of factor values and values you want to compare

Crust	Thick	□ Any	...versus...	Thick	□ Any	□ Subject Effect
Cheese	Jack	□ Any	...versus...	Jack	□ Any	□ Subject Effect
Topping	Pepperoni	□ Any	...versus...	Pepperoni	□ Any	□ Subject Effect
Gender	F	□ Any	...versus...	F	□ Any	☑ Subject Effect

Click OK to enter each comparison, and reappear for more comparisons
Click Cancel, or leave unchanged and click OK when done.

Willingness to Pay Calculates how much a price must change allowing for the new feature settings to produce the same predicted outcome. The result is calculated using the **Baseline** settings (for each background setting) and then determining the outcome after altering the **Role**, including.

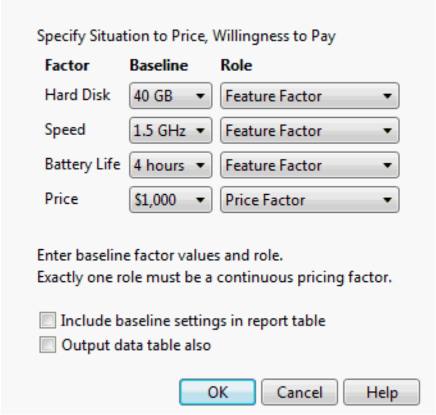
– Feature Factor - a feature in the experiment that you want to price.

– Price Factor - a continuous price factor in the experiment.

– Background Constant - something that you want to hold constant at a baseline value.

– Background Variable - something that you want to iterate across values.

Figure 5.16 Willingness to Pay Example



Specify Situation to Price, Willingness to Pay

Factor	Baseline	Role
Hard Disk	40 GB	Feature Factor
Speed	1.5 GHz	Feature Factor
Battery Life	4 hours	Feature Factor
Price	$1,000	Price Factor

Enter baseline factor values and role.
Exactly one role must be a continuous pricing factor.

□ Include baseline settings in report table
□ Output data table also

The **Include baseline settings in report table** option adds the baseline settings with a price change of zero, which is useful if you make an output table of these prices displaying all the baseline settings as well as the featured settings.

Profiler Produces a response surface viewer that takes vertical cross-sections across each factor, one at a time.

Save Utility Formula Makes a new column with a formula for the utility, or linear model, that is estimated. This is in the profile data table, except if there are subject effects. In that case, it makes a new data table for the formula. This formula can be used with various profilers with subsequent analyses.

Save Gradients by Subject Constructs a new table that has a row for each subject containing the average (Hessian-scaled-gradient) steps on each parameter. This corresponds to using a Lagrangian multiplier test for separating that subject from the remaining subjects. These values can later be clustered, using the built-in-script, to indicate unique market segments represented in the data.

Model Dialog Shows the Choice dialog box, which can be used to modify and re-fit the model. You can specify new data sets, new IDs, and new model effects.

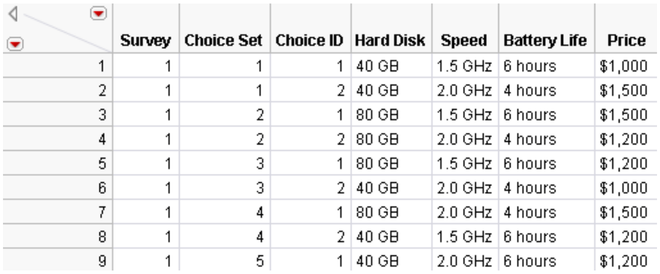
Example: Valuing Trade-offs

The Choice Modeling platform is also useful for determining the relative importance of product attributes. Even if the attributes of a particular product that are important to the consumer are known, information about preference trade-offs with regard to these attributes might be unknown. By gaining such information, a market researcher or product designer is able to incorporate product features that represent the optimal trade-off from the perspective of the consumer.

The advantages of this approach to product design can be found in the following example. It is already known that four attributes are important for laptop design--hard-disk size, processor speed, battery life, and selling price. The data gathered for this study are used to determine which of four laptop attributes (Hard Disk, Speed, Battery Life, and Price) are most important. It also assesses whether there are Gender or Job differences seen with these attributes.

1. Select **Analyze** > **Consumer Research** > **Choice** to open the launch dialog box.
2. Open Laptop Profile.jmp from the sample data directory and **Select Data Table** under Profile Data. Select Laptop Profile.jmp. A partial listing of the Profile Data table is shown in Figure 5.17. The complete data set consists of 24 rows, 12 for Survey 1 and 12 for Survey 2. Survey and Choice Set define the grouping columns and Choice ID represents the four attributes of laptops: Hard Disk, Speed, Battery Life, and Price.

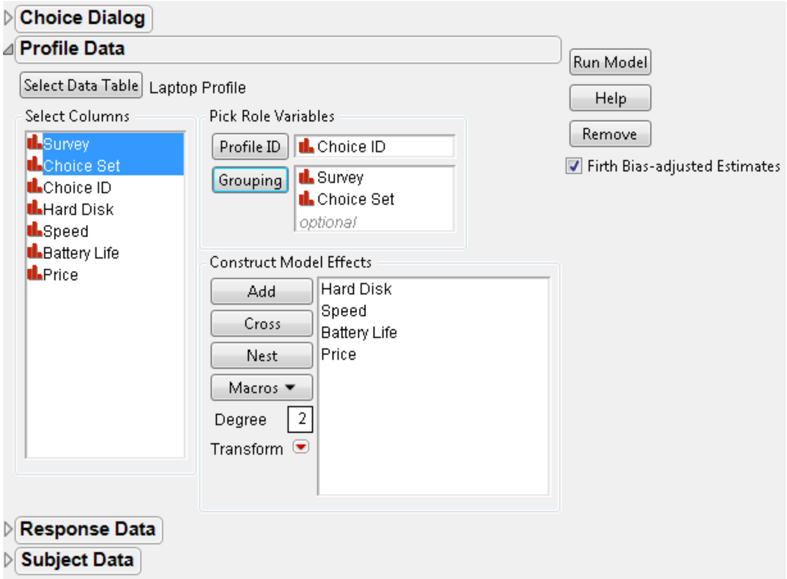
Figure 5.17 Profile Data Set for the Laptop Example



	Survey	Choice Set	Choice ID	Hard Disk	Speed	Battery Life	Price
1	1	1	1	40 GB	1.5 GHz	6 hours	$1,000
2	1	1	2	40 GB	2.0 GHz	4 hours	$1,500
3	1	2	1	80 GB	1.5 GHz	6 hours	$1,500
4	1	2	2	80 GB	2.0 GHz	4 hours	$1,200
5	1	3	1	80 GB	1.5 GHz	6 hours	$1,200
6	1	3	2	40 GB	2.0 GHz	4 hours	$1,000
7	1	4	1	80 GB	2.0 GHz	4 hours	$1,500
8	1	4	2	40 GB	1.5 GHz	6 hours	$1,200
9	1	5	1	40 GB	2.0 GHz	6 hours	$1,200

3. Select Choice ID for Profile ID, and **ADD** Hard Disk, Speed, Battery Life, and Price for the model effects.
4. Select Survey and Choice Set as the **Grouping** columns. The Profile Data dialog box is shown in Figure 5.18.

Figure 5.18 Profile Data Dialog Box for Laptop Study



Choice Dialog

Profile Data

Select Data Table: Laptop Profile

Select Columns: Survey, Choice Set, Choice ID, Hard Disk, Speed, Battery Life, Price

Pick Role Variables

Profile ID: Choice ID

Grouping: Survey, Choice Set, optional

Construct Model Effects: Hard Disk, Speed, Battery Life, Price

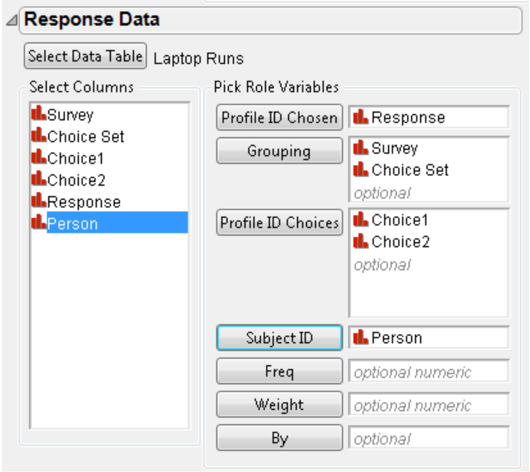
Degree: 2

☑ Firth Bias-adjusted Estimates

5. Click **Response Data** > **Select Data Table** > **Other** > **OK** and select Laptop Runs.jmp from the sample data directory.

6. Select Response as the **Profile ID Chosen**, Choice1, and Choice2 as the **Profile ID Choices**, Survey and Choice Set as the **Grouping** columns, and Person as **Subject ID**. The Response Data dialog box is shown in Figure 5.19.

Figure 5.19 Response Data Dialog Box for Laptop Study



Response Data

Select Data Table: Laptop Runs

Select Columns: Survey, Choice Set, Choice1, Choice2, Response, Person

Pick Role Variables

Profile ID Chosen: Response

Grouping: Survey, Choice Set, optional

Profile ID Choices: Choice1, Choice2, optional

Subject ID: Person

Freq: optional numeric

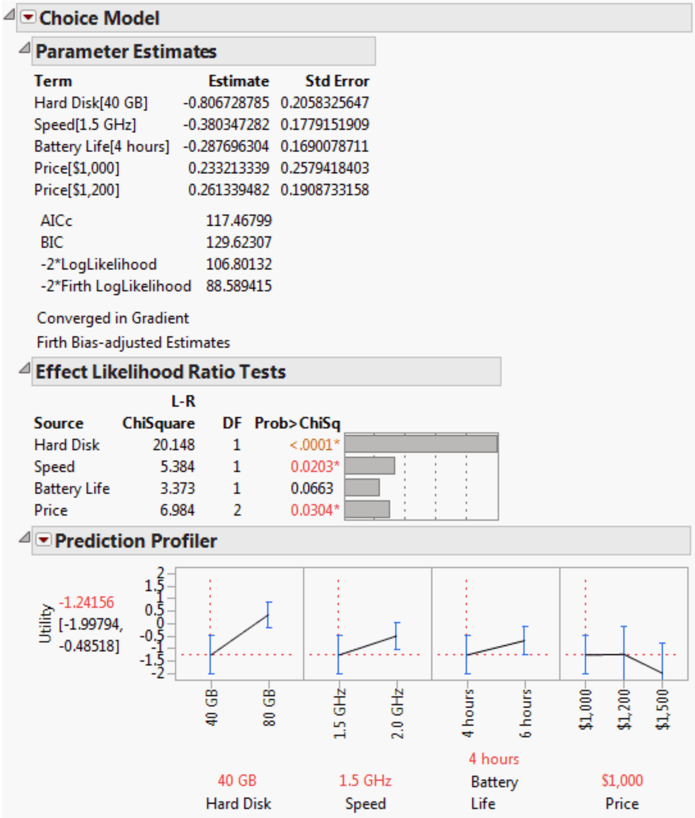
Weight: optional numeric

By: optional

7. To run the model without subject effects, click **Run Model**.

8. Choose **Profiler** from the red triangle menu.

Figure 5.20 Laptop Results without Subject Effects



Choice Model

Parameter Estimates

Term	Estimate	Std Error
Hard Disk[40 GB]	-0.806728785	0.2058325647
Speed[1.5 GHz]	-0.380347282	0.1779151909
Battery Life[4 hours]	-0.287696304	0.1690078711
Price[$1,000]	0.233213339	0.2579418403
Price[$1,200]	0.261339482	0.1908733158

AICc	117.46799
BIC	129.62307
-2*LogLikelihood	106.80132
-2*Firth LogLikelihood	88.589415

Converged in Gradient

Firth Bias-adjusted Estimates

Effect Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Hard Disk	20.148	1	<.0001*
Speed	5.384	1	0.0203*
Battery Life	3.373	1	0.0663
Price	6.984	2	0.0304*

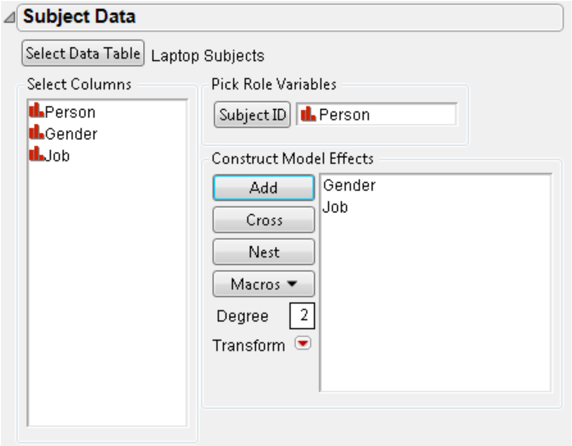
Prediction Profiler

Results of this study show that while all the factors are important, the most important factor in the laptop study is Hard Disk. The respondents prefer the larger size. Note that respondents did not think a price increase from $1000 to $1200 was important, but an increase from $1200 to $1500 was considered important. This effect is easily visualized by examining the factors interactively with the Prediction Profiler. Such a finding can have implications for pricing policies, depending on external market forces.

To include subject effect for the laptop study, simply add to the Choice Modeling dialog box:

1. Under Subject Data, **Select Data Table** > **Other** > **OK** > Laptop Subjects.jmp.
2. Select Person as **Subject ID** and Gender and Job as the model effects. The Subject Data dialog box is shown in Figure 5.21.

Figure 5.21 Subject Dialog Box for Laptop Study



Subject Data

Select Data Table: Laptop Subjects

Select Columns: Person, Gender, Job

Pick Role Variables: Subject ID: Person

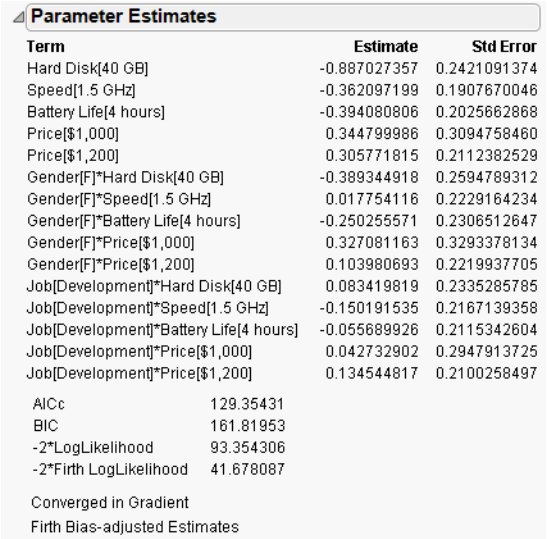
Construct Model Effects: Gender, Job

Degree: 2

3. Click **Run Model**.

Results are shown in Figure 5.22, Figure 5.23, and Figure 5.26.

Figure 5.22 Laptop Parameter Estimate Results with Subject Data



Parameter Estimates

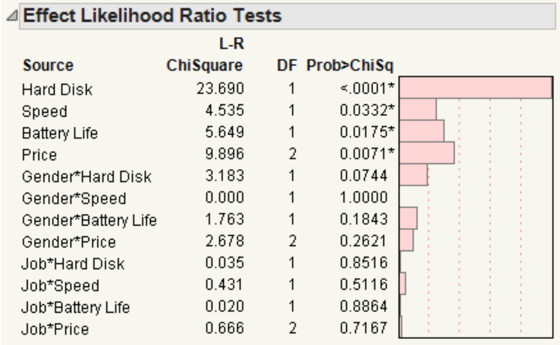
Term	Estimate	Std Error
Hard Disk[40 GB]	-0.887027357	0.2421091374
Speed[1.5 GHz]	-0.362097199	0.1907670046
Battery Life[4 hours]	-0.394080806	0.2025662868
Price[$1,000]	0.344799986	0.3094758460
Price[$1,200]	0.305771815	0.2112382529
Gender[F]*Hard Disk[40 GB]	-0.389344918	0.2594789312
Gender[F]*Speed[1.5 GHz]	0.017754116	0.2229164234
Gender[F]*Battery Life[4 hours]	-0.250255571	0.2306512647
Gender[F]*Price[$1,000]	0.327081163	0.3293378134
Gender[F]*Price[$1,200]	0.103980693	0.2219937705
Job[Development]*Hard Disk[40 GB]	0.083419819	0.2335285785
Job[Development]*Speed[1.5 GHz]	-0.150191535	0.2167139358
Job[Development]*Battery Life[4 hours]	-0.055689926	0.2115342604
Job[Development]*Price[$1,000]	0.042732902	0.2947913725
Job[Development]*Price[$1,200]	0.134544817	0.2100258497

AICc	129.35431
BIC	161.81953
-2*LogLikelihood	93.354306
-2*Firth LogLikelihood	41.678087

Converged in Gradient

Firth Bias-adjusted Estimates

Figure 5.23 Laptop Likelihood Ratio Test Results with Subject Data

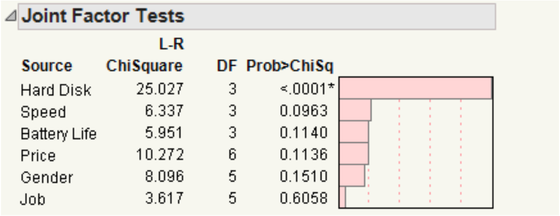


Effect Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Hard Disk	23.690	1	<.0001*
Speed	4.535	1	0.0332*
Battery Life	5.649	1	0.0175*
Price	9.896	2	0.0071*
Gender*Hard Disk	3.183	1	0.0744
Gender*Speed	0.000	1	1.0000
Gender*Battery Life	1.763	1	0.1843
Gender*Price	2.678	2	0.2621
Job*Hard Disk	0.035	1	0.8516
Job*Speed	0.431	1	0.5116
Job*Battery Life	0.020	1	0.8864
Job*Price	0.666	2	0.7167

4. Selecting **Joint Factor Tests** from the platform menu gives the table shown in Figure 5.24.

Figure 5.24 Joint Factor Test for Laptop

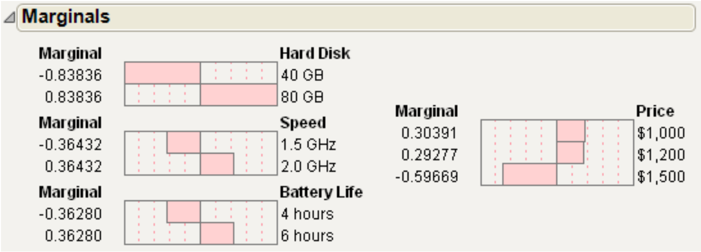


Joint Factor Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Hard Disk	25.027	3	<.0001*
Speed	6.337	3	0.0963
Battery Life	5.951	3	0.1140
Price	10.272	6	0.1136
Gender	8.096	5	0.1510
Job	3.617	5	0.6058

5. Selecting **Effect Marginals** from the platform menu displays the table shown in Figure 5.25. The marginal effects of each level for each factor are displayed. Notice that the marginal effects for each factor across all levels sum to zero.

Figure 5.25 Marginal Effects for Laptop



Marginals

Marginal	Hard Disk
-0.83836	40 GB
0.83836	80 GB

Marginal	Speed
-0.36432	1.5 GHz
0.36432	2.0 GHz

Marginal	Battery Life
-0.36280	4 hours
0.36280	6 hours

Marginal	Price
0.30391	$1,000
0.29277	$1,200
-0.59669	$1,500

Figure 5.26 Laptop Profiler Results for Females with Subject Data

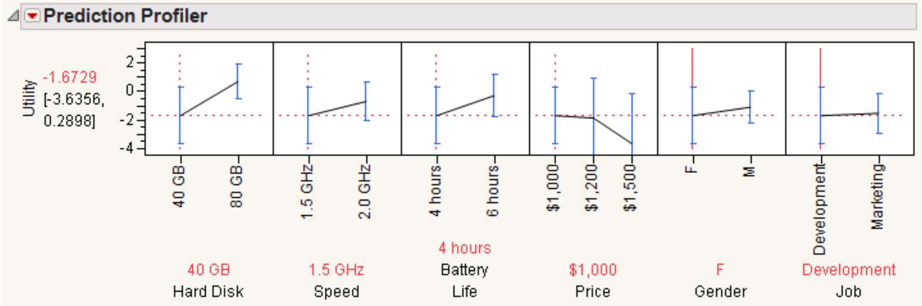
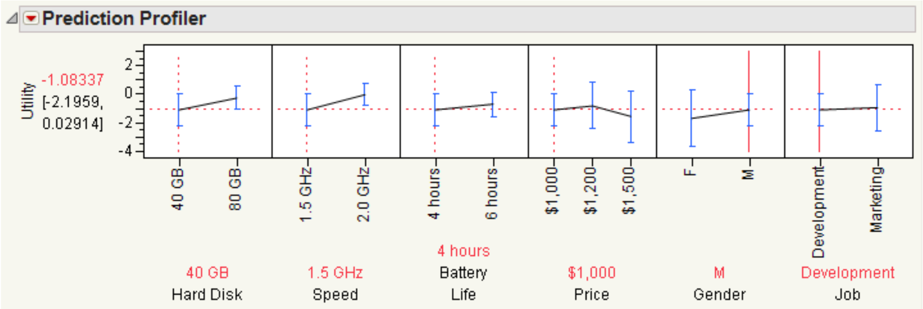


Figure 5.27 Laptop Profiler Results for Males with Subject Data



The interaction effect between Gender and Hard Disk is marginally significant, with a *p*-value of 0.0744 (See Figure 5.23 on page 102). In the Prediction Profiler, check the slope for Hard Disk for both levels of Gender. You see that the slope is steeper for females than for males.

One-Table Analysis

The Choice Modeling platform can also be used if all of your data are in one table. For this one-table scenario, you use only the Profile Data section of the Choice Dialog box. Subject-specific terms can be used in the model, but not as main effects. Two advantages, both offering more model-effect flexibility than the three-table specification, are realized by using a one-table analysis:

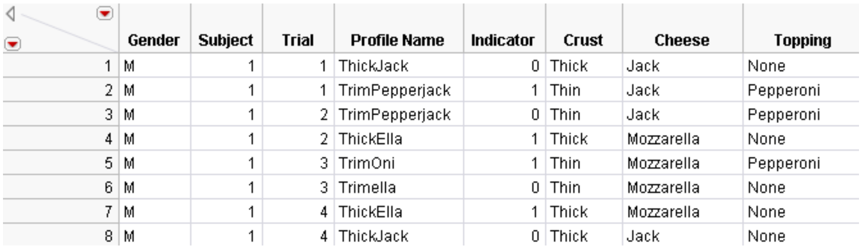
- Interactions can be selectively chosen instead of automatically getting all possible interactions between subject and profile effects as seen when using three tables.
- Unusual combinations of choice sets are allowed. This means, for example, that the first trial can have a choice set of two, the second trial can consist of a choice set of three, the third trial can have a choice set of five, and so on. With multiple tables, in contrast, it is assumed that the number of choices for each trial is fixed.

A choice response consists of a set of rows, uniquely identified by the Grouping columns. An indicator column is specified for Profile ID in the Choice Dialog box. This indicator variable uses the value of 1 for the chosen profile row and 0 elsewhere. There must be exactly one “1” for each Grouping combination.

Example: One-Table Pizza Data

This example illustrates how the pizza data are organized for the one-table situation. Figure 5.28 shows a subset of the combined pizza data. Open Pizza Combined.jmp from the sample data directory to see the complete table. Each subject completes four choice sets, with each choice set or trial consisting of two choices. For this example, each subject has eight rows in the data set. The indicator variable specifies the chosen profile for each choice set. The columns Subject and Trial together identify the choice set, so they are the **Grouping** columns.

Figure 5.28 Partial Listing of Combined Pizza Data for One-Table Analysis

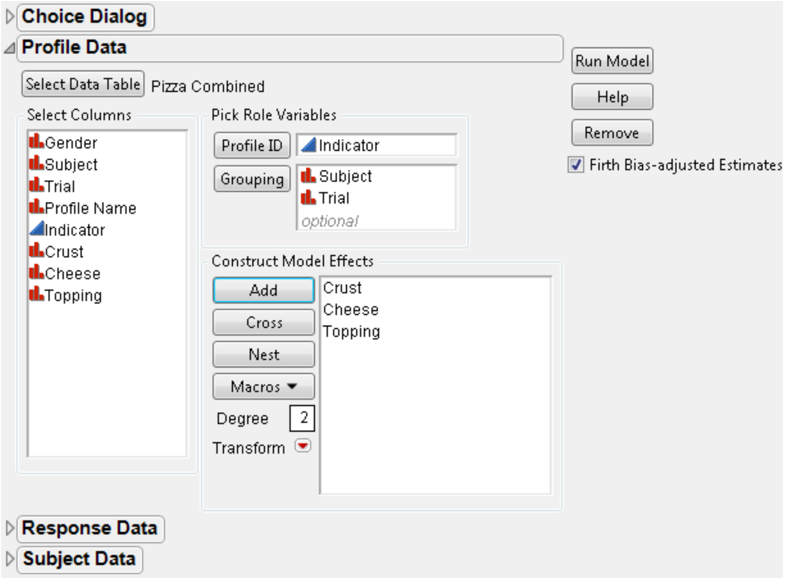


	Gender	Subject	Trial	Profile Name	Indicator	Crust	Cheese	Topping
1	M	1	1	ThickJack	0	Thick	Jack	None
2	M	1	1	TrimPepperjack	1	Thin	Jack	Pepperoni
3	M	1	2	TrimPepperjack	0	Thin	Jack	Pepperoni
4	M	1	2	ThickElla	1	Thick	Mozzarella	None
5	M	1	3	TrimOni	1	Thin	Mozzarella	Pepperoni
6	M	1	3	Trimella	0	Thin	Mozzarella	None
7	M	1	4	ThickElla	1	Thick	Mozzarella	None
8	M	1	4	ThickJack	0	Thick	Jack	None

To analyze the data in this format, open the Profile Data section in the Choice Dialog box, shown in Figure 5.29.

1. Specify Pizza Combined.jmp as the data set.
2. Specify Indicator as the **Profile ID**, Subject and Trial as the **Grouping** variables, and add Crust, Cheese, and Topping as the main effects.
3. Click **Run Model**.

Figure 5.29 Choice Dialog Box for Pizza Data One-Table Analysis



Choice Dialog

Profile Data

Select Data Table: Pizza Combined

Select Columns: Gender, Subject, Trial, Profile Name, Indicator, Crust, Cheese, Topping

Pick Role Variables

Profile ID: Indicator

Grouping: Subject, Trial, *optional*

Construct Model Effects: Crust, Cheese, Topping

Degree: 2

☑ Firth Bias-adjusted Estimates

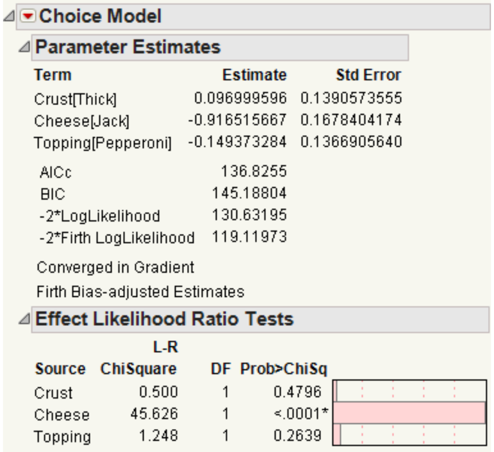
Response Data

Subject Data

A new dialog box appears asking if this is a one-table analysis with all of the data in the Profile Table.

4. Click **Yes** to fit the model, as shown in Figure 5.30.

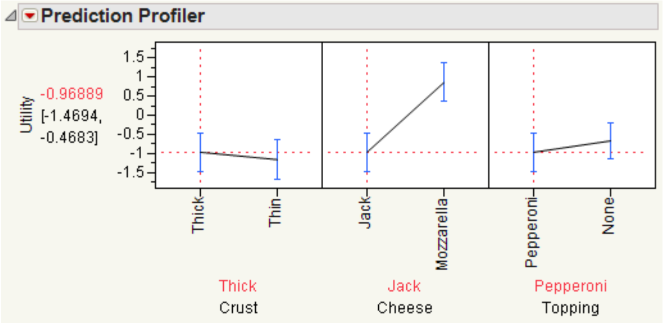
Figure 5.30 Choice Model for Pizza Data One-Table Analysis



```
Choice Model
 Parameter Estimates
 Term                    Estimate      Std Error
 Crust[Thick]          0.096999596   0.1390573555
 Cheese[Jack]         -0.916515667   0.1678404174
 Topping[Pepperoni]   -0.149373283   0.1366905640

  AICc                    136.8255
  BIC                    145.18804
  -2*LogLikelihood       130.63195
  -2*Firth LogLikelihood 119.11973

  Converged in Gradient
  Firth Bias-adjusted Estimates
 Effect Likelihood Ratio Tests
                  L-R
 Source   ChiSquare   DF  Prob>ChiSq
 Crust        0.500    1      0.4796
 Cheese      45.626    1     <.0001*
 Topping      1.248    1      0.2639
```

5. Select **Profiler** from the drop-down menu to obtain the results shown in Figure 5.31. Notice that the parameter estimates and the likelihood ratio test results are identical to the results obtained for the Choice Model with only two tables, shown in Figure 5.7.

Figure 5.31 Prediction Profiler for Pizza Data One-Table Analysis



Segmentation

Market researchers sometimes want to analyze the preference structure for each subject separately in order to see whether there are groups of subjects that behave differently. However, there are usually not enough data to do this with ordinary estimates. If there are sufficient data, you can specify “By groups” in the Response Data or you could introduce a Subject identifier as a subject-side model term. This approach, however, is costly if the number of subjects is large. Other segmentation techniques discussed in the literature include Bayesian and mixture methods.

You can also use JMP to segment by clustering subjects using response data. For example, after running the model using the Pizza Profiles.jmp, Pizza Responses.jmp, and the optional Pizza Subjects.jmp data sets, select the drop-down menu for the Choice Model platform and select **Save Gradients by Subject**. A new data table is created containing the average Hessian-scaled gradient on each parameter, and there is one row for each subject.

Note: This feature is regarded as an experimental method, because, in practice, little research has been conducted on its effectiveness.

These gradient values are the subject-aggregated Newton-Raphson steps from the optimization used to produce the estimates. At the estimates, the total gradient is zero, and

$\Delta = \mathrm{H}^{-1}g = 0$ where g is the total gradient of the log-likelihood evaluated at the MLE, and

H^{-1} is the inverse Hessian function or the inverse of the negative of the second partial derivative of the log-likelihood.

But, the disaggregation of Δ results in

$$\Delta = \Sigma_{ij}\Delta_{ij} = \Sigma\mathrm{H}^{-1}g_{ij} = 0$$

where i is the subject index, j is the choice response index for each subject,

Δij are the partial Newton-Raphson steps for each run, and

g_{ij} is the gradient of the log-likelihood by run.

The mean gradient step for each subject is then calculated as:

$\bar{\Delta}_i = \Sigma_j \frac{\Delta_{ij}}{n_i}$ where n_i is the number of runs per subject.

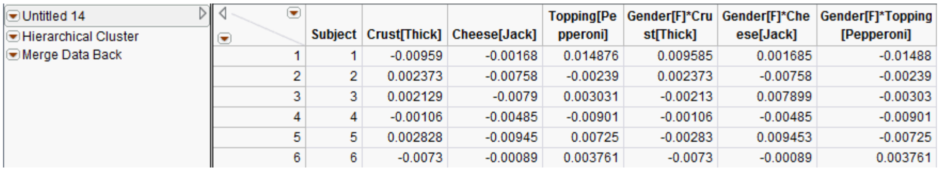
These $\overline{\Delta}_i$ are related to the force that subject i is applying to the parameters.

If groups of subjects have truly different preference structures, these forces are strong, and they can be used to cluster the subjects.

The $\overline{\Delta}_i$ are the gradient forces that are saved.

A partial data table with these subject forces is shown in Figure 5.32.

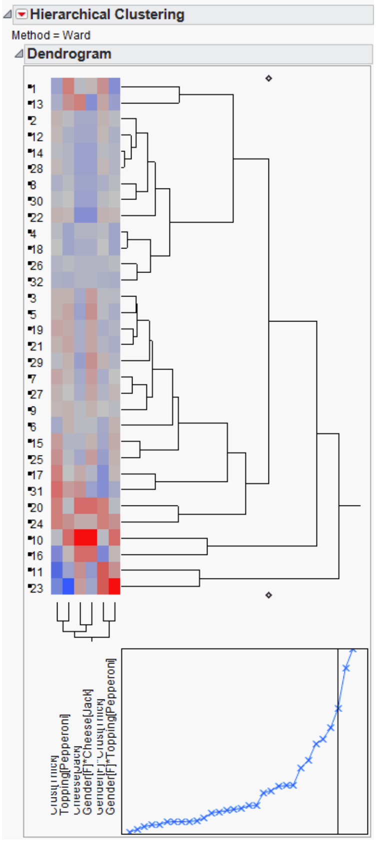
Figure 5.32 Gradients by Subject for Pizza Data



	Subject	Crust[Thick]	Cheese[Jack]	Topping[Pepperoni]	Gender[F]*Crust[Thick]	Gender[F]*Cheese[Jack]	Gender[F]*Topping [Pepperoni]
1	1	-0.00959	-0.00168	0.014876	0.009585	0.001685	-0.01488
2	2	0.002373	-0.00758	-0.00239	0.002373	-0.00758	-0.00239
3	3	0.002129	-0.0079	0.003031	-0.00213	0.007899	-0.00303
4	4	-0.00106	-0.00485	-0.00901	-0.00106	-0.00485	-0.00901
5	5	0.002828	-0.00945	0.00725	-0.00283	0.009453	-0.00725
6	6	-0.0073	-0.00089	0.003761	-0.0073	-0.00089	0.003761

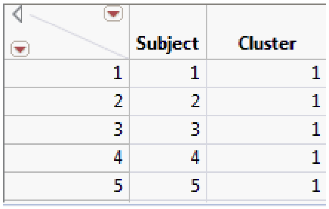
You can cluster these values by clicking on the drop-down menu of **Hierarchical Clustering** in the new data table and selecting **Run Script**. The resulting dendrogram of the clusters is shown in Figure 5.33.

Figure 5.33 Dendrogram of Subject Clusters for Pizza Data



Now, select the number of clusters desired by moving the diamond indicator at the top or bottom of the dendrogram. Alternatively, you can select **Number of Clusters** in the platform drop-down menu and enter a number. You can save the cluster IDs by clicking on the drop-down menu of Hierarchical Clustering and selecting **Save Clusters**. A new column called Cluster is created in the data table containing the gradients. Each subject has been assigned a Cluster value that is associated with other subjects having similar gradient forces. Refer to the Cluster platform chapter in the *Multivariate Methods* book for a discussion of other Hierarchical Clustering options. The gradient columns can be deleted because they were used only to obtain the clusters. Your data table then contains only Subject and Cluster variables.

Figure 5.34 Merge Clusters Back into Original Table

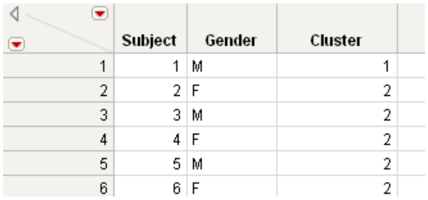


	Subject	Cluster
1	1	1
2	2	1
3	3	1
4	4	1
5	5	1

If you click **Run Script** under the **Merge Data Back** menu, as shown in the partial gradient-by-subject table in Figure 5.32, the cluster information becomes a part of the Subject data table.

The columns in the Subject data table are now Subject, Gender, and Cluster, as shown in Figure 5.35.

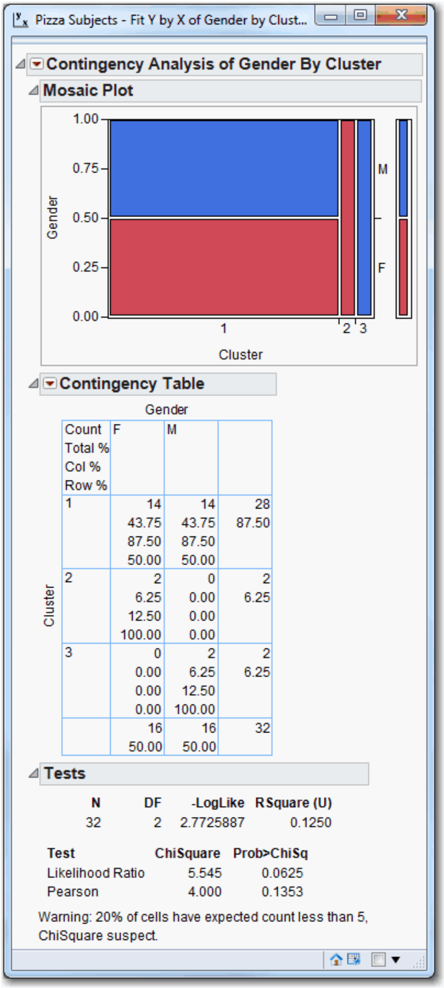
Figure 5.35 Subject Data with Cluster Column



	Subject	Gender	Cluster
1	1	M	1
2	2	F	2
3	3	M	2
4	4	F	2
5	5	M	2
6	6	F	2

This table can then be used for further analysis. For example, select **Analyze** > **Fit Y by X**. Then, specify Gender as the **Y, Response**, and Cluster as **X, Factor**. For the pizza example, this analysis is depicted in Figure 5.36.

Figure 5.36 Contingency Analysis of Gender by Cluster for Pizza Example


Pizza Subjects - Fit Y by X of Gender by Clust...
Contingency Analysis of Gender By Cluster
Mosaic Plot
Gender
1.00
0.75
0.50
0.25
0.00
M
F
1
2
3
Cluster

Contingency Table

Count Total % Col % Row %	Gender F	 M	
Cluster 1	14 43.75 87.50 50.00	14 43.75 87.50 50.00	28 87.50
2	2 6.25 12.50 100.00	0 0.00 0.00 0.00	2 6.25
3	0 0.00 0.00 0.00	2 6.25 12.50 100.00	2 6.25
	16 50.00	16 50.00	32

Tests

```
 N     DF   -LogLike   RSquare (U)
32      2  2.7725887      0.1250

Test               ChiSquare   Prob>ChiSq
Likelihood Ratio       5.545       0.0625
Pearson                4.000       0.1353

Warning: 20% of cells have expected count less than 5,
ChiSquare suspect.
```

Figure 5.36 shows that Cluster 1 contains half male and female, Cluster 2 is only female, and Cluster 3 is all male. If desired, you could now refit and analyze the model with the addition of the Cluster variable.

Special Data Rules

Default Choice Set

If in every trial, you can choose any of the response profiles, you can omit the **Profile ID Choices** selection under **Pick Role Variables** in the Response Data section of the Choice Dialog

Box. Then the Choice Model platform assumes that all choice profiles are available on each run.

Subject Data with Response Data

If you have subject data in the Response data table, just select this table as the **Select Data Table** under the Subject Data. In this case, a **Subject ID** column does not need to be specified. In fact, it is not used. It is generally assumed that the subject data repeats consistently in multiple runs for each subject.

Logistic Regression

Ordinary logistic regression can be done with the Choice Modeling platform.

Note: The Fit Y by X and Fit Model platforms are more convenient to use than the Choice Modeling platform for logistic regression modeling. This section is used only to demonstrate that the Choice Modeling platform can be used for logistic regression, if desired.

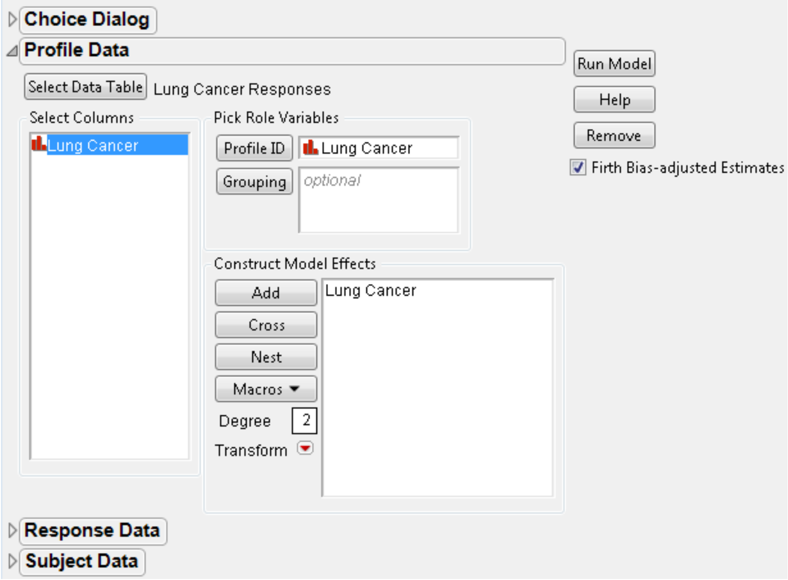
If your data are already in the choice-model format, you might want to use the steps given below for logistic regression analysis. However, three steps are needed:

- Create a trivial Profile data table with a row for each response level.
- Put the explanatory variables into the Response data.
- Specify the Response data table, again, for the Subject data table.

An example of using the Choice Modeling platform for logistic regression follows:

1. Select **Analyze** > **Consumer Research** > **Choice** > **Select Data Table** > **Other** > **OK**.
2. Open the sample data set Lung Cancer Responses.jmp. Notice this data table has only one column (Lung Cancer) with two rows (Cancer and NoCancer).
3. Select Lung Cancer as the **Profile ID** and **Add** Lung Cancer as the model effect. The Profile Data dialog box is shown in Figure 5.37.

Figure 5.37 Profile Data for Lung Cancer Example



Profile Data

Select Data Table: Lung Cancer Responses

Select Columns: Lung Cancer

Pick Role Variables

Profile ID: Lung Cancer

Grouping: optional

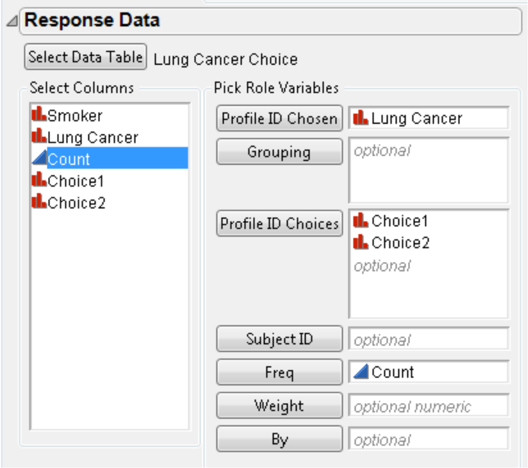
Construct Model Effects: Lung Cancer

Degree: 2

☑ Firth Bias-adjusted Estimates

4. Click the disclosure icon for **Response Data** > **Select Data Table** > **Other** > OK.
5. Open the sample data set Lung Cancer Choice.jmp.
6. Select Lung Cancer for **Profile ID Chosen**, Choice1 and Choice2 for **Profile ID Choices**, and Count for **Freq**. The Response Data launch dialog box is shown in Figure 5.38.

Figure 5.38 Response Data for Lung Cancer Example



Response Data

Select Data Table: Lung Cancer Choice

Select Columns:
Smoker
Lung Cancer
Count
Choice1
Choice2

Pick Role Variables

Profile ID Chosen: Lung Cancer

Grouping: optional

Profile ID Choices: Choice1, Choice2, optional

Subject ID: optional

Freq: Count

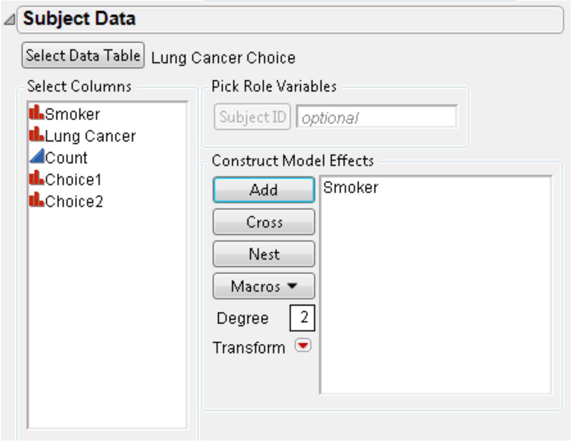
Weight: optional numeric

By: optional

7. Click the disclosure icon for **Subject Data** > **Select Data Table** > Lung Cancer Choice.jmp > **OK**.

8. Add Smoker as the model effect. The Subject Data launch dialog box is shown in Figure 5.39.

Figure 5.39 Subject Data for Lung Cancer Example



Lung Cancer Choice

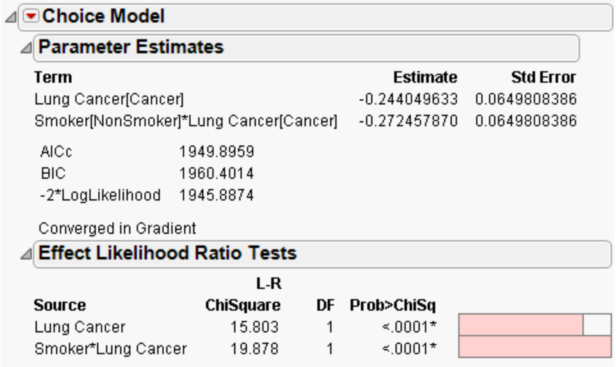
Degree 2

Construct Model Effects: Smoker

9. Uncheck **Firth Bias-adjusted Estimates** and **Run Model**.

   Choice Modeling results are shown in Figure 5.40.

Figure 5.40 Choice Modeling Logistic Regression Results for the Cancer Data

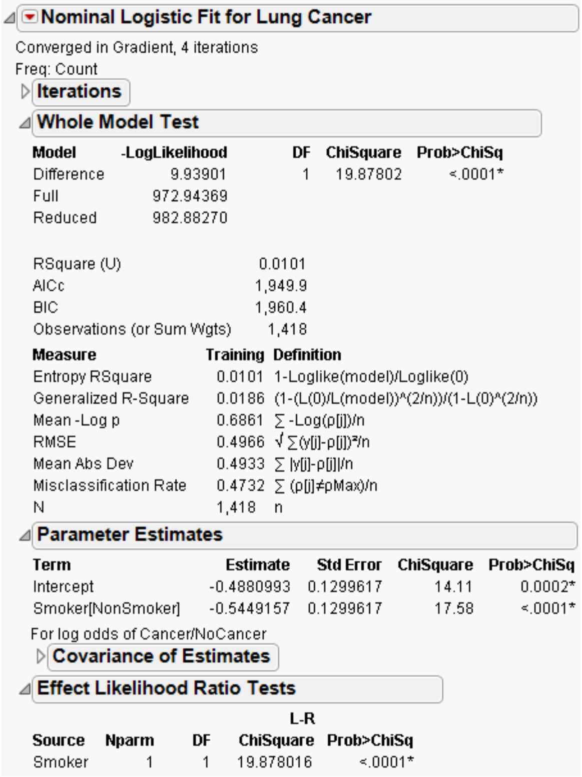


```
Choice Model
 Parameter Estimates
 Term                                  Estimate     Std Error
 Lung Cancer[Cancer]                 -0.244049633  0.0649808386
 Smoker[NonSmoker]*Lung Cancer[Cancer] -0.272457870  0.0649808386

  AICc              1949.8959
  BIC               1960.4014
  -2*LogLikelihood  1945.8874

  Converged in Gradient
 Effect Likelihood Ratio Tests
                         L-R
 Source               ChiSquare   DF  Prob>ChiSq
 Lung Cancer             15.803    1     <.0001*
 Smoker*Lung Cancer      19.878    1     <.0001*
```

Compare these results with those of logistic regression under the Fit Model platform:

1. Open Lung Cancer.jmp in the sample data directory.
2. Select **Analyze** > **Fit Model**. Automatic specification of the columns is: Lung Cancer for *Y*, Count for **Freq**, and Smoker for **Add** under **Construct Model Effects**. The **Nominal Logistic** personality is automatically selected.
3. Click **Run**. The nominal logistic fit for the data is shown in Figure 5.41.

Figure 5.41 Fit Model Nominal Logistic Regression Results for the Cancer Data



```
Nominal Logistic Fit for Lung Cancer
Converged in Gradient, 4 iterations
Freq: Count
Iterations
Whole Model Test
Model        -LogLikelihood      DF   ChiSquare   Prob>ChiSq
Difference          9.93901       1    19.87802      <.0001*
Full              972.94369
Reduced           982.88270

RSquare (U)                       0.0101
AICc                             1,949.9
BIC                              1,960.4
Observations (or Sum Wgts)         1,418
Measure                  Training  Definition
Entropy RSquare            0.0101  1-Loglike(model)/Loglike(0)
Generalized R-Square       0.0186  (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p                0.6861  ∑ -Log(p[j])/n
RMSE                       0.4966  √ ∑(y[j]-p[j])²/n
Mean Abs Dev               0.4933  ∑ |y[j]-p[j]|/n
Misclassification Rate     0.4732  ∑ (p[j]≠pMax)/n
N                           1,418  n
Parameter Estimates
Term                   Estimate    Std Error  ChiSquare  Prob>ChiSq
Intercept            -0.4880993    0.1299617      14.11     0.0002*
Smoker[NonSmoker]    -0.5449157    0.1299617      17.58     <.0001*
For log odds of Cancer/NoCancer
Covariance of Estimates
Effect Likelihood Ratio Tests
                                   L-R
Source   Nparm     DF        ChiSquare   Prob>ChiSq
Smoker       1      1        19.878016      <.0001*
```

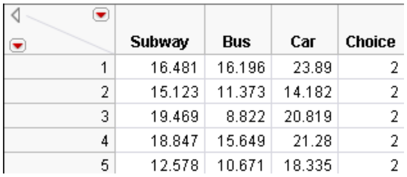
Notice that the likelihood ratio chi-square test for Smoker*Lung Cancer in the Choice model matches the likelihood ratio chi-square test for Smoker in the Logistic model. The reports shown in Figure 5.40 and Figure 5.41 support the conclusion that smoking has a strong effect on developing lung cancer. See the Logistic Regression chapter in the *Fitting Linear Models* book for details.

Transforming Data

Transforming Data to Two Analysis Tables

Although data are often in the Response/Profile/Subject form, the data are sometimes specified in another format that must be manipulated into the normalized form needed for choice analysis. For example, consider the data from Daganzo, found in Daganzo Trip.jmp. This data set contains the travel time for three transportation alternatives and the preferred transportation alternative for each subject. A partial listing of the data set is shown in Figure 5.42.

Figure 5.42 Partial Daganzo Travel Time Table for Three Alternatives

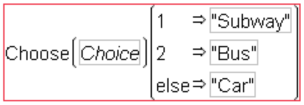


	Subway	Bus	Car	Choice
1	16.481	16.196	23.89	2
2	15.123	11.373	14.182	2
3	19.469	8.822	20.819	2
4	18.847	15.649	21.28	2
5	12.578	10.671	18.335	2

Each Choice number listed must first be converted to one of the travel mode names. This transformation is easily done by using the **Choose** function in the formula editor, as follows:

1. Create a new column labeled Choice Mode. Specify the modeling type as **Nominal**. Right-click the Choice Mode column and select **Formula**.
2. Click **Conditional** under the **Functions (grouped)** command, select **Choose**, and press the comma key twice to obtain additional arguments for the function.
3. Click Choice for the Choose expression (expr), and double click each clause entry box to enter "Subway"," Bus", and "Car" (with the quotation marks) as shown in Figure 5.43.

Figure 5.43 Choose Function for Choice Mode Column of Daganzo Data

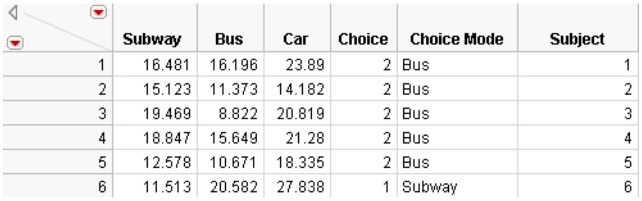


```
              1    ⇒ "Subway"
Choose(Choice) 2    ⇒ "Bus"
              else ⇒ "Car"
```

The choice response is now in the correct format.

4. Because each row contains a choice made by each subject, another column containing a sequence of numbers should be created to identify the subjects. This can be done by creating a column with the Subject label. Then, enter a 1 in the first row of the column, a 2 in the second row of the column. Finally, highlight the first and second rows of the column, right-click, and select **Fill > Continue sequence to end of table**. A partial listing of the modified table is shown in Figure 5.44.

Figure 5.44 Daganzo Data with New Choice Mode and Subject Columns

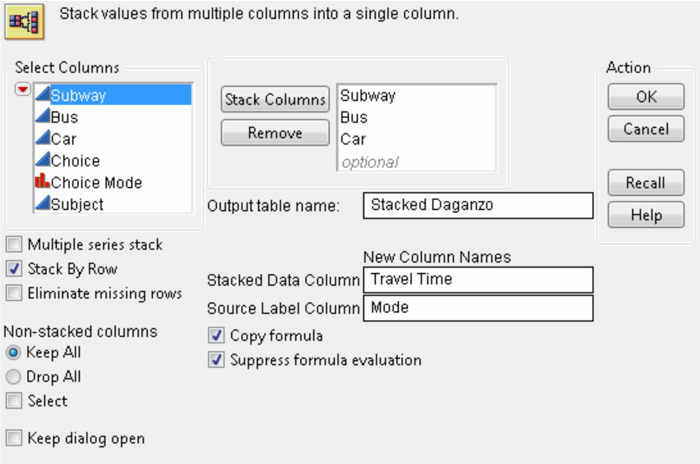


	Subway	Bus	Car	Choice	Choice Mode	Subject
1	16.481	16.196	23.89	2	Bus	1
2	15.123	11.373	14.182	2	Bus	2
3	19.469	8.822	20.819	2	Bus	3
4	18.847	15.649	21.28	2	Bus	4
5	12.578	10.671	18.335	2	Bus	5
6	11.513	20.582	27.838	1	Subway	6

In order to construct the Profile data, each alternative needs to be expressed in a separate row.

5. Use the Stack operation by clicking on **Tables** > **Stack** and filling in the entry fields as shown in Figure 5.45. Give this new data table a name, such as Stacked Daganzo.jmp, so that you can use this table for future analyses. Click **OK**. A partial view of the resulting table is shown in Figure 5.46.

Figure 5.45 Stack Operation for Daganzo Data



Stack values from multiple columns into a single column.

Select Columns: Subway, Bus, Car, Choice, Choice Mode, Subject

Stack Columns: Subway, Bus, Car, *optional*

Output table name: Stacked Daganzo

□ Multiple series stack
☑ Stack By Row
□ Eliminate missing rows

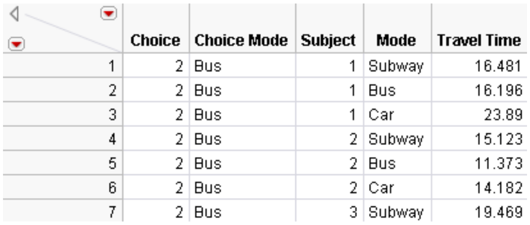
New Column Names
Stacked Data Column: Travel Time
Source Label Column: Mode

Non-stacked columns
● Keep All
○ Drop All
□ Select

☑ Copy formula
☑ Suppress formula evaluation

□ Keep dialog open

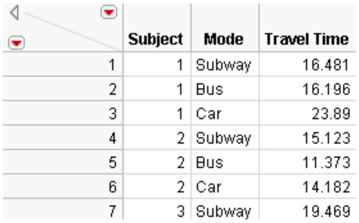
Figure 5.46 Partial Stacked Daganzo Table



	Choice	Choice Mode	Subject	Mode	Travel Time
1	2	Bus	1	Subway	16.481
2	2	Bus	1	Bus	16.196
3	2	Bus	1	Car	23.89
4	2	Bus	2	Subway	15.123
5	2	Bus	2	Bus	11.373
6	2	Bus	2	Car	14.182
7	2	Bus	3	Subway	19.469

6. Make a subset of the stacked data with just Subject, Mode, and Travel Time by selecting these columns and selecting **Tables** > **Subset**. Select **Selected Columns** and click **OK**. A partial data table is shown in Figure 5.47.

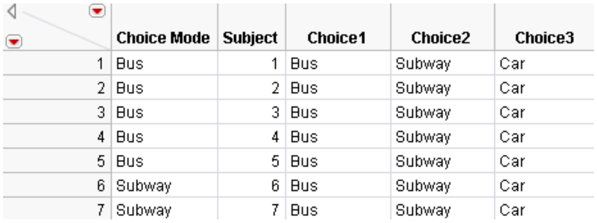
Figure 5.47 Partial Subset Table of Stacked Daganzo Data



	Subject	Mode	Travel Time
1	1	Subway	16.481
2	1	Bus	16.196
3	1	Car	23.89
4	2	Subway	15.123
5	2	Bus	11.373
6	2	Car	14.182
7	3	Subway	19.469

7. Make another subset of the original data with just Subject and Choice Mode. Then, add three constant columns for the choice set: Choice1, Choice2, and Choice3, as shown in Figure 5.48.

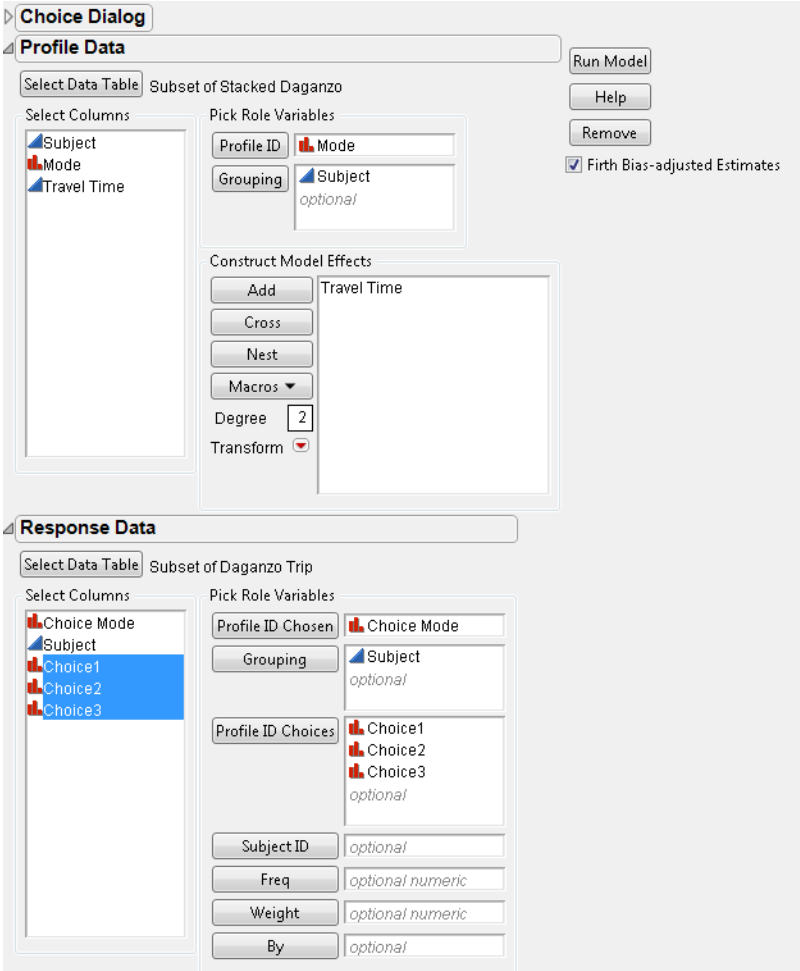
Figure 5.48 Partial Subset Table of Daganzo Data with Choice Set



	Choice Mode	Subject	Choice1	Choice2	Choice3
1	Bus	1	Bus	Subway	Car
2	Bus	2	Bus	Subway	Car
3	Bus	3	Bus	Subway	Car
4	Bus	4	Bus	Subway	Car
5	Bus	5	Bus	Subway	Car
6	Subway	6	Bus	Subway	Car
7	Subway	7	Bus	Subway	Car

8. Specify the model, as shown in Figure 5.49.

Figure 5.49 Choice Dialog Box for Subset of Daganzo Data



Profile Data

Select Data Table Subset of Stacked Daganzo

Select Columns
Subject
Mode
Travel Time

Pick Role Variables
Profile ID: Mode
Grouping: Subject
optional

Construct Model Effects
Travel Time
Degree 2

☑ Firth Bias-adjusted Estimates

Response Data

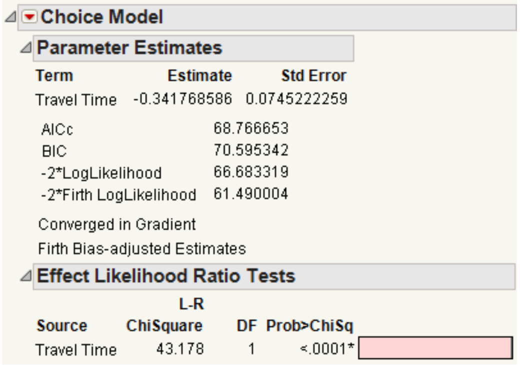
Select Data Table Subset of Daganzo Trip

Select Columns
Choice Mode
Subject
Choice1
Choice2
Choice3

Pick Role Variables
Profile ID Chosen: Choice Mode
Grouping: Subject
optional
Profile ID Choices: Choice1
Choice2
Choice3
optional
Subject ID: optional
Freq: optional numeric
Weight: optional numeric
By: optional

9. Run the model. The resulting parameter estimate now expresses the utility coefficient for Travel Time and is shown in Figure 5.50.

Figure 5.50 Parameter Estimate for Travel Time of Daganzo Data



```
Choice Model
  Parameter Estimates
  Term          Estimate       Std Error
  Travel Time   -0.341768586   0.0745222259

   AICc                       68.766653
   BIC                        70.595342
   -2*LogLikelihood           66.683319
   -2*Firth LogLikelihood     61.490004

  Converged in Gradient
  Firth Bias-adjusted Estimates
  Effect Likelihood Ratio Tests
                         L-R
  Source         ChiSquare   DF   Prob>ChiSq
  Travel Time       43.178    1      <.0001*
```

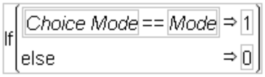
The negative coefficient implies that increased travel time has a negative effect on consumer utility or satisfaction. The likelihood ratio test result indicates that the Choice model with the effect of Travel Time is significant.

Transforming Data to One Analysis Table

Rather than creating two or three tables, it can be more practical to transform the data so that only one table is used. For the one-table format, the subject effect is added as above. A response indicator column is added instead of using three different columns for the choice sets (Choice1, Choice2, Choice3). The transformation steps for the one-table scenario include:

1. Create or open Stacked Daganzo.jmp from the steps shown in "Transforming Data to Two Analysis Tables" on page 114.
2. Add a new column labeled Response and right-click the column. Select **Formula**.
3. Select **Conditional** > **If** from the formula editor and select the column Choice Mode for the expression.
4. Enter "=" and select Mode.
5. Type 1 for the **Then Clause** and 0 for the **Else Clause**. Click **OK**. The completed formula should look like Figure 5.51.

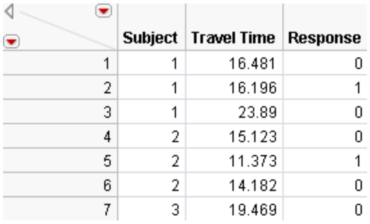
Figure 5.51 Formula for Response Indicator for Stacked Daganzo Data



```
If  Choice Mode == Mode  ⇒ 1
    else                 ⇒ 0
```

6. Subset the data table by selecting Subject, Travel Time, and Response and then select **Tables** > **Subset**. Select **Selected Columns** and click **OK**. A partial listing of the new data table is shown in Figure 5.52.

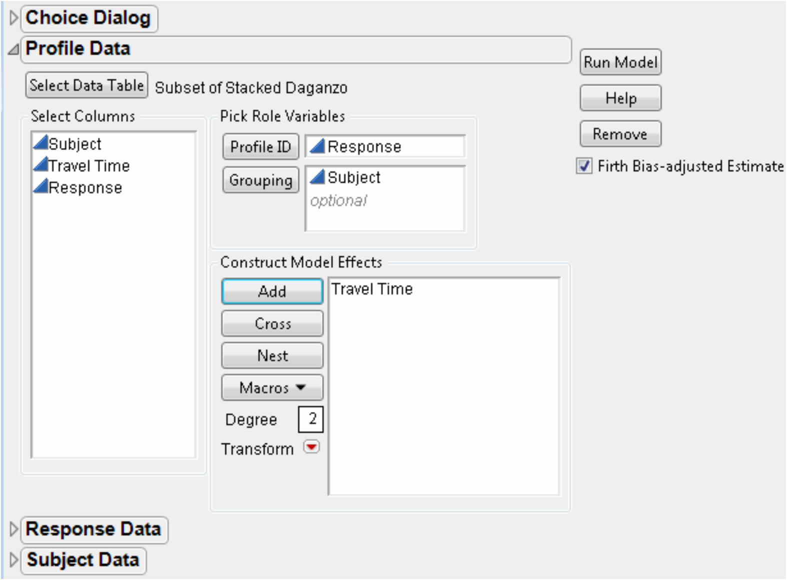
Figure 5.52 Partial Table of Stacked Daganzo Data Subset



	Subject	Travel Time	Response
1	1	16.481	0
2	1	16.196	1
3	1	23.89	0
4	2	15.123	0
5	2	11.373	1
6	2	14.182	0
7	3	19.469	0

7. Select **Analyze** > **Consumer Research** > **Choice** to open the launch dialog box and specify the model as shown in Figure 5.53.

Figure 5.53 Choice Dialog Box for Subset of Stacked Daganzo Data for One-Table Analysis



Select Data Table: Subset of Stacked Daganzo

Profile ID: Response

Grouping: Subject
optional

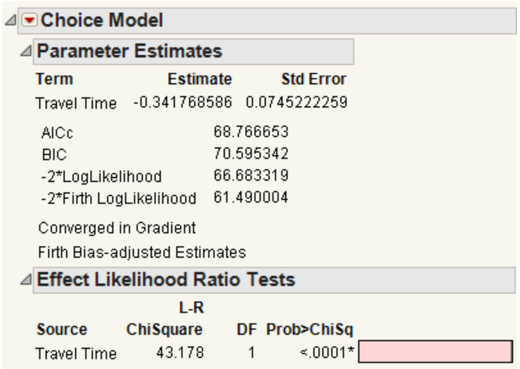
Construct Model Effects: Travel Time

Degree 2

☑ Firth Bias-adjusted Estimate

8. Select **Run Model.** A pop-up dialog window asks whether this is a one-table analysis with all the data in the Profile Table. Select **Yes** to obtain the parameter estimate expressing the utility **Travel Time** coefficient, shown in Figure 5.54.

Figure 5.54 Parameter Estimate for Travel Time of Daganzo Data from One-Table Analysis



```
Choice Model
  Parameter Estimates
  Term          Estimate      Std Error
  Travel Time   -0.341768586  0.0745222259

   AICc                     68.766653
   BIC                      70.595342
   -2*LogLikelihood         66.683319
   -2*Firth LogLikelihood   61.490004

  Converged in Gradient
  Firth Bias-adjusted Estimates
  Effect Likelihood Ratio Tests
                    L-R
  Source        ChiSquare   DF  Prob>ChiSq
  Travel Time      43.178    1     <.0001*
```

Notice that the result is identical to that obtained for the two-table model, shown earlier in Figure 5.50.

This chapter illustrates the use of the Choice Modeling platform with simple examples. This platform can also be used for more complex models, such as those involving more complicated transformations and interaction terms.

Logistic Regression for Matched Case-Control Studies

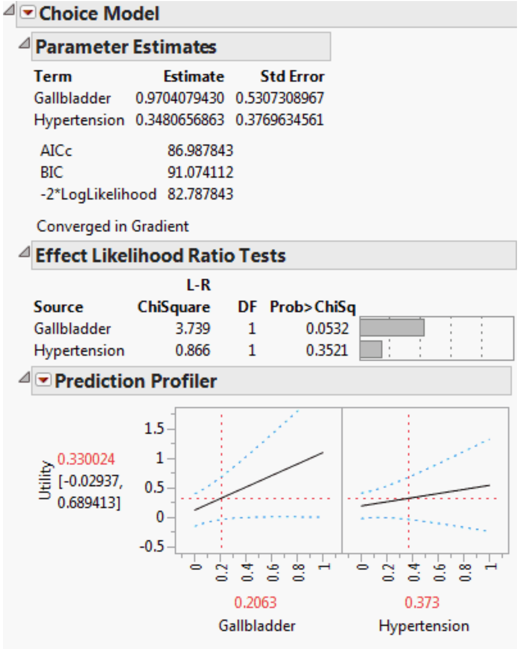
This section provides an example using the Choice platform to perform logistic regression on the results of a study of endometrial cancer with 63 matched pairs. The data are from the Los Angeles Study of the Endometrial Cancer Data in Breslow and Day (1980) and the SAS/STAT(R) 9.2 User's Guide, Second Edition (2006). The goal of the case-control analysis was to determine the relative risk for gallbladder disease, controlling for the effect of hypertension. The Outcome of 1 indicates the presence of endometrial cancer, and 0 indicates the control. Gallbladder and Hypertension data indicators are also 0 or 1. To perform the analysis, follow the steps below:

1. Open Endometrial Cancer.jmp.
2. Select **Analyze > Consumer Research > Choice**.
3. Select the **Select Data Table** button.
4. Select Endometrial Cancer as the profile data table.
5. Assign Outcome to the **Profile ID** role.
6. Assign Pair to the **Grouping** role.
7. Add the following columns as model effects: Gallbladder, Hypertension.
8. Deselect the **Firth Bias-Adjusted Estimates** check box.
9. Select **Run Model**.

10. When you are asked if this is a one-table analysis, answer **Yes**.
11. On the Choice Model red triangle menu, select **Profiler**.

    The report is shown in Figure 5.55.

Figure 5.55 Logistic Regression on Endometrial Cancer Data



Choice Model

Parameter Estimates

Term	Estimate	Std Error
Gallbladder	0.9704079430	0.5307308967
Hypertension	0.3480656863	0.3769634561

AICc	86.987843
BIC	91.074112
-2*LogLikelihood	82.787843

Converged in Gradient

Effect Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Gallbladder	3.739	1	0.0532
Hypertension	0.866	1	0.3521

Prediction Profiler

Likelihood Ratio tests are given for each factor. Note that Gallbladder is nearly significant at the 0.05 level (p-value = 0.0532). Use the Prediction Profiler to visualize the impact of the factors on the response.

Statistical Details

Parameter estimates from the choice model identify consumer *utility*, or marginal utilities in the case of a linear utility function. Utility is the level of satisfaction consumers receive from products with specific attributes and is determined from the parameter estimates in the model.

The choice statistical model is expressed as follows:

Let $X[k]$ represent a subject attribute design row, with intercept

Let $Z[j]$ represent a choice attribute design row, without intercept

Then, the probability of a given choice for the $k'th$ subject to the $j'th$ choice of m choices is:

$$P_i[jk] = \frac{\exp(\beta'(X[k] \otimes Z[j]))}{\sum_{l=1}^{m} \exp(\beta'(X[k] \otimes Z[l]))}$$

where:

- $\otimes$ is the Kronecker row-wise product
- the numerator calculates for the $j'th$ alternative actually chosen
- the denominator sums over the m choices presented to the subject for that trial

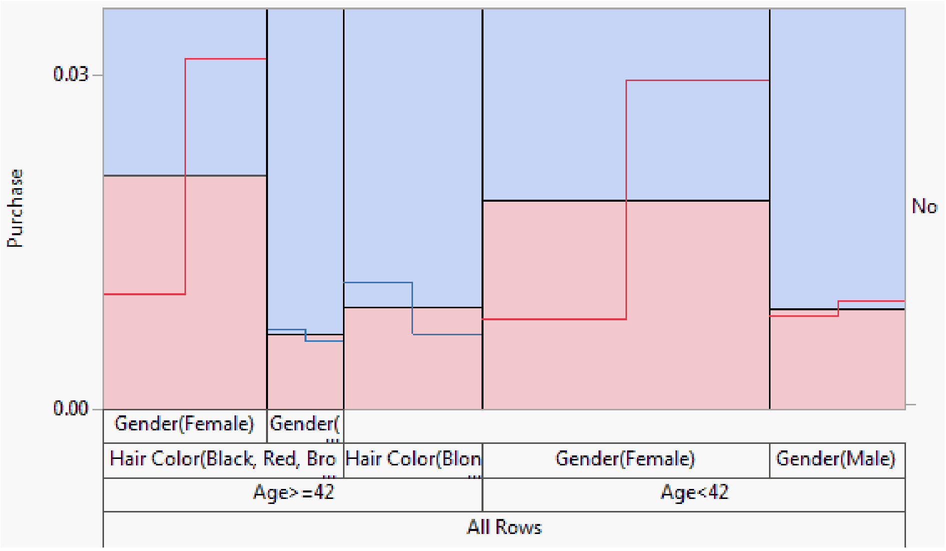
Chapter 6

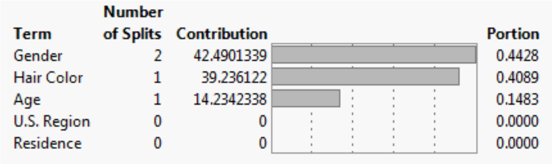
Uplift Models

Model the Incremental Impact of Actions on Consumer Behavior

Use uplift modeling to optimize marketing decisions, to define personalized medicine protocols, or, more generally, to identify characteristics of individuals who are likely to respond to an intervention. Also known as incremental modeling, true lift modeling, or net modeling, uplift modeling differs from traditional modeling techniques in that it finds the interactions between a treatment and other variables. It directs focus to individuals who are likely to react positively to an action or treatment.

Figure 6.1 Example of Uplift for a Hair Product Marketing Campaign





Term	Number of Splits	Contribution		Portion
Gender	2	42.4901339		0.4428
Hair Color	1	39.236122		0.4089
Age	1	14.2342338		0.1483
U.S. Region	0	0		0.0000
Residence	0	0		0.0000

Contents

Uplift Platform Overview

Use the Uplift platform to model the incremental impact of an action, or *treatment*, on individuals. An uplift model helps identify groups of individuals who are most likely to respond to the action. Identification of these groups leads to efficient and targeted decisions that optimize resource allocation and impact on the individual. (See Radcliffe and Surry, 2011.)

The Uplift platform fits partition models. While traditional partition models find splits to optimize a prediction, uplift models find splits to maximize a treatment difference.

The uplift partition model accounts for the fact that some individuals receive the treatment, while others do not. It does this by fitting a linear model to each possible (binary) split. A continuous response is modeled as a linear function of the split, the treatment, and the interaction of the split and treatment. A categorical response is expressed as a logistic function of the split, the treatment, and the interaction of the split and treatment. In both cases, the interaction term measures the difference in uplift between the groups of individuals in the two splits.

The criterion used by the Uplift platform in defining splits is the significance of the test for interaction over all possible splits. However, predictor selection based solely on *p*-values introduces bias favoring predictors with many levels. For this reason, JMP adjusts *p*-values to account for the number of levels. (See the paper “Monte Carlo Calibration of Distributions of Partition Statistics” on the JMP website.) The splits in the Uplift platform are determined by maximizing the adjusted *p*-values for *t* tests of the interaction effects. The logworth for each adjusted *p*-value, namely -$\log_{10}$(adj *p*-value), is reported.

Example of the Uplift Platform

The Hair Care Product.jmp sample data table results from a marketing campaign designed to increase purchases of a hair coloring product targeting both genders. For purposes of designing the study and tracking purchases, 126,184 "club card" members of a major beauty supply chain were identified. Approximately half of these members were randomly selected and sent a promotional offer for the product. Purchases of the product over a subsequent three-month period by all club card members were tracked.

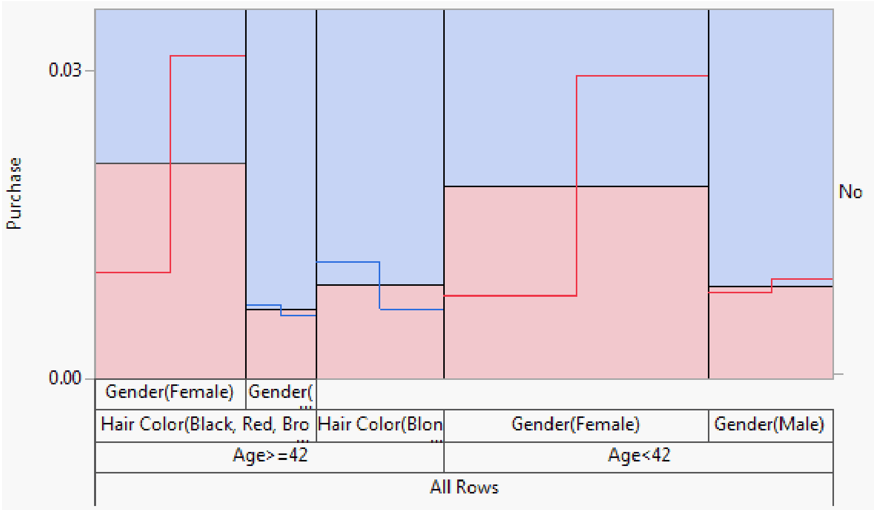
The data table shows a Promotion column, indicating whether the member received promotional material. The column Purchase indicates whether the member purchased the product over the test period. For each member, the following information was assembled: Gender, Age, Hair Color (natural), U.S. Region, and Residence (whether the member is located in an urban area). Also shown is a Validation column consisting of about 33% of the subjects.

For a categorical response, the Uplift platform interprets the first level in its value ordering as the response of interest. This is why the column Purchase has the Value Ordering column property. This property ensures that "Yes" responses are first in the ordering.

1. Open the Hair Care Product.jmp sample data table.
2. Select **Analyze > Consumer Research > Uplift**.
3. From the Select Columns list:
   – Select Promotion and click **Treatment**.
   – Select Purchase and click **Y, Response**.
   – Select Gender, Age, Hair Color, U.S. Region, and Residence, and click **X, Factor**.
   – Select Validation and click **Validation**.
4. Click **OK**.
5. Below the Graph in the report that appears, click **Go**.

   Based on the validation set, the optimal Number of Splits is determined to be five. The Graph is shown in Figure 6.2. Note that the vertical scale has been modified in order to show the detail.

Figure 6.2 Graph after Five Splits

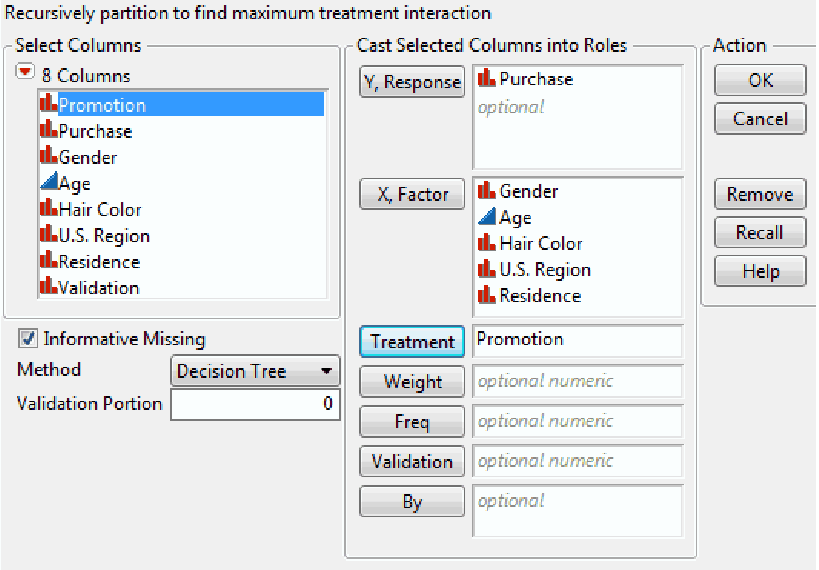


The graph indicates that uplift in purchases occurs for younger people (*Age* < 42), both females and males, and for older females (*Age* ≥ 42) with black, red, or brown hair. For blond-haired subjects and for non-blond males in the *Age* ≥ 42 group, the promotion has a negative effect.

Launch the Uplift Platform

To launch the Uplift platform, select **Analyze > Consumer Research > Uplift**. Figure 6.3 shows a launch window for the Hair Care Product.jmp sample data table. The columns that you enter for Y, Response, and X, Factor can be continuous or categorical. In typical usage, the Treatment column is categorical, and often has only two levels. If your Treatment column contains more than two levels, the first level is treated as Treatment1 and the remaining levels are combined in Treatment2.

Figure 6.3 Launch Window for Uplift



Recursively partition to find maximum treatment interaction

Select Columns

8 Columns

- Promotion
- Purchase
- Gender
- Age
- Hair Color
- U.S. Region
- Residence
- Validation

☑ Informative Missing

Method: Decision Tree

Validation Portion: 0

Cast Selected Columns into Roles

- Y, Response: Purchase
- X, Factor: Gender, Age, Hair Color, U.S. Region, Residence
- Treatment: Promotion
- Weight: optional numeric
- Freq: optional numeric
- Validation: optional numeric
- By: optional

You can specify your own Validation column, or designate a random portion of your data to be selected as a Validation Portion. Note that the only Method currently supported by Uplift is Decision Tree.

The Uplift Model Report

The report opens by showing the Graph and the initial node of the Tree, as well as controls for splitting.

Uplift Model Graph

The graph represents the response on the vertical axis. The horizontal axis corresponds to observations, arranged by nodes. For each node, a black horizontal line shows the mean response. Within each split, there is a subsplit for treatment shown by a red or blue line. These lines indicate the mean responses for each of the two treatment groups within the split. The value ordering of the treatment column determines the placement order of these lines. As nodes are split, the graph updates to show the splits beneath the horizontal axis. Vertical lines divide the splits.

Beneath the graph are the control buttons: **Split**, **Prune**, and **Go**. The Go button only appears if there is a validation set. Also shown is the name of the Treatment column and its two levels, called Treatment1 and Treatment2. If more than two levels are specified for the Treatment column, all but the first level are treated as a single level and combined into Treatment2.

To the right of the Treatment column information is a report showing summary values relating to prediction. (Keep in mind that prediction is not the objective in uplift modeling.) The report updates as splitting occurs. If a validation set is used, values are shown for both the training and the validation sets.

RSquare The RSquare for the regression model associated with the tree. Note that the regression model includes interactions with the treatment column.

N The number of observations.

Number of Splits The number of times splitting has occurred.

AICc The Corrected Akaike Information Criterion (AICc), computed using the associated regression model. AICc is only given for continuous responses.

Uplift Decision Tree

The decision tree shows the splits used to model uplift. See Figure 6.4 for an example using the Hair Care Product.jmp sample data table. Each node contains the following information:

Treatment The name of the treatment column is shown, with its two levels.

Rate Only appears for two-level categorical responses. For each treatment level, the proportion of subjects in this node who responded.

Mean Only appears for continuous responses. For each treatment level, the mean response for subjects in this node.

Count The number of subjects in this node in the specified treatment level.

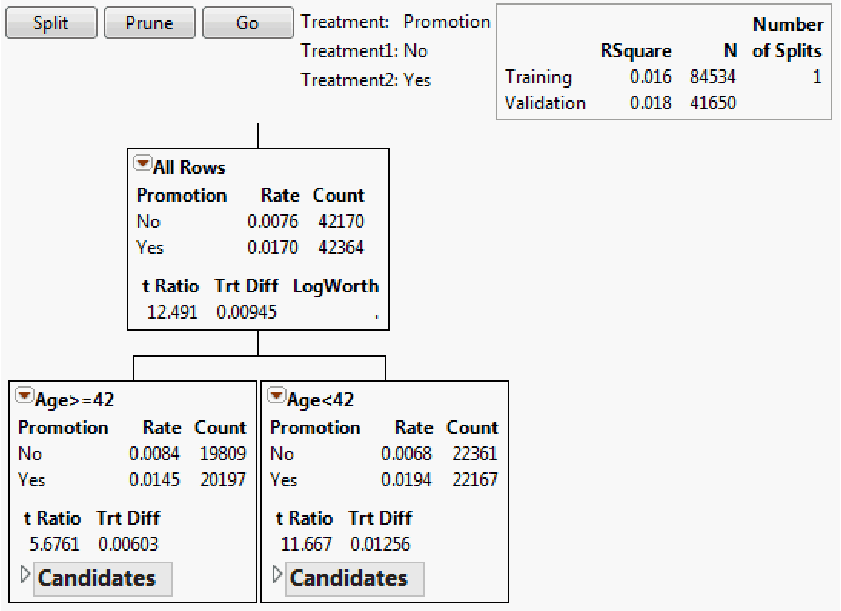
t Ratio The *t* ratio for the test for a difference in response across the levels of Treatment for subjects in this node. If the response is categorical, it is treated as continuous (values 0 and 1) for this test.

Trt Diff The difference in response means across the levels of Treatment. This is the uplift, assuming that:

- The first level in the treatment column's value ordering represents the treatment.
- The response is defined so that larger values reflect greater impact.

LogWorth The value of the logworth for the subsequent split based on the given node.

Figure 6.4 Nodes for First Split



Treatment: Promotion
Treatment1: No
Treatment2: Yes

	RSquare	N	Number of Splits
Training	0.016	84534	1
Validation	0.018	41650	

All Rows

Promotion	Rate	Count
No	0.0076	42170
Yes	0.0170	42364

t Ratio	Trt Diff	LogWorth
12.491	0.00945	.

Age>=42

Promotion	Rate	Count
No	0.0084	19809
Yes	0.0145	20197

t Ratio	Trt Diff
5.6761	0.00603

Age<42

Promotion	Rate	Count
No	0.0068	22361
Yes	0.0194	22167

t Ratio	Trt Diff
11.667	0.01256

Candidates Report

Each node also contains a Candidates report. This report gives:

Term The model term.

LogWorth The maximum logworth over all possible splits for the given term. The logworth corresponding to a split is -$\log_{10}$ of the adjusted *p*-value.

F Ratio When the response is continuous, this is the F Ratio associated with the interaction term in a linear regression model. The regression model specifies the response as a linear function of the treatment, the binary split, and their interaction. When the response is categorical, this is the ChiSquare value for the interaction term in a nominal logistic model.

Gamma When the response is continuous, this is the coefficient of the interaction term in the linear regression model used in computing the *F* ratio. When the response is categorical, this is an estimate of the interaction constructed from Firth-adjusted log-odds ratios.

Cut Point If the term is continuous, this is the point that defines the split. If the term is categorical, this describes the first (left) node.

Uplift Report Options

With the exception of the options described below, all of the red triangle options for the Uplift report are described in the documentation for the Partition platform. For details about these options, see the Partition Models chapter in the *Specialized Models* book.

Minimum Size Split

This option presents a dialog box where you enter a number or a fractional portion of the total sample size to define the minimum size split allowed. To specify a number, enter a value greater than or equal to 1. To specify a fraction of the sample size, enter a value less than 1. The default value for the Uplift platform is set to the minimum of 25 or the floor of the number of rows divided by 2,000.

Column Uplift Contributions

This table and plot address a column's contribution to the uplift tree structure. A column's contribution is computed as the sum of the F Ratio values associated with its splits. Recall that these values measure the significance of the treatment-by-split interaction term in the linear regression model.

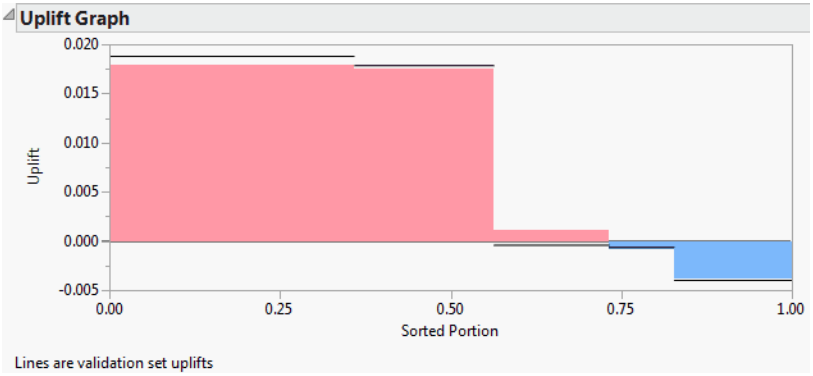
Uplift Graph

Consider the observations in the training set. Define uplift for an observation as the difference between the predicted probabilities or means across the levels of Treatment for the observation's terminal node. These uplift values are sorted in descending order. On its vertical axis, the Uplift Graph shows the uplift values. On its horizontal axis, the graph shows the proportion of observations with each uplift value.

See Figure 6.5 for an example of an Uplift Graph for the Hair Care Product.jmp sample data table after three splits and an additional specific split by Gender for the Hair Color (Black, Red Brown) group. Note that, for two groups of subjects (non-blond males in the *Age* ≥ 42 group, and blond-haired subjects in this group), the promotion has a negative effect.

The horizontal lines shown on the Uplift Graph delineate the graph for the validation set. Specifically, the decision tree is evaluated for the validation set and the Uplift Graph is constructed from the estimated uplifts.

Figure 6.5 Uplift Graph



Save Columns

Save Difference Saves the estimated difference in mean responses across levels of Treatment for the observation's node. This is the estimated uplift.

Save Difference Formula Saves the formula for the Difference, or uplift.

Chapter 7

Item Analysis

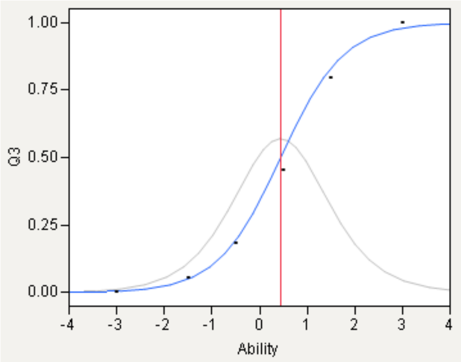
Analyze Test Results by Item and Subject

Item Response Theory (IRT) is a method of scoring tests. Although classical test theory methods have been widely used for a century, IRT provides a better and more scientifically based scoring procedure.

Its advantages include:

- Scoring tests at the item level, giving insight into the contributions of each item on the total test score.
- Producing scores of both the test takers and the test items on the same scale.
- Fitting nonlinear logistic curves, more representative of actual test performance than classical linear statistics.

Figure 7.1 Item Analysis Example



Contents

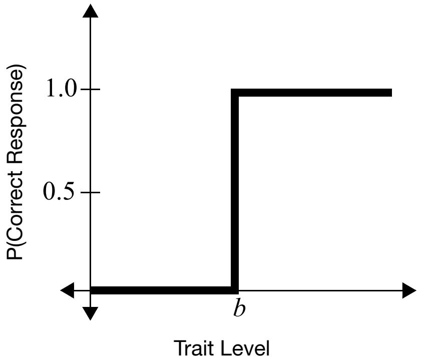
Item Analysis Platform Overview

Psychological measurement is the process of assigning quantitative values as representations of characteristics of individuals or objects, so-called *psychological constructs*. *Measurement theories* consist of the rules by which those quantitative values are assigned. Item Response Theory (IRT) is a measurement theory.

IRT uses a mathematical function to relate an individual's probability of correctly responding to an item to a trait of that individual. Frequently, this trait is not directly measurable and is therefore called a *latent trait*.

To see how IRT relates traits to probabilities, first examine a test question that follows the Guttman "perfect scale" as shown in Figure 7.2. The horizontal axis represents the amount of the theoretical trait that the examinee has. The vertical axis represents the probability that the examinee will get the item correct. (A missing value for a test question is treated as an incorrect response.) The curve in Figure 7.2 is called an *item characteristic curve* (ICC).

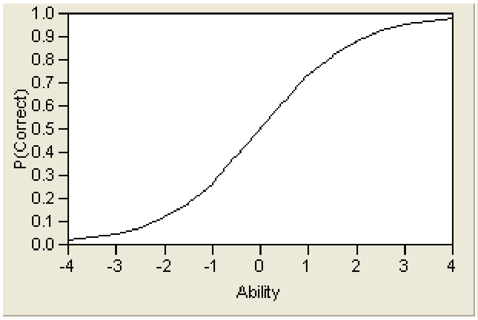
Figure 7.2 Item Characteristic Curve of a Perfect Scale Item



This figure shows that a person who has ability less than the value *b* has a 0% chance of getting the item correct. A person with trait level higher than *b* has a 100% chance of getting the item correct.

Of course, this is an unrealistic item, but it is illustrative in showing how a trait and a question probability relate to each other. More typical is a curve that allows probabilities that vary from zero to one. A typical curve found empirically is the S-shaped logistic function with a lower asymptote at zero and upper asymptote at one. It is markedly nonlinear. An example curve is shown in Figure 7.3.

Figure 7.3 Example Item Response Curve

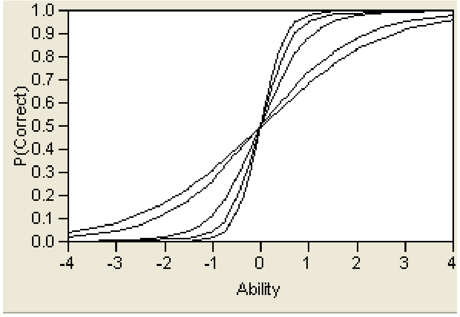


The logistic model is the best choice to model this curve, because it has desirable asymptotic properties, yet is easier to deal with computationally than other proposed models (such as the cumulative normal density function). The model itself is

$$P(\theta) = c + \frac{1-c}{1 + e^{-(a)(\theta - b)}}$$

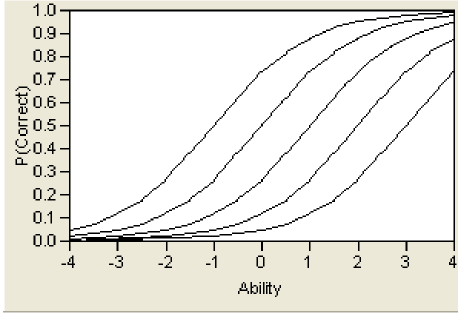
In this model, referred to as a Three-Parameter Logistic (*3PL*) model, the variable *a* represents the steepness of the curve at its inflection point. Curves with varying values of *a* are shown in Figure 7.4. This parameter can be interpreted as a measure of the discrimination of an item—that is, how much more difficult the item is for people with high levels of the trait than for those with low levels of the trait. Very large values of *a* make the model practically the step function shown in Figure 7.2. It is generally assumed that an examinee will have a higher probability of getting an item correct as their level of the trait increases. Therefore, *a* is assumed to be positive and the ICC is monotonically increasing. Some use this positive-increasing property of the curve as a test of the appropriateness of the item. Items whose curves do not have this shape should be considered as candidates to be dropped from the test.

Figure 7.4 Logistic Model for Several Values of *a*



Changing the value of *b* merely shifts the curve from left to right, as shown in Figure 7.5. It corresponds to the value of θ at the point where $P(\theta)=0.5$. The parameter *b* can therefore be interpreted as item difficulty where (graphically), the more difficult items have their inflection points farther to the right along their *x*-coordinate.

Figure 7.5 Logistic Curve for Several Values of *b*

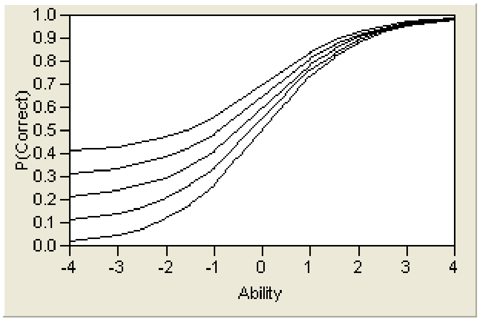


Notice that

$$\lim_{\theta \to -\infty} P(\theta) = c$$

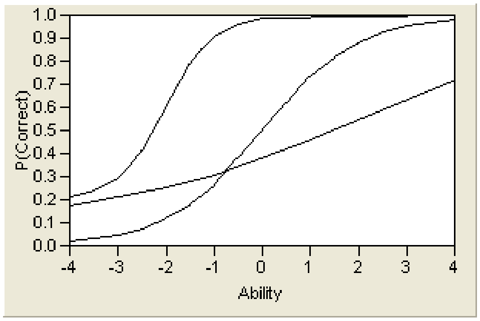
and therefore *c* represents the lower asymptote, which can be nonzero. ICCs for several values of *c* are shown graphically in Figure 7.6. The *c* parameter is theoretically pleasing, because a person with no ability of the trait might have a nonzero chance of getting an item right. Therefore, *c* is sometimes called the *pseudo-guessing parameter*.

Figure 7.6 Logistic Model for Several Values of *c*



By varying these three parameters, a wide variety of probability curves are available for modeling. A sample of three different ICCs is shown in Figure 7.7. Note that the lower asymptote varies, but the upper asymptote does not. This is because of the assumption that there might be a lower guessing parameter, but as the trait level increases, there is always a theoretical chance of 100% probability of correctly answering the item.

Figure 7.7 Three Item Characteristic Curves



Note, however, that the 3PL model might by unnecessarily complex for many situations. If, for example, the *c* parameter is restricted to be zero (in practice, a reasonable restriction), there are fewer parameters to predict. This model, where only *a* and *b* parameters are estimated, is called the 2PL model.

Another advantage of the 2PL model (aside from its greater stability than the 3PL) is that *b* can be interpreted as the point where an examinee has a 50% chance of getting an item correct. This interpretation is not true for 3PL models.

A further restriction can be imposed on the general model when a researcher can assume that test items have equal discriminating power. In these cases, the parameter *a* is set equal to 1, leaving a single parameter to be estimated, the *b* parameter. This *1PL* model is frequently called the *Rasch model*, named after Danish mathematician Georg Rasch, the developer of the model. The Rasch model is quite elegant, and is the least expensive to use computationally.

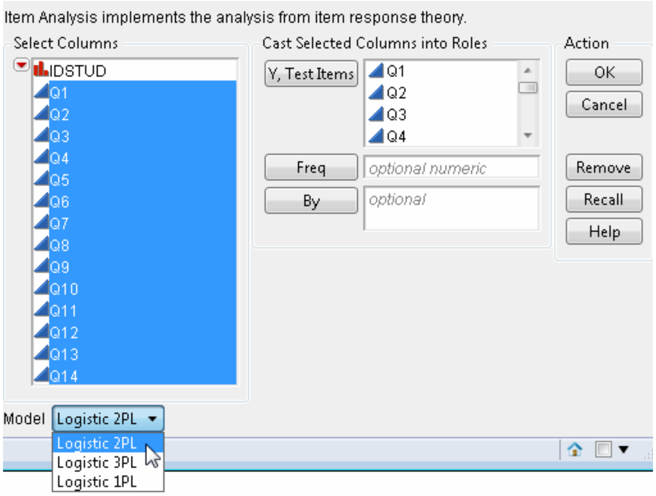
Caution: You must have a lot of data to produce stable parameter estimates using a 3PL model. 2PL models are frequently sufficient for tests that intuitively deserve a guessing parameter. Therefore, the 2PL model is the default and recommended model.

Launch the Item Analysis Platform

For example, open the sample data file MathScienceTest.jmp. These data are a subset of the data from the Third International Mathematics and Science Study (TIMMS) conducted in 1996.

To launch the Item Analysis platform, select **Analyze > Consumer Research > Item Analysis**. This shows the dialog in Figure 7.8.

Figure 7.8 Item Analysis Launch Dialog



Item Analysis implements the analysis from item response theory.

Model Logistic 2PL

Y, Test Items Are the questions from the test instrument.

Freq Specifies a variable used to specify the number of times each response pattern appears.

By Performs a separate analysis for each level of the specified variable.

Specify the desired model (1PL, 2PL, or 3PL) by selecting it from the **Model** drop-down menu.

For this example, specify all fourteen continuous questions (Q1, Q2,..., Q14) as **Y, Test Items** and click **OK**. This accepts the default 2PL model.

Special Note on 3PL Models

If you select the 3PL model, a dialog pops up asking for a penalty for the *c* parameters (thresholds). This is not asking for the threshold itself. The penalty that it requests is similar to the type of penalty parameter that you would see in ridge regression, or in neural networks.

The penalty is on the sample variance of the estimated thresholds, so that large values of the penalty force the estimated thresholds' values to be closer together. This has the effect of speeding up the computations, and reducing the variability of the threshold (at the expense of some bias).

In cases where the items are questions on a multiple choice test where there are the same number of possible responses for each question, there is often reason to believe (*a priori*) that the threshold parameters would be similar across items. For example, if you are analyzing the results of a 20-question multiple choice test where each question had four possible responses, it is reasonable to believe that the guessing, or threshold, parameters would all be near 0.25. So, in some cases, applying a penalty like this has some "physical intuition" to support it, in addition to its computational advantages.

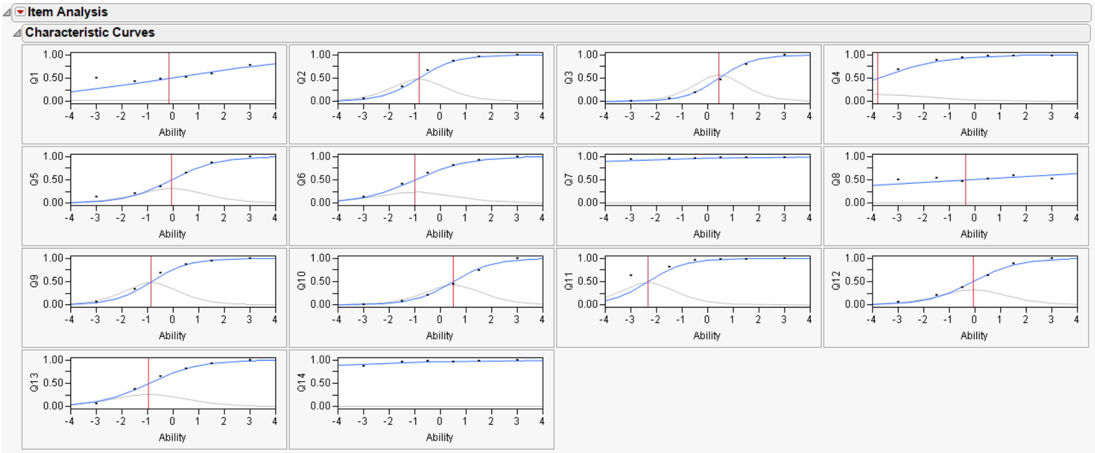
The Item Analysis Report

The following plots appear in Item Analysis reports.

Characteristic Curves

Item characteristic curves for each question appear in the top section of the output. Initially, all curves are shown stacked in a single column. They can be rearranged using the **Number of Plots Across** command, found in the drop down menu of the report title bar. For Figure 7.9, four plots across are displayed.

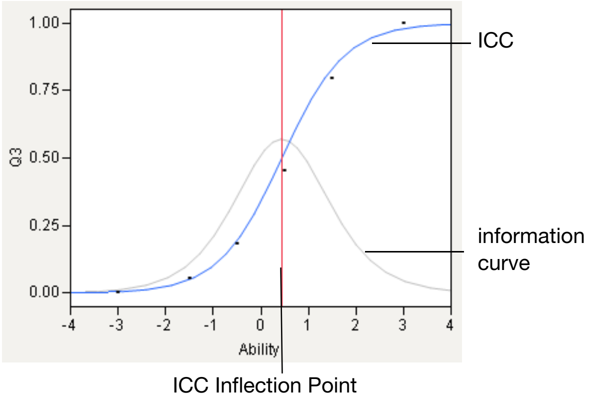
Figure 7.9 Component Curves



A vertical red line is drawn at the inflection point of each curve. In addition, dots are drawn at the actual proportion correct for each ability level, providing a graphical method of judging goodness-of-fit.

Gray information curves show the amount of information each question contributes to the overall information of the test. The information curve is the slope of the ICC curve, which is maximized at the inflection point.

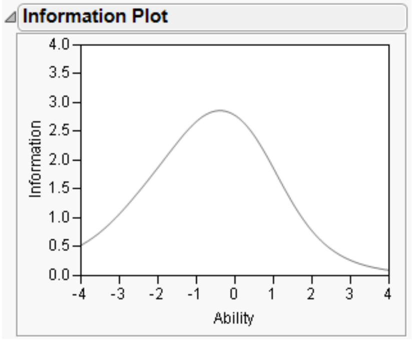
Figure 7.10 Elements of the ICC Display



Information Curves

Questions provide varying levels of information for different ability levels. The gray information curves for each item show the amount of information that each question contributes to the total information of the test. The total information of the test for the entire range of abilities is shown in the Information Plot section of the report (Figure 7.11).

Figure 7.11 Information Plot

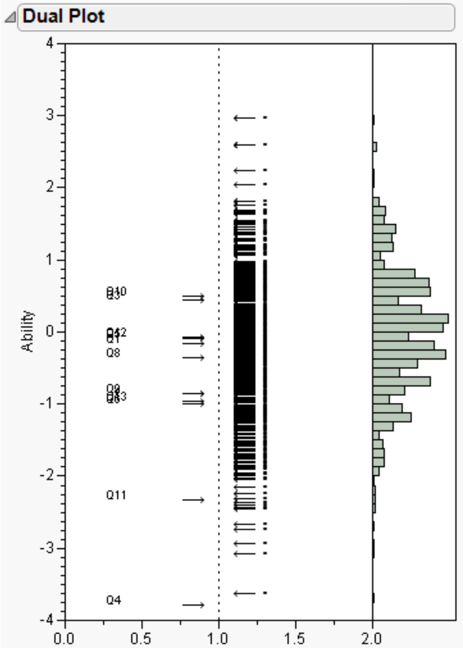


Dual Plots

The information gained from item difficulty parameters in IRT models can be used to construct an increasing scale of questions, from easiest to hardest, on the same scale as the examinees. This structure gives information about which items are associated with low levels of the trait, and which are associated with high levels of the trait.

JMP shows this correspondence with a *dual plot*. The dual plot for this example is shown in Figure 7.12.

Figure 7.12 Dual Plot

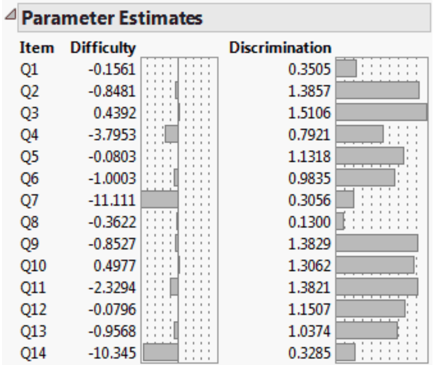


Questions are plotted to the left of the vertical dotted line, examinees on the right. In addition, a histogram of ability levels is appended to the right side of the plot.

This example shows a wide range of abilities. Q10 is rated as difficult, with an examinee needing to be around half a standard deviation above the mean in order to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels, with Q11 and Q4 appearing as easier. There are some questions that are off the displayed scale (Q7 and Q14).

The estimated parameter estimates appear below the Dual Plot, as shown in Figure 7.13.

Figure 7.13 Parameter Estimates



Parameter Estimates

Item	Difficulty	Discrimination
Q1	-0.1561	0.3505
Q2	-0.8481	1.3857
Q3	0.4392	1.5106
Q4	-3.7953	0.7921
Q5	-0.0803	1.1318
Q6	-1.0003	0.9835
Q7	-11.111	0.3056
Q8	-0.3622	0.1300
Q9	-0.8527	1.3829
Q10	0.4977	1.3062
Q11	-2.3294	1.3821
Q12	-0.0796	1.1507
Q13	-0.9568	1.0374
Q14	-10.345	0.3285

Item Identifies the test item.

Difficulty Is the *b* parameter from the model. A histogram of the difficulty parameters is shown beside the difficulty estimates.

Discrimination Is the *a* parameter from the model, shown only for 2PL and 3PL models. A histogram of the discrimination parameters is shown beside the discrimination estimates.

Threshold Is the *c* parameter from the model, shown only for 3PL models.

Item Analysis Platform Options

The following three commands are available from the drop-down menu on the title bar of the report.

Number of Plots Across Brings up a dialog to specify how many plots should be grouped together on a single line. Initially, plots are stacked one-across. Figure 7.9 on page 142 shows four plots across.

Save Ability Formula Creates a new column in the data table containing a formula for calculating ability levels. Because the ability levels are stored as a formula, you can add rows to the data table and have them scored using the stored ability estimates. In addition, you can run several models and store several estimates of ability in the same data table.

The ability is computed using the `IRT Ability` function. The function has the following form

```
IRT Ability (Q1, Q2,...,Qn, [a1, a2,..., an, b1, b2,..., bn, c1, c2, ...,
cn]);
```

where `Q1`, `Q2`,...,`Qn` are columns from the data table containing items, `a1`, `a2`,..., `an` are the corresponding discrimination parameters, `b1`, `b2`,..., `bn` are the corresponding difficulty parameters for the items, and `c1`, `c2`, ..., `cn` are the corresponding threshold parameters. Note that the parameters are entered as a matrix, enclosed in square brackets.

Script Contains options that are available to all platforms. See *Using JMP*.

Technical Details

Note that $P(\theta)$ does not necessarily represent the probability of a positive response from a *particular* individual. It is certainly feasible that an examinee might definitely select an incorrect answer, or that an examinee might know an answer for sure, based on the prior experiences and knowledge of the examinee, apart from the trait level. It is more correct to think of $P(\theta)$ as the probability of response for a set of individuals with ability level θ. Said another way, if a large group of individuals with equal trait levels answered the item, $P(\theta)$ predicts the proportion that would answer the item correctly. This implies that IRT models are item-invariant; theoretically, they would have the same parameters regardless of the group tested.

An assumption of these IRT models is that the underlying trait is unidimensional. That is to say, there is a single underlying trait that the questions measure that can be theoretically measured on a continuum. This continuum is the horizontal axis in the plots of the curves. If there are several traits being measured, each of which have complex interactions with each other, then these unidimensional models are not appropriate.

Appendix A

References

Akaike, H. (1974), “Factor Analysis and AIC,” *Pschychometrika*, 52, 317–332.

Akaike, H. (1987), “A new Look at the Statistical Identification Model,” *IEEE Transactions on Automatic Control*, 19, 716–723.

Dwass, M. (1955), “A Note on Simultaneous Confidence Intervals,” *Annals of Mathematical Statistics* 26: 146–147.

Farebrother, R.W. (1981), “Mechanical Representations of the L1 and L2 Estimation Problems,” *Statistical Data Analysis*, 2nd Edition, Amsterdam, North Holland: edited by Y. Dodge.

Fieller, E.C. (1954), “Some Problems in Interval Estimation,” *Journal of the Royal Statistical Society, Series B*, 16, 175-185.

Firth, D. (1993), “Bias Reduction of Maximum Likelihood Estimates,” Biometrika 80:1, 27–38.

Goodnight, J.H. (1978), “Tests of Hypotheses in Fixed Effects Linear Models,” *SAS Technical Report R–101*, Cary: SAS Institute Inc, also in Communications in Statistics (1980), A9 167–180.

Goodnight, J.H. and W.R. Harvey (1978), “Least Square Means in the Fixed Effect General Linear Model,” *SAS Technical Report R–103*, Cary NC: SAS Institute Inc.

Heinze, G. and Schemper, M. (2002), “A Solution to the Problem of Separation in Logistic Regression,” *Statistics in Medicine* 21:16, 2409–2419.

Hocking, R.R. (1985), *The Analysis of Linear Models*, Monterey: Brooks–Cole.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, Second Edition, New York: John Wiley and Sons.

Kaiser, H.F. (1958), “The varimax criterion for analytic rotation in factor analysis” *Psychometrika*, 23, 187–200.

McFadden, D. (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” in P. Zarembka, ed., *Frontiers in Econometrics*, pp. 105–142.

Radcliffe, N. J., and Surry, P. D. (2011), “Real-World Uplift Modelling with Significance-Based Uplift Trees,” Stochastic Solutions White Paper, Portrait Technical Report TR-2011-1.

Reichheld, F. F. (2003) “The One Number You Need to Grow,” *Harvard Business Review,* Vol. 81 No. 12, 46-54.

Wright, S.P. and R.G. O’Brien (1988), “Power Analysis in an Enhanced GLM Procedure: What it Might Look Like,” *SUGI 1988, Proceedings of the Thirteenth Annual Conference*, 1097–1102, Cary NC: SAS Institute Inc.

Index

Consumer Research