Human Dimensions of Wildlife, 13:361–379, 2008

ISSN: 1087-1209 print / 1533-158X online
DOI: 10.1080/10871200802339631

Routledge
Taylor & Francis Group

Structuring Survey Data to Facilitate Analysis and Interpretation

JAY BEAMAN[1] AND JERRY J. VASKE[2]

[1]Auctor Consulting Associates, Ltd., Cheyenne, Wyoming, USA
[2]Human Dimensions of Natural Resources, Colorado State University, Fort Collins, Colorado, USA

Human dimensions survey data are commonly stored in flat files where the rows correspond to individuals and the columns are variables. As the number of variables increases (e.g., 1,000+) or when compressed variables are used, the complexity of understanding the data increases substantially. This article illustrates how data can be restructured into relational entities to facilitate analyses. Using Sportsperson data from the 2006 National Survey of Fishing, Hunting and Wildlife-Associated Recreation (FHWAR), approximately 1,750 flat file variables were reduced to fewer than 60 relational variables. In contrast to the compressed flat file variables that cannot be directly used in SPSS or SAS, variables in the relational entities can be analyzed. Three examples are given to illustrate using the relational entities. General implications of using relational data structures in analysis and data collection are introduced.

Keywords data structure, relational entities, FHWAR

Introduction

Human dimensions researchers commonly store survey data in flat files where the rows correspond to individual respondents and the columns are variables for a respondent's answers to survey questions. The size of such files depends on the number of respondents and the number of variables. Today's computers allow analyses with thousands of respondents and thousands of variables. Many flat file data sets with a small number of variables (e.g., <100) are relatively easy to understand and analyze. As the number of variables increases (e.g., 1,000+), the complexity of analyzing the relationships between variables can be substantial. For example, the 2006 National Survey of Fishing, Hunting, and Wildlife-Associated Recreation (FHWAR) includes a Sportsperson flat data file that has 3,765 variables. Many "compressed" variables carry information about the values of three variables (e.g., "days" of participation is combined with an "activity" such as big game hunting in a particular "state") in a single variable. These compressed variables cannot be directly analyzed by statistical software commonly used by human dimensions researchers.

This article pursues creating and using a relational database structure and rationalizes its use. The article is divided into three major sections. The first section briefly introduces relational database concepts using journal article data. The second introduces a relational database structure using FHWAR as an example. The restructured database reduces approximately 1,750 flat file variables in the 2006 Sportsperson data to fewer than 60 relational

Address correspondence to Jay Beaman, Auctor Consulting Associates, Ltd., 1015 Hoy, Cheyenne, WY 82009, USA. E-mail: jaybman@aceweb.com

variables. Section three contains SAS and SPSS (i.e., Statistical Package for the Social Sciences) examples to facilitate understanding the benefits of having FHWAR data in a relational structure. We conclude with some practical implications of restructuring FHWAR data and human dimensions data in general. Research avenues flowing from restructuring FHWAR data are suggested.

Data Structures

Flat File Data Structures

Two types of data structures are considered in this article: (a) flat files and (b) relational databases. A flat file structure is illustrated with variables for storing information about articles published in *Human Dimensions of Wildlife* (*HDW*) (Table 1). Each row of Table 1 represents a journal article published in *HDW*. Each column is a variable characterizing a given article. For example, all *HDW* articles have one or more authors, a title, specifics about date of publication (e.g., year, volume, issue number, pages), as well as other potential descriptor variables (e.g., keywords).

The flat file data structure for *HDW* articles resulted in multiple columns with similar information and numerous empty cells. For example, because the article by Diefenbach et al. (2005) had seven co-authors (last row, Table 1), seven columns (variables) were devoted to author information. Because 19 of the 26 articles had only one or two authors, more than 67% of the author fields were blank. If the table allowed for separate columns for each author's first name and initials, more variables would be necessary and more "empty" cells would occur. A similar situation arises for variables such as keywords. Some articles (e.g., book reviews) do not contain any keywords; others could have six or seven keywords.

Relational Databases

Problems such as numerous empty cells and not enough variables for some information (e.g., authors) can be avoided by structuring data as a relational database (Avedon, 1992). A relational database is formally defined as a set of tables containing data for predefined categories (Codd, 1970). Information in a relational database is stored in separate files (i.e., tables) that are linked to one another. In a relational database terminology, a table is referred to as an entity (**E**). The rows (i.e., tuples) in a table represent information about an object (e.g., journal article or respondent). The columns (i.e., attributes) represent variables. Two types of relations (**R**) can occur (Chen, 1976). Some relations are a set of tuples; a table with attributes. These relations store data. Other relations are algebraic (e.g., PERSON_ID in table A equals PERSON_ID in table B). These relations use data stored in entities.

Figure 1 shows a structure for storing journal article information relationally using four entities linked by three relations. In the author entity, *AuthorID* uniquely identifies an author. Attributes could be last name, first name, second initial, and other potentially useful information (e.g., affiliation, e-mail address, phone numbers). The article entity contains information about the articles. Each row in this entity represents a particular article. The unique *ArticleID* appears in a row with the article title and other article specific information (e.g., volume, issue, pages). Author data is linked to article information using relation **R1**.

The relation **R1** is a table in which multiple authors (i.e., *AuthorID*) are associated with a given article (*ArticleID*). This is referred to as a "many to one" relation. For example, article number 2059 (Table 1) would occur in three rows in **R1**. Each row is for one of the three authors of article 2059. If the *AuthorID*s were 314, 59, and 233 for Chase, Siemer, and

Table 1
Example of a "flat file" data structure for journal articles

ID	Author1	Author2	Author3	Author4	Author5	Author6	Author7	Title	Year	Vol.	No.	Pages
1060	Gill							The wildlife professional subculture: The case of the crazy aunt	1996	1	1	60–69
1070	Decker	Krueger	Baer	Knuth	Richmond			From clients to stakeholders: A philosophical shift for fish and wildlife management	1996	1	1	70–82
1085	Heberlein	Thomson						Changes in U.S. hunting participation, 1980–90	1996	1	1	85–86
1032	Tynon							Quality hunting experiences	1997	2	1	32–46
1068	LaPage							The wolf as social indicator	1997	2	1	68–70
2037	Whittaker							Capacity norms on bear viewing platforms	1997	2	2	37–49
4053	Green							Defensiveness about hunting?	1998	3	4	53–54
2027	Eliason							The illegal taking of wildlife	1999	4	2	27–39
2059	Chase	Siemer	Decker					Suburban deer management	1999	4	2	59–60
3074	Hamilton							The case for abundant species management	1999	4	3	74–85
3086	Faast	Simon-Brown						A social ethic for fish and wildlife management	1999	4	3	86–92
4084	Lewis	Leitch						Value-added wildlife management	1999	4	4	84–85
4062	Green	Stowe						Quality deer management: Ethical and social issues	2000	5	4	62–71

(*Continued*)

Table 1
(*Continued*)

ID	Author1	Author2	Author3	Author4	Author5	Author6	Author7	Title	Year	Vol.	No.	Pages
4072	Gill							Managing wildlife ethics issues ethically	2000	5	4	72–82
2081	Hayslette	Armstrong	Mirarchi					Mourning dove hunting in Alabama	2001	6	2	81–95
3173	Finn	Loomis						The importance of catch motives to anglers	2001	6	3	173–187
3189	Miller	Graefe						Effect of harvest success on hunter attitudes	2001	6	3	189–203
3197	Hadlock	Beckwith						Providing incentives for endangered species recovery	2002	7	3	197–213
2081	Riley	Siemer	Decker	Carpenter	Organ	Berchielli		Adaptive impact management	2003	8	2	81–95
2097	Van Deelen	Etter						Effort and the functional response of deer hunters	2003	8	2	97–108
3165	McFarlane	Watson	Boxall					Women hunters in Alberta, Canada	2003	8	3	165–180
3199	Scott	Thigpen						Understanding the birder as tourist	2003	8	3	199–218
3219	Manfredo	Vaske	Teel					The potential for conflict index	2003	8	3	219–228
2087	Ditton	Sutton						Substitutability in recreational fishing	2004	9	2	87–102
2095	Aslin	Bennett						Two tool boxes for wildlife management	2005	10	2	95–107
3201	Diefenbach	Finley	Luloff	Stedman	Swope	Zinn	San Julian	Bear and deer hunter density and distribution	2005	10	3	201–212

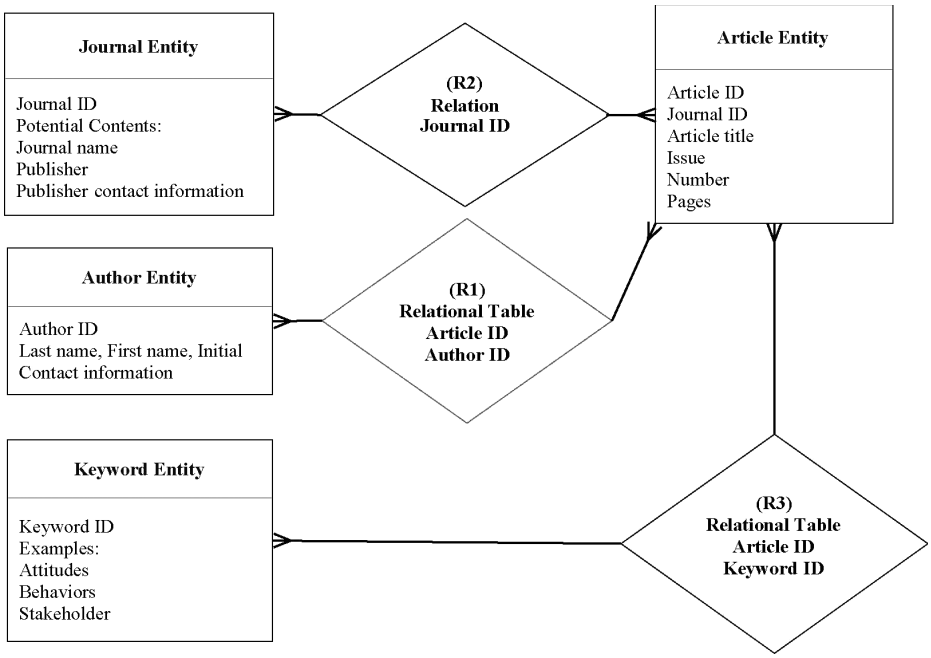

Figure 1. A relational data structure for journal articles.

Decker, respectively, the **R1** rows would be (2059, 314), (2059, 59) and (2059, 233). The author entity and **R1** (author–article relation) can have any number of rows so all authors of an article can be identified. If information about an author changes (e.g., affiliation, e-mail address), the author's information is updated by changing one record in the author entity.

R2 relates the article entity and the journal entity containing journal information. Table 1 only included selected articles from *HDW* for illustration purposes; a multi-journal database would have articles from a variety of journals. An article is associated with one journal. Articles from different journals would have different values of *JournalID* (a variable in the journal entity). The link between the article entity and the journal entity is defined by associating journals with articles based on *JournalID*. The journal entity stores detailed information about a journal that may prove useful (e.g., journal name, publisher, publisher contact information). Because a journal appears once in the journal entity, data about a journal is entered in one table; any necessary changes (e.g., the journal is no longer published) are made in one place. The number of details (attributes) about journals in the journal entity does not affect the size of article entity.

The relation of an article to keywords is stored in **R3**. The pair (*ArticleID*, *KeywordID*) associates an article with keywords in the keyword entity. Because there is no practical limit on the number of (*ArticleID*, *KeywordID*) pairs, there are no restrictions on the number of keywords used to describe an article.

Data Storage Affecting Ease of Use: FHWAR Example

The FHWAR Survey

Natural resource agencies recognized the need for comprehensive information on the economic impacts of sport fishing and hunting in the 1950s (Grambsch & Fisher, 1989;

International Association of Fish and Wildlife Agencies [IAFWA], 1953). IAFWA approved the first National Survey of Fishing, Hunting and Wildlife-Associated Recreation at its annual meeting in September 1954 (IAFWA, 1956). The Fish and Wildlife Service contracted with U.S. Census Bureau for the first national survey in 1955. The survey has continued approximately every five years since 1955. FHWAR data such as sportsperson expenditures and species sought in different states are used by state fish and wildlife agencies in: (a) planning and management, (b) budget negotiations with lawmakers and policymakers, and (c) legal actions such as settlements in natural resource damage cases. The FHWAR survey is currently recognized as the most comprehensive data for understanding the economic impacts of fishing and hunting in the United States (McDowell & Mock, 2004). Despite the need for FHWAR data, working with these data is challenging.

Current FHWAR Flat File Data Structure

FHWAR data files are distributed on a CD in three ASCII "text" files: (a) Screening file, (b) Sportsperson (fishing and hunting) file, and (c) Wildlife Watcher file.[1] The ASCII text files are long lines of text for each respondent. The text files can be converted to SAS data sets using programs (e.g., convert3.sas) provided by the Census Bureau. SPSS will convert the SAS files to SPSS. Conversion programs, however, do not create variable labels or value labels for variables. Variable descriptions and value labels are provided on the CD in Microsoft (MS) Word documents (i.e., fh2.doc, fh3.doc, fh4.doc). For the 2006 data, there are 11 pages of text for the Screening file, 146 pages for the Sportsperson file, and 36 pages for the Wildlife Watcher file. Incorporating variable descriptors from the document files into SAS (or SPSS) requires considerable effort.

The 2006 Screening file contains 144,509 records and 56 variables. The Sportsperson file includes 21,942 records with 3,765 variables. In this article, sportspersons are a sample of individuals (16+ years old) selected from the screening sample based on their likelihood of fishing or hunting (see USDI & USDC, 2006, p. 149).[2] The Wildlife Watcher file has 11,285 records and 772 variables. Wildlife viewers were selected from the screening sample to report on non-consumptive wildlife related activities. With more than 4,500 variables (i.e., 56 + 3,765 + 772 = 4,593), finding a variable for analysis can involve searching approximately 200 pages of variable and value documentation. The search task is further complicated by some variable names that have no intuitively obvious meaning (e.g., NCU_STD1, NCUTOD1).[3]

Some queries of the Sportsperson data are relatively straightforward. For example, obtaining the number of males who hunted in 2006 by state of residence requires only a few SPSS or SAS commands (see Example 1—Hypothesis 1—Figure 3). Addressing other questions, however, is more complex. For example, the percent of Colorado males that hunted cannot be acquired using only Sportsperson data (see Example 1—Hypothesis 2—Figure 4). Screening or other data are necessary to determine the number of Colorado males by which to divide the number of hunters to compute a percent.

Further complexity exists because some blocks of variables are "compressed." For example, because the maximum number of states any respondent reported hunting in was eight, there are eight variables for recording hunting in different states (HUNTSTD1 through HUNTSTD8). If the first state mentioned by a hunter was Colorado, a code of "CO" is stored in HUNTSTD1. If another person reported hunting in Wyoming first and Colorado second, the codes would be HUNTSTD1 = WY and HUNTSTD2 = CO, respectively. Similarly, there are eight variables for big game hunting, eight for small game hunting, eight for migratory bird hunting, and eight for hunting other animals. Variables for big game hunting are

BGHNT1 through BGHNT8. For these variables, values are Yes (1) or No (Blank). The 1 through 8 in the variable names BGHNT1 through BGHNT8 refer to the states. The state associated with BGHNT1 is in HUNTSTD1 (BGHNT2 is in HUNTSTD2, etc.). Responses about big game hunting in 50 states are compressed into eight variables.

Compressing variables complicates analysis. For example, to examine big game hunting in a particular state (e.g., Colorado) information in BGHNT1 through BGHNT8 must be "decompressed" to produce variables for use in SAS or SPSS. Decompression is a multistep process. Step 1 involves examining HUNTSTD1 through HUNTSTD8 to determine if there is a value of CO. If CO is not found, the respondent did not hunt in Colorado. If HUNTSTD2=CO, the value of BGHNT2 is determined. If BGHNT2=1, the person hunted big game in Colorado; otherwise no. To obtain each trivariate piece of information (i.e., state, activity, participation status), the researcher must be familiar with using SAS code, SPSS syntax, or another programming language (e.g., Basic or C) to decompress the variables and store the information for analysis. For example, for big game hunting in Colorado, a variable, BGH_CO, could be created with 1=yes and 0=no. For 50 states, 50 "BGH" variables would be created to store state specific information. Fifty small game variables (e.g., SMG_CO, SMG_WY) could be created for participation data for small game hunting. Similarly, groups of 50 variables could be created for storing days of participation and numbers of trips. Decompression is required to access all FHWAR data involving state-specific responses. When groups of compressed variables in the Sportsperson data file are uncompressed to blocks of 50, there are about 20,000 FHWAR variables.

Overall, FHWAR data are challenging to use because there are over 4,500 variables that are described in about 200 pages of documentation. Some of the variable names may have no intuitive meaning or convey trivariate information (e.g., *days* of participation in an *activity* in a *state*) that must be decompressed for analysis with statistical software (e.g., SAS, SPSS).

FHWAR Sportsman Data in a Relational File Structure

Just as some articles have multiple authors, some sportspersons have annual fishing or hunting expenditures of various types (e.g., lodging) in particular states. Figure 2 illustrates the relationship between four entities: *PERSON*, *SPORTSPERSON*, *HUNTING_ACTIVITY*, and *TRIP_EXPENDITURES*.[4] The *PERSON* entity has data about persons in the United States and includes: six control variables (e.g., PERSON_WEIGHT, CENSUS_DIVISION), (b) 10 demographic variables (e.g., AGE, SEX), (c) eight hunting variables (e.g., HUNTED_2005), (d) eight fishing variable (e.g., FISHED_2005), (e) six residential wildlife watching variables (e.g., HOME_OBSERVE_2005), and (f) five non-residential wildlife watching variables (e.g., TRIP_WATCH_2005). All variables in the *PERSON* entity were obtained from the Screening data.[5]

Sportsperson flat file data were used to create three entities: *SPORTSPERSON*, *HUNTING_ACTIVITY*, and *TRIP_EXPENDITURES*.[6] The *SPORTSPERSON* entity contains: (a) six control variables (e.g., PERSON_ID, SPORTSPERSON_WEIGHT), (b) 11 demographic variables (e.g., AGE, SEX), and (c) 15 national summary variables (e.g., HUNTED_2006). Information from the compressed variables in the original flat file structure and other hunting activity information are in the *HUNTING_ACTIVITY* entity. The *HUNTING_ACTIVITY* entity replaces 840 compressed variables with 12 relational variables. The *TRIP_EXPENDITURES* entity reduces 844 compressed variables to 10 variables (e.g., TRIP_EXPEND_CATEGORIES, DOLLARS). *TRIP_EXPENDITURES* contains FHWAR fishing and hunting trip expenditure responses. By changing data structure, fewer than 60 variables replace 1,750 Sportsperson flat file variables.

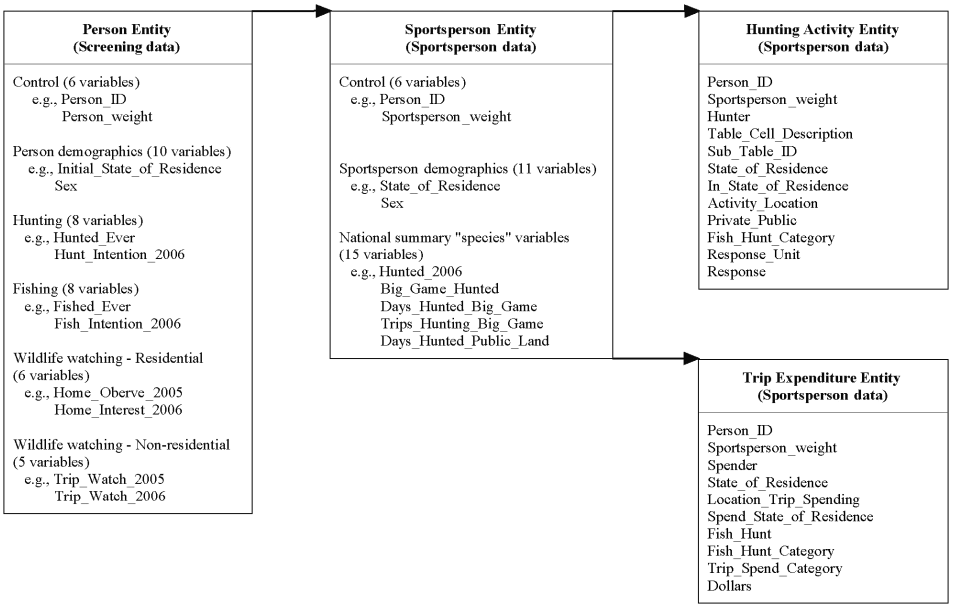

Figure 2. A relational data structure for selected components of FHWAR data.

All four entities, along with some SAS and SPSS syntax files and associated documentation files, are available online.[7] Value and variable labels have been incorporated into the SPSS data files. SAS variable labels and formats for value labels are provided and information is given on using the SAS files with their formats.

Some variables appear in more than one entity. PERSON_ID, for example, provides the "key" for linking the entities. Demographic information is included in both the *PERSON* entity and the *SPORTSPERSON* entity to facilitate analyses. SPORTSPERSON_WEIGHT is included in *SPORTSPERSON*, *HUNTING_ACTIVITY*, and *TRIP_EXPENDITURES* because this weight applies to these entities.

Using the Sportsperson Data in a Relational File Structure

Examples provide a convenient way to understand using the *PERSON* entity and the restructured Sportsperson data. This section includes three examples to illustrate using the relationally structured FHWAR data. The figures associated with these examples include SAS code or SPSS syntax and output. The two analyses associated with the first example can be performed using flat file data in much the same way they are performed with *PERSON*, and *SPORTSPERSON* entities. Analyses presented in the other examples are difficult to perform using flat file Sportsperson data.

Example 1

Analysis Problem. Consider two research hypotheses:

H_1 Wyoming *male sportspersons* are more likely to have hunted in 2006 than Colorado male sportspersons.

H_2 Wyoming *males* are more likely to have hunted in 2006 than Colorado males.

Only the *SPORTSPERSON* entity is needed to examine hypothesis 1 because Sportsperson data can be used to estimate the number of male hunters and male sportspersons per state.[8] Sportsperson data, however, cannot be used to estimate the number of males in a state (hypothesis 2) because male sportspeople are a sub-population of males. The test for examining hypothesis 2 requires joining the *SPORTSPERSON* entity with *PERSON* entity information.

SPSS Syntax and Output—Hypothesis 1. The first line of the SPSS syntax accesses sportsperson data in Sportsperson.sav (Figure 3). The second line weights the data by the variable, SPORTSPERSON_WEIGHT. The "temporary select if" statements constrain the analysis to males in Colorado and Wyoming. The first CROSSTABS procedure relates HUNTED_2006 to STATE_OF_RESIDENCE with the weight turned on to produce percents for hypothesis 1. The second CROSSTABS examines the relationship between the same two variables, but with the weight turned off to get an approximate statistical test (χ^2) and an effect size (ϕ). Using unweighted data for the statistical tests was appropriate because the cell frequencies for the unweighted data have approximately the same distribution as when weighted. In some cases, if weights are "adjusted," SPSS can be used to produce valid statistics (see Vaske, 2008).

The output supports hypothesis 1 (Figure 3); 35% of male Wyoming sportspersons hunted whereas 17% of male sportspersons in Colorado hunted, $\chi^2 = 14.72$, $p < .001$. The effect size, however, suggests only a "minimal" (see Vaske, Gliner, & Morgan, 2002) relationship between the two variables, $\phi = .183$.

SPSS Syntax and Output—Hypothesis 2. Testing hypothesis 2 requires data from both the *PERSON* and the *SPORTSPERSON* entities (Figure 4). The SPSS AGGREGATE command is used to determine: (a) number of male sportspersons in Colorado and Wyoming who hunt (from *SPORTSPERSON*) and (b) number of males in Colorado and Wyoming (from *PERSON*). The SPSS MATCH command is used to join the two files produced by aggregation. A COMPUTE statement calculates the percent of males that are hunters.

The results indicated that 22% of Wyoming males hunted whereas only 7% of Colorado males hunted. These percentages provide evidence in support of hypothesis 2. A formal test that this difference is significant can be derived from the estimates of the standard deviations in the percents. Obtaining standard errors and making a test using weighted data is not difficult but outside the scope of this article (see Appendix D, USDI & USDC, 2006).

Lessons Learned—Hypotheses 1 and 2. First, percents based on the Sportsperson data are not percents based on the population of males 16+ in a state. The percent of the population hunting cannot be determined only using numbers obtained from the Sportsperson data (flat file or relational entity). Second, population estimates (e.g., males in Colorado 16+) could be obtained from either the Census Bureau or the *PERSON* entity. If the latter is used, individuals less than 16 years of age must be eliminated from Person data in estimating population sizes (see SPSS syntax, Figure 4). Third, statistical tests and estimates must be pursued with care when weights are applied.

SPSS syntax—Hypothesis 1

SPSS syntax	Description
GET FILE='C:\Sportsperson.sav'. Select if State_of_Residence = 8 or State_of_Residence = 56). Select if (Sex = 0). WEIGHT BY Sportsperson_Weight.	Opens the Sportsperson Entity Selects Colorado and Wyoming residents Selects males Weights data by Sportsperson Weight
CROSSTABS /TABLES=Hunted_2006 BY State_of_Residence /CELLS=COUNT COLUMN /COUNT TRUNCATE CELL.	Produces a weighted 2-way crosstabulation
WEIGHT OFF.	Turns off weights
CROSSTABS /TABLES=Hunted_2006 BY State_of_Residence /CELLS=COUNT COLUMN /STATISTICS = CHISQ, PHI /COUNT TRUNCATE CELL.	Produces an unweighted 2-way crosstabulation (with χ^2 and ϕ) to test hypothesis

Output—Hypothesis 1

		State Residence		
Hunted in 2006		Colorado	Wyoming	Total
No	Count	623,989	82,105	706,094
	% within State of Residence	83%	65%	80%
Yes	Count	128,607	44,986	173,593
	% within State of Residence	17%	35%	20%
Total	Count	752,596	127,091	879,687
	% within State of Residence	100%	100%	100%

$\chi^2 = 14.72$, $p < .001$, $\phi = .183$, based on unweighted data.

Figure 3. Using SPSS and *SPORTSPERSON* entity variables—Hypothesis 1.

Example 2

Analysis Problem. This example uses the *TRIP_EXPENDITURES* entity and examines two hypotheses:

H_3 In-state annual lodging expenditures for *big game hunting* will vary between Colorado and Wyoming residents.

H_4 In-state annual lodging expenditures for *hunting and fishing* will vary between Colorado and Wyoming residents.

SPSS syntax—Hypothesis 2

SPSS syntax	Description
GET FILE='C:\Sportsperson.sav'.	Opens the *SPORTSPERSON* entity
WEIGHT BY Sportsperson_Weight.	Weights data by Sportsperson Weight
Select if (State_of_Residence = 8 or State_of_Residence = 56).	Selects Colorado and Wyoming residents
Select if (Sex = 0).	Selects males
AGGREGATE /OUTFILE='C:\Hunters.sav' /BREAK=State_of_Residence /Hunted_2006_sum=SUM(Hunted_2006).	AGGREGATE—Aggregates groups of cases in the *SPORTSPERSON* data entity into single cases based on the respondents' state of residence (Colorado or Wyoming), sums the number reporting hunting, and produces an aggregated data file (Hunters.sav)
GET FILE='C:\Person.sav'.	Opens the *PERSON* entity
Select if (Initial_State_of_Residence=8 or Initial_State_of_Residence=56).	Selects Colorado and Wyoming residents
Select if (Sex=0).	Selects males
Select if (Age GT 15).	Selects individuals older than 15 years of age
WEIGHT BY Person_Weight.	Weights data by Person Weight
AGGREGATE /OUTFILE='C:\State_Population.sav' /BREAK=Initial_State_of_Residence /Person_sum=SUM(Person).	Aggregates groups of cases in the *PERSON* data entity based on their state of residence, sums the number of people in the two states (Colorado and Wyoming), and produces an aggregated data file (State_Population.sav)
GET FILE='C:\State_Population.sav'.	Opens the State_Population data file
Compute State_of_Residence=Initial_State_of_Residence.	Compute equates respondents state of residence during wave 3 interviewing to their state of residence during the screening interview[1]
MATCH FILES /FILE=* /FILE='C:\Hunters.sav'.	MATCH combines cases from State_Population data file with Hunters data file
Compute Percent = 100 * Hunted_2006_sum / Person_sum.	Calculates the percent of male hunters
List variables = State_of_Residence Hunted_2006_sum Person_sum Percent.	For each state, LIST displays number of males hunting in 2006, number of males in the state's population, and the percent of males hunting in 2006

[1]In the sportsperson data 99.4% reported the same initial and wave 3 state of residence.

Output—Hypothesis 2

State of Residence	Number Males Hunting (2006)	State Population Males (2006)	Percent Males Hunting (2006)
Colorado	128,607	1,788,928	7.19
Wyoming	44,986	201,872	22.28

Figure 4. Using SPSS, *SPORTSPERSON,* and *PERSON* entity variables—Hypothesis 2.

SAS Code and Output—Hypothesis 3. The annual amount spent by a person on lodging in particular states for big game hunting is in *TRIP_EXPENDITURES*. Using a WHERE clause, the SAS code selects Colorado and Wyoming residents who had in-state lodging expenditures *greater than zero* (Figure 5). The WHERE clause in PROC TABULATE requests means and standard errors. Only one TABULATE procedure is necessary for producing information needed to test hypothesis 3.

Average big game hunting expenditures were $204 for Wyoming and $157 for Colorado (Figure 5). The difference ($204 – $157 = $47) was not statistically significant because the difference was less than the standard error ($86) in the Colorado mean. Hypothesis 3 was not supported.

SAS Code and Output—Hypothesis 4. Hypothesis 4 is concerned with expenditures reported for multiple hunting (e.g., big game hunting, small game hunting) and fishing activities. Unlike

SAS Code—Hypotheses 3

SAS code	Description
Proc Tabulate data=FHWAR6_2.Trip_ Expenditures;	Analysis based on *TRIP EXPENDITURES* entity
Where State_of_Residence IN (8,56) and Dollars > 0 and Trip_Expend_Category=2 and Spend_State_of_Residence=1 and Fish_Hunt_Type=–1;	Selects Colorado and Wyoming residents with lodging expenditures greater than 0, expenditures in state, and expenditures for big game hunting.
Weight Sportsperson_Weight;Var Dollars;	Weight DOLLARS by SPORTSPERSON_WEIGHT
Class State_of_Residence; Table State_of_Residence=' ', Dollars =' '*(mean='Mean'*f=dollar8.2 stderr='Standard Error'*f=8.2 sum='Total Dollars'*f=dollar12. n='# Cases'*f=5.)/rts=30 ;Run;	For each state of residence for weighted dollars give a mean, standard error, total dollars and number of cases used in making estimates.

Output—Hypotheses 3

	Annual Lodging Expenditures for Large Game Hunting[1]			
State	Mean	Standard Error	Total Dollars	Number of Cases
Colorado	$156.67	$86.29	$3,741,820	10
Wyoming	$203.83	$39.51	$1,633,084	10

[1]Numbers based on expenditures greater than zero dollars.

Figure 5. Using SAS and *TRIP_EXPENDITURE* entity variables—Hypothesis 3.

hypothesis 3 that only considered lodging expenditures greater than zero, hypothesis 4 examined two scenarios: (a) zero dollars excluded and (b) zero dollars included.

The SAS code (Figure 6) starts with a DATA step that aggregates in-state lodging expenditures reported by Colorado and Wyoming residents. Lodging expenditures are read in for different types of hunting (e.g., big game hunting, small game hunting) and fishing. The "BY PERSON_ID" statement facilitates adding a person's expenditure responses to a Sum initially set to zero. When all data for an individual have been read, DOLLARS is set to Sum. Two IF statements conditionally output information with the variable ZERO showing if an expenditure is "Zero dollars excluded" or "Zero dollars included." TABULATE is again used to output means, standard errors, total expenditures, and numbers of cases.

Figure 6 shows the results for zero dollars excluded and included. Exclusion or inclusion of zero dollars is important because some respondents may report zero dollars for in-state lodging expenditures, whereas others may not respond to these questions and thus are treated as missing values. When zero dollars were excluded, the average annual lodging expenditures were \$156 (*SE* = \$27) in Colorado and \$117 (*SE* = \$23) in Wyoming. Including zeros in the analysis increased the number of cases (*n* = 213 vs. 61 for Colorado, 155 vs. 29 for Wyoming) and reduced the means (*M* = \$41 for Colorado, *M* = \$21 for Wyoming with zeros included). As expected, totals of dollars are not influenced by including zeros.

Some differences between these means were significant. For example, for Colorado, the difference of \$114 between means with zero excluded (*M* = \$156) and with zero included (*M* = \$41) had a standard deviation of $28 = (27^2+9^2)^{1/2}$. This difference of \$114 was greater than four standard deviations and thus statistically significant. Average hunting and fishing lodging expenditures for Colorado and Wyoming (hypothesis 4), however, did not differ significantly when zeros were excluded (difference about one standard deviations). When zeros were included, \$42 is about 1.96 standard deviations larger than \$21. Based on the normal approximation and a two-tail test, this is significant at the 5% level. Acceptance or rejection of hypothesis 4 depends on whether zeros are excluded or included in computing means. Although not pursued here, including/excluding zero expenditures affects standard error in means.

Lessons Learned—Hypotheses 3 and 4. First, testing either hypothesis 3 or 4 in the Sportsperson flat file structure would necessitate using information in 639 trip expenditure variables. Similar to the large game hunting illustration, the trip expenditure data must be decompressed. Using the relational structure, only six variables (SPORTSPERSON_WEIGHT, FISH_HUNT, STATE_OF_RESIDENCE, SPEND_STATE_OF_RESIDENCE, TRIP_EXPEND_CATEGORY, DOLLARS) were needed for the analysis.

Second, Example 2 used SAS for illustration purposes because standard errors were calculated. Similar analysis in SPSS would yield weighted means identical to the SAS means. The weighted standard errors from SPSS, however, are incorrect unless the SPORTSPERSON_WEIGHT is adjusted appropriately for each mean computed (see Vaske, 2008, or online examples). Information in FHWAR publications could also be used to estimate standard errors (see Appendix D, USDI & USDC, 2006).

Third, including zeros in the means influences interpreting results. If including/excluding zeros can be justified and a valid approach gives a more reliable estimate, it should presumably be used. Because a response of zero versus no response (i.e., missing data) can be affected by a variety of factors (e.g., number of hunting and fishing trips, distances traveled), dealing appropriately with zero responses is not trivial.

SAS Code—Hypotheses 4

SAS code	Description
DATA fhwar6_2.yy;Set FHWAR6_2.Trip_Expenditures;	Uses *TRIP EXPENDITURES* entity
Where State_of_Residence IN (8,56) and Trip_Expend_Category=2 and Spend_State_of_Residence=1;	Selects Colorado & Wyoming residents Selects lodging expenditures Selects in-state of residence expenditures
BY Person_ID; IF First.Person_ID then sum = 0; retain sum;	BY creates the ability to identify when data for a person starts and stops. When data for a person starts set sum=0.

Sum=Sum+Dollars;
IF Last.Person_ID then do;
Dollars=sum;
IF Dollars > 0 then do; zero="Zero dollars excluded";
output;
end;
IF Dollars >= 0 then do; zero="Zero dollars included";
output;
end;
End;
	Add to Sum.
	For last data for a person output total as Dollars (=sum)
	If Dollars>0 output here.
	For every total output here (include 0).
	Use data created to form a table.

Proc Tabulate data=fhwar6_2.yy;	
Weight Sportsperson_weight;var dollars;	Weight DOLLARS by SPORTSPERSON_WEIGHT
Class Zero State_of_Residence; Table Zero*State_of_Residence=",	Classify by Zero (zero dollars included or excluded) by State (CO and WY) and aggregate responses.
Dollars ="*(mean='Mean' * f=dollar8.2 stderr='Standard Error' * f=8.2 sum='Total Dollars' * f=dollar12. N='# Cases' * f=5.) / rts=50 ; Run;	For classification give means, standard error, total dollars and number of cases

Output—Hypothesis 4

	Annual Lodging Expenditures for Hunting and Fishing			
	Mean	Standard Error	Total Dollars	Number of Cases
Zero Dollars Excluded				
Colorado	$156.01	27.04	$22,244,878	61
Wyoming	$117.13	22.37	$2,316,021	29
Zero Dollars Included				
Colorado	$41.68	8.82	$22,244,878	213
Wyoming	$21.48	5.48	$2,316,021	155

Figure 6. Using SAS and *TRIP EXPENDITURE* entity variables—Hypothesis 4.

Example 3

Analysis Problem. This example joins the *SPORTSPERSON* entity with *HUNTING_ACTIVITY* entity and examines the hypothesis:

H_5 Male sportspersons' participation in general types of hunting will vary between Colorado and Wyoming.

SPSS Syntax and Output—Hypothesis 5. The variable SUB_TABLE_ID in the *HUNTING_ACTIVITY* entity allows a convenient method for selecting particular data for analysis (Figure 7).[9] The syntax selects information about participation in big game, small game, migratory bird and "other game" types of hunting in state of residence (SUB_TABLE_ID=5). A new data file is saved containing two variables from *HUNTING_ACTIVITY* (PERSON_ID, FISH_HUNT_TYPE). The MATCH FILES command[10] joins the two variable file with data from the *SPORTSPERSON* entity using PERSON_ID as the linking variable. Once the files are joined, Colorado and Wyoming male hunters are selected. The data are weighted by SPORTSPERSON_WEIGHT and a crosstabulations is run. The SPORTSPERSON_WEIGHT is then turned off. A new weight variable (A_WEIGHT) is calculated based on SPORTSPERSON_WEIGHT to compensate for the population sizes in Colorado and Wyoming. Data are weighted by A_WEIGHT and a second CROSSTABS analysis performed. Results of the two crosstabulations are displayed in the output table.

When SPORTSPERSON_WEIGHT is used, 67% of big game hunters are from Colorado and 33% from Wyoming (first CROSSTABS, left side of output table, Figure 7). Because the weighted data represents 128,607 male hunters in Colorado and only 44,986 in Wyoming (see Example 1, Figure 4), row percents are not useful in testing hypothesis 5. Percents reflect both a difference in population size and any difference that exists in the participation rate. Using the A_WEIGHT adjusted to give 1,000 total in each state, a different pattern of findings emerges (second CROSSTABS, right side of output table, Figure 7). Unlike the first CROSSTABS (weighted by SPORTSPERSON_WEIGHT) that indicated Colorado residents were more likely to hunt big game than Wyoming residents; the second CROSSTABS (weighted by A_WEIGHT) revealed the opposite; Wyoming males were more likely to hunt big game than Colorado males.

Lessons Learned—Hypothesis 5. First, analysis of the relational data was based on nine variables. Comparable computations in the flat file structure would involve seven of the variables in the relational entities and 32 compressed variables. As before, decompression is necessary for state level flat file variables.

Second, the column totals shown in the output (i.e., 220,997 and 61,407, Figure 7) are not numbers of hunters. Individuals can participate in multiple hunting activities (e.g., big game, small game). In other words, people are counted more than once in the totals. The numbers of hunters were 128,607 in Colorado and 44,986 in Wyoming (Figure 4). To determine the percent of a group (e.g., Colorado males or Wyoming males) participating in a type of hunting, the number of participants must be divided by the size of the group.

Discussion

This article illustrated some advantages of restructuring the FHWAR flat file data as *PERSON*, *SPORTSPERSON*, *HUNTING_ACTIVITY*, and *TRIP_EXPENDITURES* relational databases. First, approximately Sportsperson 1,750 flat file variables were reduced to fewer than 60 relational variables. Second, the obtuse variable names in the flat file

SPSS syntax—Hypothesis 5

SPSS syntax	Description of SPSS syntax
GET FILE='C:\Hunting_Activity.sav'.	Opens the *HUNTING_ACTIVITY* entity
Select if (sub_table_ID=5).	Selects sub-table 5—Type of game hunted (e.g., big game, small game) in (state) in 2006
Select if (In_State_of_Residence=1).	Selects activity in-state of residence
SAVE OUTFILE='C:\Table_5.sav' / KEEP Person_ID Fish_Hunt_Type / COMPRESSED.	Saves 2 variables (PERSON_ID, FISH_HUNT_TYPE) in Table 5 into a new file (Table_5.sav)
GET FILE='C:\Table_5.sav'.	Opens the Table 5 data file
MATCH FILES / File=* / File ='C:\FHWAR_2006_Sportsperson.sav' / FIRST=Start	Joins Table 5 with the *SPORTSPERSON* Entity
/ KEEP=Person_ID Sportsperson_Weight State_of_Residence Sex Hunted_2006 Fish_Hunt_Type	Keeps selected variables from both entities
/ BY Person_ID. EXECUTE.	Matches files based on PERSON_ID
DO IF (Start=1) . COMPUTE #SPW=Sportsperson_Weight. COMPUTE #RES=State_of_Residence. COMPUTE #Sex =Sex. COMPUTE #H06=Hunted_2006. ELSE IF (Start=0). COMPUTE Sportsperson_Weight=#SPW. COMPUTE State_of_Residence=#RES. COMPUTE Sex=#Sex. COMPUTE Hunted_2006=#H06. END IF. EXECUTE.	Computes temporary variables (e.g., #SPW) for use during join Replaces temporary variables with original variable names
Select if (Hunted_2006=1) Select if (Sex=0). Select if (State_of_Residence=8 or State_of_Residence=56).	Selects individuals who hunted in 2006 Selects males Selects Colorado and Wyoming residents
SAVE OUTFILE='C:\State_GameType.sav'.	Saves the resulting file (State_GameType.sav)
WEIGHT BY Sportsperson_Weight. CROSSTABS /TABLES=Fish_Hunt_Type BY State_of_Residence /CELLS=COUNT ROW.	Weights by the Sportsperson Weight Produces a 2-way crosstabulation (left side of output table)
WEIGHT OFF	Turns off the Sportsperson Weight

If (State_of_Residence=8)
A_Weight=1000 * Sportsperson_Weight/128607.
If (State_of_Residence=56)
A_Weight=1000 * Sportsperson_Weight/44986.
	Creates an adjusted Sportsperson Weight (A_Weight) for number of males hunting in 2006 per state (see Figure 4. Output—Hypothesis 2 for denominators)

WEIGHT BY A_Weight.	Weights by the adjusted A_Weight
CROSSTABS /TABLES=Fish_Hunt_Type BY State_of_Residence /CELLS=COUNT ROW.	Produces a 2-way crosstabulation (right side of output table)

Figure 7. Using SPSS to join variables from *HUNTING_ACTIVITY* and *SPORTSPERSON* entities—Hypothesis 5.

Output – Hypothesis 5

		Unadjusted Weighted Data		State Population Size Adjusted Weighted Data	
		Colorado	Wyoming	Colorado	Wyoming
Big game hunting	Weighted Number	79,437	39,982	618	889
	% within category	67%	33%	41%	59%
Small game hunting	Weighted Number	90,374	14,162	703	315
	% within category	86%	14%	69%	31%
Migratory bird hunting	Weighted Number	46,703	5,794	363	129
	% within category	89%	11%	74%	26%
Other animal hunting	Weighted Number	4,483	1,469	35	33
	% within category	75%	25%	52%	48%
Total	Weighted Number	220,997	61,407		
	% within category	78%	22%		

Figure 7. (*Continued*)

were replaced with intuitive names. Third, and most important, in contrast to the compressed flat file variables that cannot be directly used in SPSS or SAS, variables in the relational entities can be used in analysis (e.g., crosstabulations, aggregation).

Restructuring the data facilitated analysis, but also highlighted the complexities in analyzing the FHWAR data. Example 1 (hypothesis 1) examined the relationship between state of residence (Colorado or Wyoming) and hunting participation in 2006 (No or Yes). Although the analysis was easy to perform with relational data, the percentages were deceiving because sportspersons are a sub-sample of individuals in the *PERSON* entity. Hypothesis 2 illustrated the steps necessary for correctly determining the denominator for estimating the percent of hunters in the two states' populations. Example 2 used trip expenditure data and highlighted decisions regarding inclusion or exclusion of zero expenditures. Example 3 addressed issues related to a single person hunting multiple types of game and demonstrated procedures for calculating correct percentages.

As noted in these examples, weighting FHWAR data is necessary to obtain population estimates. Statistics such as χ^2 or estimates of standard error produced based on weights should be viewed cautiously. Weights and other analysis considerations, however, do not arise because of structuring data relationally. Rather, obtaining valid results using FHWAR or other complex data necessitates precisely specifying the research question/hypothesis. Our results illustrate that some analyses that would be complex with flat file data can be quite simple with relationally structured data. Analysis is facilitated by values not being embedded in variables, which occurs when bivariate or trivariate information is flattened.

Restructuring all of 2006 FHWAR data as well as data from 1991, 1996, and 2001 would yield similar analysis capabilities and new practical opportunities for state fish and wildlife agencies. For example, the 2006 Sportsperson flat file contains approximately 300 variables related to hunting/fishing licenses, waterfowl stamps, and special fees. These variables are compressed and contain answers to general questions such as: (a) For

which state(s) did you buy a license to hunt? (b) How many hunting licenses did you have for (state) in 2006? and (c) Concerning your (first/second/third/fourth/fifth) license of (state), how much did it cost? Responses to these questions must be decompressed before they can be used for analysis. More importantly, the utility of such general questions is questionable. State agencies have accurate records of how many licenses of a particular type (e.g., hunting only, fishing only, combination hunting and fishing) were sold and know how much each specific license costs. Because license types vary substantially by state, most responses in the flat file structure cannot be used to examine the relationship between license purchase records and FHWAR data (for an exception see Beaman & Vaske, 2005). With hundreds of licenses, permits, and stamps sold, it is not practical to be specific about licenses in flat file data. Moving to a relational structure for obtaining license data has advantages. First, interviewers could ask questions about actual state specific licenses. All state license information would be "pre-stored" in a single entity. The size of this entity would not impact other relational data entities. Second, questions about the cost of a specific license would not be necessary because correct information about licenses and their cost would be pre-stored. Third, and most important, establishing a relationship between state-specific license sales and FHWAR data would provide a foundation for benchmarking and calibrating meaningful annual estimates based on FHWAR.

Although numerous national and state level FHWAR reports have been produced (see, for example, http://www.census.gov/prod/www/abs/fishing.html), our search of the literature identified relatively few scientific journal articles that have used FHWAR data. This may partially be attributable to the challenges associated with using FHWAR's flat file data structures. By making the data available in user-friendly relational formats (i.e., SPSS and SAS) more academic and agency researchers have access to the information.[11] We hope that this article encourages restructuring other components of the 2006 data (e.g., fishing activities, wildlife viewing), as well as data from previous data collection years, into relational form. An entity based merger of FHWAR data from 1991, 1996, 2001, and 2006 would facilitate analyzing trends in hunting, fishing, and wildlife viewing.

Notes

1. Data for the 2006 FHWAR survey were collected in three waves with in-person and telephone screening of households. The first wave (April and May 2006) was a "Screening" interview. See U.S. Department of the Interior [USDI], Fish and Wildlife Service, and U.S. Department of Commerce, U.S. Census Bureau [USDC] (2006) for a complete description of the methodology.
2. As defined by USDI and USDC (2006), there are four sportspersons categories: (1) active, (2) likely, (3) inactive, and (4) non-participant (see p. 149 for details). Page 2 of the USDI and USDC (2006) report, however, defines "sportspersons as those who fished or hunted."
3. NCU_STD1=In which state(s) did you take trips or outings to observe, photograph, or feed wildlife? NCUTOD1=How many trips lasting a single day or multiple days did you take in or to "state" from January 1, 2006 to December 31, 2006 primarily to observe, photograph, or feed wildlife?
4. For illustration purposes four entities were constructed. Other entities (e.g., *FISHING_ACTIVITY, WILDLIFE_WATCHER, LICENSES*) could be created. In addition, all expenditure data could be in a generalized *TRIP_EXPENDITURES* entity by expanding the expenditure categories.
5. When the Screening data are weighted, the data approximate the United States population over five years of age (USDI & USDC, 2006).
6. Because the Census Bureau's Sportsperson sample is a sub-sample of the screening sample, a sportsperson weight is used to generalize to a hypothetical population of hunters and anglers 16+ years of age in the United States (see Appendix D of USDI & USDC, 2006). The Census Bureau

cautions against using a specific weight if the variables come from different flat files (i.e., Screening, Sportsperson, Wildlife Watcher). The caution applies to the relational entities created from the flat files. The sportsperson weight should be used when analyzing the *SPORTSPERSON, HUNTING_ACTIVITY,* and *TRIP_EXPENDITURES* entities.

7. The website for downloading these files is http://welcome.warnercnr.colostate.edu/~jerryv/.
8. A male sportsperson can include anglers, individuals who hunted in 2006, and individuals who were included in the sportsperson file based on their likelihood of hunting, but did not actually hunt.
9. Online documentation gives values for variables such as SUB_TABLE_ID.
10. In SPSS using the menu option of adding variables does not work to join one record to many. "MATCH" can be used as shown in Figure 7.
11. Although this article endorsed using relational databases, SQL (Structured Query Language, see Groff & Weinberg, 2002) was not pursued because many human dimensions researchers are more fluent in SPSS or SAS than in SQL.

References

Avedon, E. M. (1992). Implementing entity-relationship theory for handling leisure and cultural information. *Society and Leisure, 14*(1), 35–46.

Beaman, J. G., & Vaske, J. J. (2005). Reliability, accuracy and bias in survey estimates in tourism surveys. *eRTR Review of Tourism Research, 3*(3), 54–60.

Chen, P. P. (1976). The entity-relationship model—Toward a unified view of data. *ACM Transactions on Database Systems, 1*, 9–36.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13*(6), 377–387.

Grambsch, A. E., & Fisher, W. L. (1989). History and development of the national survey of fishing, hunting and wildlife-associated recreation. *Wildlife Society Bulletin, 17*, 538–543.

Groff, J. R., & Weinberg, P. M. (2002). *The complete Reference SQL* (2nd edition). New York: McGraw-Hill.

International Association of Fish and Wildlife Agencies. (1953). Proceedings of the 43rd Convention of the International Association of Game, Fish and Conservation Commissioners, September 11–12. Dallas, Texas, USA.

International Association of Fish and Wildlife Agencies. (1956). Proceedings of the 46th Convention of the International Association of Game, Fish and Conservation Commissioners, September 13–14. Toronto, Ontario, Canada.

McDowell R., & Mock, J. (2004). The national survey of fishing, hunting and wildlife-associated recreation: A history of partnership between state fish & wildlife agencies, the US Fish & Wildlife Service, & non-governmental conservation organizations since 1955. Unpublished manuscript. Washington DC: International Association of Fish and Wildlife Agencies. January.

U.S. Department of the Interior, Fish and Wildlife Service, and U.S. Department of Commerce, U.S. Census Bureau [USDI & USDC]. (2006). *2006 National Survey of Fishing, Hunting and Wildlife-Associated Recreation (FHW/06-NAT).* Washington, DC.

Vaske, J. J. (2008). *Survey research and analysis: Applications in parks, recreation and human dimensions*. State College, PA: Venture Publishing Inc.

Vaske, J. J., Gliner, J. A., & Morgan, G. A. (2002). Communicating judgments about practical significance: Effect size, confidence intervals and odds ratios. *Human Dimensions of Wildlife, 7*, 287–300.