RAKING FIRE DATA

Michael A. Greene, Linda E. Smith, Mark S. Levenson, Signe Hiser and Jean C. Mah
U. S. Consumer Product Safety Commission

Keywords: missing data, imputation, raking.

Abstract

The U. S. Consumer Product Safety Commission staff uses raking to impute unknown or missing values in fire incident reports. In typical analyses, up to six variables are imputed. These variables describe the cause of the fire and the way that it propagates through a building. Typically, 25 to 40 percent of the values of these variables are unknown. All the variables imputed are categorical, with up to 100 values. The full crosstabulation may run 100,000 cells or more, with modal cell count of zero. In this context, raking can become unstable by either failing to converge, or can produce results where some cell counts are less than their original, pre-imputed values. This paper describes some strategies for raking high dimensional sparse tables.

1. Introduction

The U.S. Consumer Product Safety Commission (CPSC) staff analyzes fire incident data to determine patterns and losses associated with fires that involve household products. Losses mean fire deaths, injuries and property damage. The purposes of these analyses are (1) to support or evaluate standards that would make household products less likely to ignite and (2) to identify products that pose new hazards and the patterns of usage that are associated with such hazards. Attributing fire losses to fire causes is an important part of this task. Recent studies at CPSC have included estimates of the number of fire incidents, injuries and deaths associated with upholstered furniture, mattresses, ranges (Smith and Greene, 2001), and disposable cigarette lighters (Smith, Singh, Greene, 2001). Fire loss estimates for a wide range of products are found in CPSC's annual Residential Fire Loss Estimates (Mah, 2001).

The source of fire incident data and fire loss data is the U. S. Fire Administration's National Fire Incidence Reporting System (NFIRS). NFIRS is a large, automated database of detailed fire incident reports from local fire departments. It covers about 40 percent of the fires, injuries and deaths in the United States. Each fire incident has a separate record in the database and is described by over 100 hundred variables. These variables are coded according to the Uniform Coding for Fire Protection (National Fire Protection Association, 1976). National level product-specific estimates are developed by scaling up the NFIRS product-specific totals by the estimated residential fire losses in the National Fire Protection Association's annual survey of Fire Loss in the United States (Karter, 1999). The procedure, known as the "national estimates method", is described in Hall and Harwood (1989).

One important aspect of NFIRS is that it contains a detailed coded description of the cause of each fire. Most of the information is contained in six variables. These are (1) *area of origin,* i.e. where the fire started; (2) *equipment involved in ignition*, which describes the product providing the heat that started the fire such as a stove, furnace, etc.; (3) *form of heat of ignition,* describing the heat energy that started the fire, e.g. spark, hot surface, etc.; (4) *type of material first ignited*, such as paper, liquid fuel, solid fuel or fabric (5) *form of material first ignited*, e.g. walls, furniture, clothing, etc. and (6) *ignition factor*, the factor that brought the heat of ignition and material first ignited together, e.g. arson, short circuit, product defects or misuse of materials.

For a specific analysis, say, fires and fire losses involving kitchen stoves, it is necessary to identify the relevant values of these variables, such as *area of origin*=kitchen, *equipment involved* = range tops or ovens, *form of heat* = natural gas, propane or electricity, *type of material first ignited*= cooking materials, and *ignition factor* =falling asleep, inadequate control, spillage, mechanical failures, etc. Then fires, deaths or injuries with these values for the variables would be counted, and then scaled up to national levels using the estimates from the NFPA survey. However, about 40 percent of the values of the NFIRS variables are missing. To avoid estimates that are negatively biased, it is important to impute missing values, before scaling up to national levels.

There are two categories of missing values for an NFIRS variable. First, the value may appear as blank, missing or unknown in the database. This means that there is no information to distinguish, for example, if the form of heat was natural gas, electricity, fireworks, or lava. These missing data are the first of two imputation problems, involving allocating these unknowns over all the counts in the cells with known values. There is also a second way that data are missing in NFIRS, denoting that there is some information available but not enough for a definitive classification. Most codes for NFIRS variables are grouped in decimal series where values within a series are similar. One value in each series is reserved for “insufficient information available to classify further.” For example, in the *form of heat* variable, smoking materials occupy the 30 decimal series. Within this series, 31 is Cigarettes, 32 is Cigars, 33 is Pipes, 39 is Heat from Smoking Materials not classified above and 30 is Heat from Smoking Materials; insufficient information available to classify further. This means that the form of heat was some smoking material, but it was not possible to determine if it was a cigarette, cigar, pipe or other material. The imputation problem here is to assign the unknown within the decimal series to the known values, but only within the decimal series.

Since the mid 1980’s, missing values have been imputed following a procedure in the national estimates method (Hall and Harwood, 1989). Variables are imputed (i.e. fit to the known marginals) one at a time without any iterations. The result is that the first variable imputed fits the marginals exactly, but the other variables do not. For example, using the national estimates method, the unknown values of form of heat would first be allocated to the known values. Then unknown values of area of origin, and then the next variable. Since only one set of marginals fit at the end of the process, the imputed values are sensitive to the order of the variables used in imputation. With no theoretical reason to prefer one ordering over another, and with up to 6! = 720 possible orderings, it is virtually guaranteed that two analysts would obtain different results from the same data. Hall and Harwood (1989) acknowledged this when they wrote “…it requires the analyst to choose one dimension as primary…both of these approaches will produce different results….” (page 108). To improve this process, raking was implemented for analysis of the fire data. The results with raking do not depend on the order in which the variables are imputed.

This paper is about CPSC staff experiences using raking. Section 2 briefly introduces raking and reviews the literature. The requirement for fairly high dimensional analysis (six variables, each with up to 40 or 50 values), results in very sparse tables, especially when analyzing fire deaths. Also, as a result of different categories of missingness (i.e. the two levels described above), it is necessary to separately rake parts of the table. The structure for this is described in Section 3. Section 4 describes problems that have been encountered as a result of the high dimensionality and sparseness. Section 5 contains the conclusion.

2. Raking Described

Consider a two dimensional table with observed cell counts, n_{ij}, unknown population cell counts, N_{ij} and estimates of the population cell counts N_{ij}^*. Marginal sums $\sum_j N_{ij} = N_{i+}$ and $\sum_i N_{ij} = N_{+j}$ are known. As pointed out in Little and Rubin (1987, 59), raking applies to the individual cell counts, n_{ij}, to iteratively calculate estimates that satisfy marginal constraints $N_{i+}^* = \sum_j N_{ij}^* = N_{i+}$ and $N_{+j}^* = \sum_i N_{ij}^* = N_{+j}$ by using multiplicative row and column constants, a_i, and b_j where $N_{ij}^* = a_i b_j n_{ij}$. That is the individual cell counts are adjusted to the marginal totals.

Iterative proportional fitting (IPF) is used to adjust the cells to marginal totals. At the first step of the procedure, estimators are calculated $N_{ij}^{(1)} = n_{ij} N_{i+} / n_{i+}$. This matches the row marginals exactly, but the column marginals are unlikely to agree with the known values. The next iteration adjusts the individual cells to the column marginals by $N_{ij}^{(2)} = N_{ij}^{(1)} N_{+j} / N_{+j}^{(1)}$. Then the row marginals are adjusted by $N_{ij}^{(3)} = N_{ij}^{(2)} N_{i+} / N_{i+}^{(2)}$. Iteration between rows and columns continues until convergence is achieved, where convergence is defined as $\left| N_{i+}^* - N_{i+} \right| < e$ and $\left| N_{+j}^* - N_{+j} \right| < e$ for some small value e. Both iterative proportional fitting and raking are attributed to Deming and Stephan (1940).

The next few tables show an hypothetical example for a 2x2 problem with an additional unknown row and unknown column. The example adjusts columns first instead of rows, but the principles are the same. Table 1 shows the table containing the unknowns.

Table 1
Original Problem

	Female	Male	Unknown	Total
Old	65	30	5	100
Young	25	50	25	100
Unknown	10	2000	70	2080
Total	100	2080	100	2280

Before raking, the unknown marginals are distributed to the known marginals in proportion to the value of the known marginals, (e.g. the Female marginal is 100 * (2280/2180)=104.6). The table without the values of the unknowns is shown in table 2 below. This is now ready for raking.

Table 2
Raking Setup

	Female	Male	Total	Population	Difference
Old	65	30	95	*1140*	1045
Young	25	50	75	*1140*	1065
Total	90	80	170		
Population	*104.6*	*2175.4*			
Difference	14.6	2095.4			

Population totals of 104.6 and 2175.4 for the columns are shown above, and are different from the computed marginals by 14.6 and 2095.4, respectively. The first iteration involves multiplying the entries in the first column by the ratio of population to computed marginals (104.6/90) and the second column by the ratio (2175.4/80). The values are shown in table 3.

Table 3
First Raking Step: Columns

	Female	Male	Total	Population	Difference
Old	75.5	815.8	891.3	*1140*	248.7
Young	29.1	1359.6	1388.7	*1140*	-248.7
Total	104.6	2175.4	2280.0		
Population	*104.6*	*2175.4*			
Difference	0.0	0.0			

While the column marginals have been adjusted to the population totals, the row marginals are now off. The appropriate multipliers for the row marginals are 1140/891.3 and 1140/1388.7, respectively. This results in table 4.

Table 4
First Raking Step: Rows

	Female	Male	Total	Population	Difference
Old	96.6	1043.4	1140	*1140.00*	0.0
Young	23.9	1116.2	1140	*1140.00*	0.0
Total	120.5	2159.6	2280		
Population	*104.6*	*2175.4*			
Difference	-15.9	15.8			

The application of row multipliers perfectly aligns the rows at the expense of the columns. The next iteration multiplies entries in the first column by 104.6/120.5 and the second column by 2175.4/2159.6. The result is in table 5.

Table 5
Second Raking Step: Columns

	Female	Male	Total	Population	Difference
Old	83.9	1051.1	1134.9	*1140.0*	5.1
Young	20.7	1124.3	1145.1	*1140.0*	-5.1
Total	104.6	2175.4	2280.0		
Population	*104.6*	*2175.4*			
Difference	0.0	0.0			

The reader can verify that one more adjustment to the rows, using multipliers 1140/1134.9 and 1140/1145.1 brings the population and calculated marginals within ±0.3 in both dimensions.

In the context of the framework that $N_{ij} = a_i b_j n_{ij}$, or written as probabilities, $p_{ij} = a_i b_j p_{ij}$, the problem can be viewed as a loglinear model $\log(p_{ij} / p_{ij}) = \log a_i + \log b_j + e_{ij} = a_i + b_j + e_{ij}$, where p_{ij} and p_{ij} are the population and sample cell probabilities, respectively. Alternatively, the fitted value $\hat{p}_{ij}$ can be expressed as $\log(\hat{p}_{ij} / p_{ij}) = a'_i + b'_j + m$, where $\sum_i a'_i = \sum_j b'_j = 0$.

Little and Wu (1991), compare other estimators of the form $g(\hat{p}_{ij} / p_{ij}) = a'_i + b'_j + m$, where g is the identity, inverse or inverse square function. Under various simulated assumptions, raking (g = log) performed as well as or better than the other estimators.

One interesting property of raking is that it preserves the sample odds ratio. Consider four cells after raking as

$$\frac{N_{11} / N_{12}}{N_{21} / N_{22}} = \frac{a_1 b_1 n_{11} / a_1 b_2 n_{12}}{a_2 b_1 n_{21} / a_2 b_2 n_{22}} = \frac{n_{11} / n_{12}}{n_{21} / n_{22}} .$$

3. Raking at CPSC

As noted above, before raking was used by CPSC staff, unknowns were allocated using the national estimates method. That method is identical to the first raking step. Table 3 above contains the values that would be obtained from that method. The national estimates method works the following way in a two dimensional table:

1. Every entry in the table is scaled by the ratio of column (or row) totals to knowns.
2. Different factors within each row (column) are created as the ratio of row (column) totals to knowns. The table is then scaled by those factors.

Because it is non iterative, the national estimates method can be done by hand for small tables or with some elementary programming for more larger ones. But, the final cell values are sensitive to the order of adjusting the cells. For example, using the national estimates method in the same order as the tables above, the young male cell is 1359.6. If we reversed the order of adjustment of the variables, the value in the table would have been 760. (The raking solution is 1119.4.)

Table 6 presents a scaled down version of a typical problem at CPSC.

Table 6
Fire Deaths by Form of Materials and Form of Heat
(Scaled Down Hypothetical CPSC Raking Problem)

	Form of Heat							
Form of Materials	**Fuel Fired In Scope**	Fuel Fired Not In Scope	Fuel Fired Unknown If In Scope	**Smoking Materials In Scope**	Smoking Materials Not In Scope	Smoking Materials Unknown If In Scope	Unknown	Total
Not Furniture	55	20	13	14	51	20	310	483
Furniture: Not in Scope	21	4	18	16	74	20	155	308
Upholstered Furniture	**4**	21	18	**12**	15	30	110	210
Unknown Furniture	8	7	1	2	21	3	105	147
Unknown	12	21	2	14	18	16	256	339
Total	100	73	52	58	179	89	936	1487

Note: Like the last few tables, the values in the table do not represent actual data.

At CPSC, the objective of fire data analyses is to identify fires and fire losses associated with various consumer products and fire causes. In this hypothetical example, the requirement is to produce estimates for Upholstered Furniture fire deaths from some “In Scope” Smoking Materials (cigars, pipes and cigarettes) and from “In Scope” Fuel Fired Form of Heat. (In Scope Smoking Materials represents the codes for cigars, pipes, cigarettes and other smoking materials all collapsed together into a single category). In Scope Fuel Fired Form of Heat includes gas and liquid fueled equipment, generally representing cigarette, cigar and pipe lighters. These categories are shown in bold in the table. Fuel fired unknown (column 3), smoking materials unknown (column 6) and unknown furniture (row 4) are the decimal series, or partial unknowns.

The first stage of raking begins where the cell counts for all the other values are adjusted to the new marginals. This continues until convergence.

In the second stage, the two subproblems are tackled separately, the first involving furniture and fuel-fired and the second involving furniture and smoking materials. Table 7 shows the furniture fuel-fired subproblem. The row representing “Not Furniture” is deleted from the table because there will be no further imputation with that row. During the second stage, "Unknown Furniture" is raked into "Upholstered Furniture" and "Furniture: Not in Scope" and the two remaining "Unknown" columns have their totals raked into the appropriate columns. These two subproblems have to be handled separately because the marginals can only be allocated to the specific rows and columns remaining within each subproblem.

Table 7
One Second Stage Subproblem: Furniture and Fuel-fired

	Form of Heat			
Form of Materials	Fuel Fired In Scope	Fuel Fired Not In Scope	Fuel Fired Unknown If In Scope	Total
Furniture:Not In Scope	49.4	12.3	45.9	107.6
Upholstered Furniture	**9.1**	62.2	44.2	115.5
Unknown Furniture	32.6	37.3	4.4	74.3
Total	91.1	111.8	94.5	297.4

Following the second stage of raking, the two "unknown" columns would be removed. The final output would be the two values for upholstered furniture, one from Fuel Fired In Scope and the second from Smoking Materials In Scope.

This is a small version of a raking problem at CPSC. In the analysis leading to estimates for range fires (Smith and Greene, 2001), the first stage data matrix had five variables (dimensions) with a total of 59,400 cells. These included the original rows and columns with Unknown values, similar to table 5. The second stage of raking involved 160 separate subproblems.

Raking is accomplished with a set of SAS macros written at Abt Associates (Izrael, Hoaglin and Battaglia, 2000). The inputs to the raking algorithm are the names of input and output datasets, the names of the input weights (original cell frequencies are used as weights), the marginal control (population) totals, the number of variables (dimensions) for raking and the names of those variables, the total for the population marginals, the convergence criteria and the number of iterations. We typically set the convergence criteria at 1 (all computed marginal sums must be within 1 of the population totals) and the maximum number of iterations at 100. It is rare to go beyond 50 iterations. Typically 10-30 iterations are required. The output is a dataset containing the resulting raked weights (cell values). Both input weight datasets and output weight datasets are in the form of an output dataset from PROC FREQ.

A set of macros written by CPSC staff set up the subproblems for the second stage of raking. We code the values of the variables as follows: the first character is a numeric digit, denoting the part of the data matrix that it belongs to and the next few characters are unique values corresponding to a unique row (column, etc.) of the submatrix. These last few characters essentially denote an instruction to the software. The character strings, "Rake" and "Out" are reserved. The word "Rake" is an instruction to change the value to "Unknown" and to rake out that row (column). "Out" means that the row or column is to be deleted before starting the raking step. Any other character string would be treated as something to be reported on after completing raking. These character strings are created using PROC FORMAT in SAS, from the original numeric digits in the database. For example, typical values for the variable Form of Heat would be "30CigarettePipe," aggregating over the individual categories for Cigarette, Cigar and Pipe; "30SmokeOther;" and "30Rake;" the last representing the category "Smoking Materials insufficient information to classify further." Upholstered furniture would be coded as "21Furniture" and unknown furniture would be coded as"20 rake.""20Rake."

Individual subproblem membership is identified by concatenating the first character of each of the values of the variables into a single character string. In two dimensions, for example, the subproblem with Smoking Materials and Furniture would be "32," the first digit, 3, from Smoking Materials and the second digit, 2, for furniture. In one pass through the data, the appropriate value of the subproblem is identified and added to each record. The structure of the record is the output from PROC FREQ, as discussed above, containing the values of each variable, the cell count and finally the subproblem to which it belongs.

A SAS macro then counts the number of unique subproblems. PROC SQL is then used to assign each distinct subproblem value to a macro variable that is indexed from 1 to the number of subproblems. This then sets up a loop on the number of unique subproblems with the following steps: (1) a dataset is created with only the observations that are in the appropriate subproblem, (2) any variable with the last four characters "Rake" has the value of that variable changed to UNKNOWN for subsequent raking, (3) the raking algorithm is then invoked for the subproblem, and (4) the output is then appended to the final dataset. The loop continues until all submatrices have been raked. This procedure implements "by group" processing that is similar to the suggestion in Izrael, Hoaglin and Battaglia.

In practical applications, we try to group values of variables whenever possible, almost always when we will report aggregates in the final output. With five variables and 60,000 cells, just a few more values of several variables could result both in large growth in the number of cells submitted to the original raking problem, and the number of subproblems. With typical frequency counts for a single year's data at about 1,300 deaths, 11,000 injuries and 150,000 fires, sparseness can become a problem. This is described in the next section.

4. Problems

Because raking is a multiplicative adjustment to the cell counts, cells containing zero counts pose a serious problem for raking. Consider the following table:

Table 8
A Zero Row Can Lead to Non-Convergence

	Female	Male	Total	Population	Difference
Old	1	25	26	*108.9*	82.9
Young	0	0	0	*66.1*	66.1
Total	1	25	26		
Population	*117.1*	*57.9*			
Difference	116.1	32.9			

There is no way to make progress with this table, because there is no way to adjust the "Young" row to the population marginals. To avoid this problem, we replace zeroes in the tables with a small quantity, say, epsilon, defined as 10^{-6}. In tables where there are occasional zeroes, this does not make a difference in the result because most numbers are considerably larger. However, in a situation such as this where the entire row or column turns out to be zero, adding a small quantity allows the program to converge.

After replacing zeroes with small quantities we have occasionally seen the estimates shrink after raking. Since the values in the table represent counts, and the population (marginal) totals are always larger than computed marginals from the tables, decreasing counts are an undesirable situation. Tables 9 and 10 contain an extreme example of this situation.

Table 9
Initial Data: A Zero Can Lead to Shrinking Estimates

	Female	Male	Total	Population	Difference
Old	1	**25**	26	*108.9*	82.9
Young	0	15	15	*66.1*	51.1
Total	1	40	41		
Population	*117.1*	*57.9*			
Difference	116.1	17.9			

Table 10
Final Data: Shrinking Estimates After Raking

	Female	Male	Total	Population	Difference
Old	109.6	**0.0**	109.6	*108.9*	-0.7
Young	7.5	57.9	65.4	*66.1*	0.7
Total	117.1	57.9	175.0		
Population	*117.1*	*57.9*			
Difference	0.0	0.0			

The actual value in the cell "Old-Male" is zero to 2 decimal places. Somehow, 25 old males have become lost in the exercise of the raking procedure.

Why does this happen? The raking algorithm, in trying to fit the marginal gender distribution, will put as many females as possible into the "Old-Female" cell because the "Young-Female" cell has an epsilon (essentially a zero). This then distorts the age distribution, because there are too many "Old" of both genders. Next, "Old-Males" are lowered below the initial estimate. The number of "Young-Males" can be increased without distorting the age distribution, because the counts in the "Young-Female" cell have to remain low. But Old-Males cannot be similarly increased. The final result is that the (raked) number of "Old-Males" are fewer than the initial counts.

Partly at the root of this problem is that initially, the age distribution of the men is relatively well known, while there is almost no data on females. Compounding this, more than two thirds of the data are known to be females from the population totals. The problem seems to be to infer the age distribution of the females from very little knowledge. Distorted results then should not be surprising.

As part of the analysis of the output, we always track decreasing cell counts. We have rarely seen anything of practical importance (say more than a 10 percent decrease in a large cell), but this does not rule out that such a problem can occur.

We received several excellent suggestions on how to approach this problem. Alan Dorfman (2001) suggested a weighted optimization approach, where the final values, N_{ij} are a function of the product of the original values, n_{ij} and the multiplier that represents the proportion of known elements relative to that cell. Graham Kalton (2001) suggested that the objective was actually to increase the count in the zero cell until the procedure did not shrink the estimate below the original values. Implicit in both approaches is an optimization scheme that minimizes the departure from the cell counts in the original table while at the same time preserving all the features of raking. We have not implemented any of these approaches, but we have carefully compared the raked output with the original values to determine when shrinkage occurs. The possibility of shrinkage has also motivated as much collapsing of categories as possible to avoid having zero or near zero cells.

A second problem results from the fact that we impute each type of fire loss separately. This has been traditional because it has been believed that the reporting pattern differs for each type of fire loss, implying that the pattern of missing data also differs by fire loss. The risk of separate, rather than joint imputation, is that there may be an imputed fire death in a cell where there actually does not turn out to be any fires. This may be a consequence of filling the table with small quantities to replace the zero cells. The small quantities can be inflated during raking. Judkins (1997) called this a “swiss cheese” pattern, pointing to the zeroes in the sparse tables as the holes. Generally, imputing fires through raking, or any other procedure, is much safer than imputing fire deaths because there are far fewer zero cells, and less of a chance of the type of misbehavior produced by sparse tables as shown above. It might also be feasible to impute fires, put imputed counts in the data for each incident, then weight the number of deaths and injuries by the imputed number of fires to arrive at the number of fire losses.

5. Conclusion

This paper has reviewed strategies and problems involved in imputing unknown fire causes using raking. Raking is an improvement over the previous “national estimates” procedure because the estimates do not depend on the order that variables are imputed. But raking imposes more demands on the underlying structure in the data. The most serious problem encountered has been that imputed values occasionally were less than original cell counts. Small cell values, which are a consequence of the high dimensionality data sets required to identify fire causes, were usually at the root of this.

The strategy used by CPSC staff has been to collapse cells together whenever possible to minimize the problems of sparse counts from high dimensionality. Other approaches involving optimization are under consideration, should these problems persist and be of sufficient magnitude to distort estimates.

References

Deming, WE and Stephan, FF (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known." *Annals of Mathematical Statistics*, 11, 427-444.

Dorfman, A (2001), personal communication.

Judkins, DR (1997), "Imputing for Swiss Cheese Patterns of Missing Data." Statistics Canada Symposium Proceedings.

Kalton, G (2001), personal communication.

Mah, J (2001), "1998 Residential Fire Loss Estimates: U. S. National Estimates of Fires, Deaths, Injuries and Property Losses from Non-Incendiary, Non-Suspicious Fires." U.S. Consumer Product Safety Commission, Washington, DC.

Hall, JR Jr and Harwood, B, (1989), "The National Estimates Approach to U. S. Fire Statistics." Fire Technology, pp. 99-113.

Izrael, D, Hoaglin, DC, and Battaglia, MP (2000), "A SAS Macro for Balancing a Weighted Sample." Paper 258 SUGI (SAS Users Group International) 25. The SAS programs are available from the authors at Abt Associates, Cambridge, MA.

Judkins, DR (1997), "Imputing for Swiss Cheese Patterns of Missing Data." Statistics Canada Symposium Proceedings.

Little, R and Rubin, D (1987), Statistical Analysis with Missing Data. New York: John Wiley and Sons.

National Fire Protection Association, (1976), "Uniform Coding for Fire Protection, 1976." Boston, MA: National Fire Protection Association.

Smith, L and Greene, MA (2001), "Range Fires." US Consumer Product Safety Commission, Washington, DC.

Smith, L, Greene, MA and Singh, H (2001), "A Study of the Effectiveness of the U.S. Safety Standard for Child-Resistant Cigarette Lighters." US Consumer Product Safety Commission, Washington, DC.