Chapter 2

Methodology

This chapter describes the method used in the study including study design and sampling technique, variables and conceptual framework, data collection and management, and statistical methods.

2.1 Study design and sampling technique

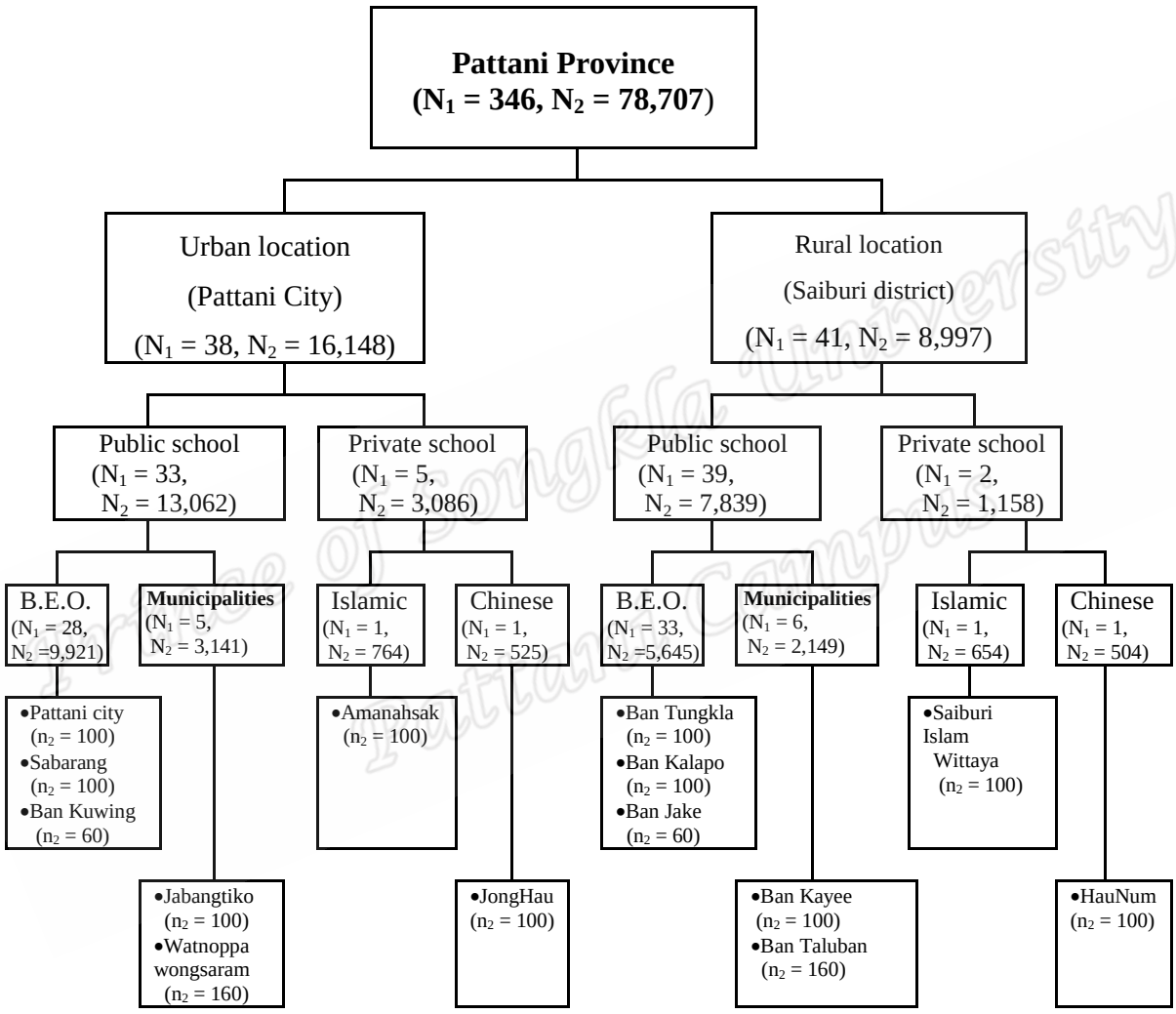
This study used a cross-sectional study design involving interviews and surveys of primary school students in a sample selected from the target population studied.

The target population comprised all students at Pattani primary schools attending school between November 1, 2005 and March 31, 2006. The participants were selected by using a multi-stage sampling method.

The first stage involved selecting school location by using purposive sampling, with the criterion being a cluster of four types of school (public school of Basic Education Office (B.E.O.), public school of municipalities, Islamic private school, and Chinese private school). Pattani City was selected as the urban location and Saiburi district as the rural one, because these were the only two districts that met the school-type cluster criterion.

In the second stage, public schools were selected by simple random sampling and private schools were selected by purposive sampling (there was only one of each such school in each district).

Finally, participants in each school grade were selected by using a systematic sampling technique which was done proportionate to population size across each class; choosing every 4th seat number where there was a single class in a grade and every 6th seat number where there was more than one class in a grade.



Note: N_1 is number of schools and N_2 is number of students

Figure 2.1: Multi-stage sampling method

Sample size calculations followed an Italian study of bullying (Baldry, 2003) and were based on the main outcome and exposure to parental violence and non-exposure to parental violence. The prevalence of bullying by the Italian primary school students

in the 'non-exposure to parental violence' group was 45.7%. This information was then used to calculate the required sample size for this study, obtaining an estimate by substituting $\alpha = 0.05$, 1-$\beta = 0.2$, $OR = 1.344$ so $Z_{\alpha/2}$ and Z_{β} are 1.96 and 0.84 respectively, $r = 1$ (ratio of non bully to bully subjects), $p_2 = 0.46$ (prevalence of bullying in non exposure to parental violence group), $p = 0.50$, $p_1 = 0.53$, into a formula for sample size given by the following (McNeil, 1996), namely

$$n_1 = \frac{\left(Z_{\alpha/2}\sqrt{\left(1+\frac{1}{r}\right)\frac{1}{p(1-p)}} + Z_{\beta}\sqrt{\frac{1}{p_1(1-p_1)}+\frac{1}{rp_2(1-p_2)}} \right)^2}{(\mathrm{In}OR)^2} \quad (2.1)$$

Where $p_1 = \frac{p_2}{p_2 + (1-p_2)/OR}$, and $p = \frac{p_1 + rp_2}{1+r}$

The sample size of the study was then calculated as

$$n_1 = \frac{\left(1.96\sqrt{\left(1+\frac{1}{1}\right)\frac{1}{0.50(1-0.50)}} + 0.84\sqrt{\frac{1}{0.53(1-0.53)}+\frac{1}{(1\times 0.46)(1-0.46)}} \right)^2}{(\mathrm{In}1.344)^2}$$

$$= 718.6$$

This gives $n_1 = n_2 = 719$. It was thereby concluded that a minimum sample size of 1,438 was required for this study.

2.2 Variables and conceptual framework

Determinant variables

The determinant variables in this study comprised school factors (school type and location), demographic factors (gender, age group, and religion), family factor (observation of parental physical abuse), entertainment factor (preference of cartoon type), and friendship factor (number of close friends).

Outcome variable

The outcome of interest was bullying behaviour which was identified as a dichotomous variable; 'not bullied others' or 'bullied others'.

Conceptual framework

The conceptual framework in this study, depicted as a path diagram in Figure 2.2, is used to summarise the variables considered in the study and their roles.

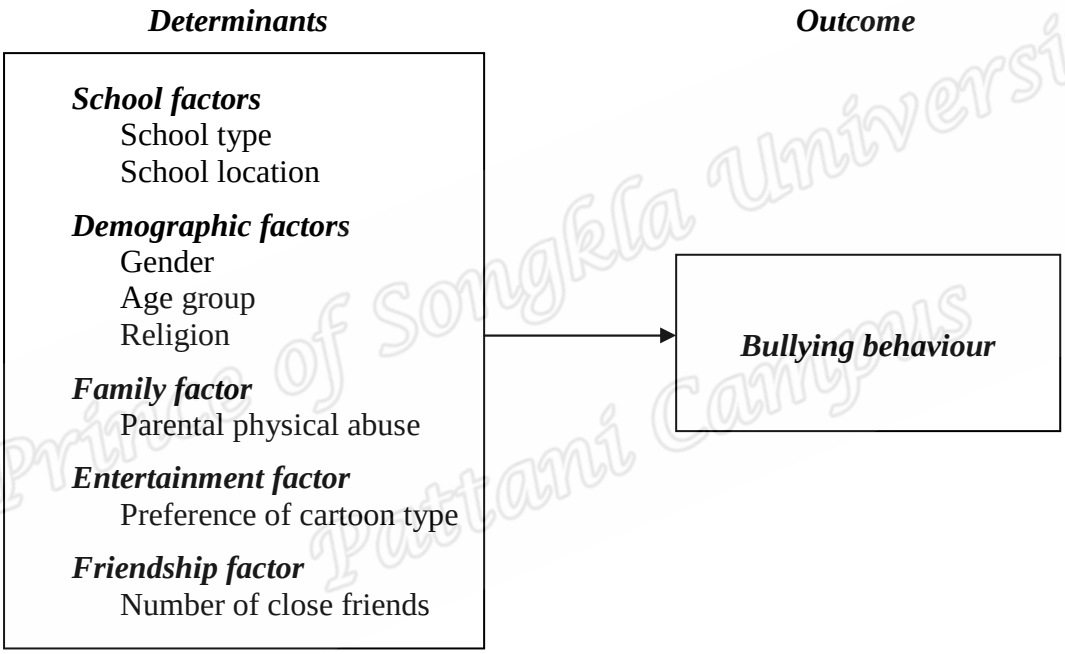


Figure 2.2: Conceptual framework showing variables in the study

2.3 Data collection and management

Data collection instrument

A questionnaire for primary data collection was adopted from Besag (1992), which comprised two sections.

Section 1: The respondent's background including school name, school location, school type, grade, gender, religion, age, weight, height, had seen parents' physical

abuse, most preferred cartoon type, number of close friends, career ambitions, method of travelling to school, willing at school, favourite zone, and unsafe zone.

Section 2: The bullying experiences including harmed anyone physically (ever been, and what way), hurt anyone verbally/mentally (ever been, and what way), circumstance of bully, gender of bullied student, age of bullied student, location of bullying, time of bullying and reason for bullying.

Questionnaire design

The questionnaire was modified to evaluate the severity of bullying as follows.

1. A literature review was undertaken and definitions were established.

2. The questionnaire was modified from Besag (1992), which used binary choice questions, for example 'Have you bullied anyone this term? (yes, no)' 'In what ways? (not at all, hit and kicked, teased, things taken from them', Appendix II). To appropriately modify for Thai students in primary school and to minimize the recall bias, this study asked the common behaviour of Thai students and used a binary type format (Appendix I).

3. A pilot study, involving 20 students from the demonstration school, Prince of Songkla University, Pattani campus and 20 students of Thesabans Ban Paknum, Saiburi, Pattani, was undertaken in order to improve the clarity of the questions and improve efficiency of data collection. The discrimination power ranged from 0.176 to 0.549 and reliability coefficient of internal consistency shows Kuder-Richardson20 of 0.876 (Appendix I).

Data collection

Verbal consent to participate in the study was obtained from students after assurance of confidentiality was given to individuals and group administered. The collection assistants were teachers in target schools, who volunteered to participate and were studying for a Graduate Diploma in Professional Teaching at Yala Islamic University. These teachers were trained in the interviewing techniques and the details of the questionnaire. They were asked to take care not to rush through the questionnaire and also to record responses accurately.

The teachers interviewed students in the classroom after permission was granted by the school principal. Each individual was interviewed with grades 1 to 3 students. Interviewed lasted approximately 20 to 30 minutes. Group administered (narrated) surveys of grades 4 to 6 students took approximately 40 to 60 minutes. With older students, the interviewer read the instructions to them and then allowed the student to write their own responses. Most of these responses were uncomplicated and involved just ticking a box.

Data management

The data were analyzed using Webstat (a set of programs for graphical and statistical analysis of data stored in an SQL database, written in HTML and VBScript), and R program (R Development Core Team, 2008).

2.4 Statistical methods

Factor analysis

Factor analysis is a data reduction technique. It is a group of procedures designed for removing duplicated information from a set of correlated variables and representing the variables with a smaller set of derived variables or factors. There are three procedures involved. The first stage involves obtaining the original data matrix. A set of subjects O_1, $O_{2,\ldots,}O_n$ are measured with a different number of variables V_1, V_2,…,V_k. The second stage involves the creation of a correlation matrix, which is calculated for each combination of two variables: V_1 with V_2, V_1 with V_3, etc., according to the following formula: If x_i is the observation from subject i on V_1 and y_i is the observation from subject i on V_2, then the correlation, r, between V_1 and V_2 is given by

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{(n-1)s_x s_y} \qquad (2.2)$$

where s_x and s_y are the sample standard deviations of V_1 and V_2, and n is the number of pairs of observations.

The last stage involves computing the factor loadings. These reveal the extent to which each of the variables contribute to the meaning of each of the factors. Within any one column of the factor matrix, some of the loadings will be high and some will be low. The variables with a high loading on a factor will be the ones that provide the meaning of the factor (Manly, 2000).

There are many ways to determine the number of factors. Educational research often use methods based on eigenvalues; eigenvalues less than one indicate that this factor

contributes less than the original variable and therefore should not be retained (Hair et al, 1998), the most statistically valid method is based on maximum likelihood estimation of the coefficients in the factor analysis decomposition.

Maximum likelihood factor analysis is a widely used method. This method enables a goodness of fit test to be conducted of a solution comprising *k* factors. It provides a test of the null hypothesis that *k* common factors are sufficient to describe the data. The algorithms for this method are given as follows.

Suppose there are *p* variables and we want to fit *k* factors. Let *R* be the *p* x *p* correlation matrix of the variables, *L* the *p* x *k* matrix of factor loadings, and ψ the vector of length *p* containing the unique variables. Then we need to find values for *L* and ψ that maximize the likelihood function, $f(1, \psi)$.

For the fixed value of ψ, we maximize (L, ψ) with respect to *L*. The value of *L* is then substituted into $f(1, \psi)$. Now *f* can be reviewed as a function of ψ. A transformation of this function gives

$$m(\psi) = \sum_{m=k+1}^{p} \left[\log \gamma_m + \frac{1}{\gamma_m} - 1 \right] \qquad (2.3)$$

where $\gamma_1 \leq \gamma_2 \ldots \leq \gamma_p$ are the eigenvalues of $\psi R^{-1} \psi$. We then minimize $m(\psi)$. This gives an estimate of ψ, which is then put into the likelihood $f(L, \psi)$. The likelihood is again maximized with respect to *L*. A new value for $m(\psi)$ is computed. This process is iterated until convergence is achieved.

After making the decision on how many factors to extract from the original set of variables we can redefine the factors so that the explained variance is redistributed

among the factors. This technique is used to sharpen the distinction in the meaning of the factors. A redefinition of the factors, with the loading on the various factors either very high or very low, and then eliminating as many medium sized loading, aids in the interpretation of factors.

Varimax rotation is one of many types of rotation that is often used and is regarded as the standard approach. This approach places more emphasis on the simplification of the factors. It tends to avoid a general factor. Using the comprehensibility method to select a number of factors; for example, analysis the subset data of bullying of Saiburi district, which consisting of 9 variables and 720 subjects, as follows.

1. The original data is listed in Table 2.1.

2. A correlation matrix is listed in Table 2.2.

	V1	V2	V3	V4	V5	V6	V7	V8	V9
O1	1	1	0	1	0	0	0	1	1
O2	1	1	1	0	0	0	0	1	0
O3	0	1	1	0	1	0	1	1	1
O4	0	0	0	0	0	0	0	1	0
O5	0	0	0	1	1	1	1	1	1
O6	0	1	0	1	0	1	0	1	0
O7	1	0	0	0	1	0	0	0	0
.									
.									
.									
O720	1	1	1	0	1	1	0	1	1

Table 2.1: Original data

Variable	V_1	V_2	V_3	V_4	V_5	V_6	V_7	V_8	V_9
V_1	1	0.50	0.27	0.03	0.12	0.07	-0.10	0.21	0.33
V_2		1	0.29	0.05	0.12	0.13	-0.11	0.21	0.40
V_3			1	0.05	0.15	0.21	0.10	0.16	0.43
V_4				1	0.18	0.28	0.30	0.21	0.25
V_5					1	0.30	0.24	0.31	0.31
V_6						1	0.42	0.18	0.47
V_7							1	0.19	0.26
V_8								1	0.30
V_9									1

Table 2.2: Correlation matrix

3. The factor loadings of maximum likelihood estimation of both varimax and none rotation are listed in Table 2.3. A two factor structure emerged. Before rotation, some variables show similar loading score in two factors structure; for example, variable 1 (V_1) has a loading score of 0.47 on factor 1 (F_1) and -0.47 on factor 2 (F_2), and variable 2 (V_2) has a loading score of 0.54 on factor 1 (F_1) and -0.50 on factor 2 (F_2). It's a maximum amount of variance but rarely provide a structure with conceptual meaning. Interpretability is enhanced by rotating so that clusters of variables are distinctly associated with a factor. The rotation has been done in such a way that variables 1, 2, 4, 5, 6, and 7 have high loading score on factor 1 (F_1) and low loading score on factor 2 (F_2), and the reverse is true for variable 3, 8 and 9.

None rotation		
Variable	F_1	F_2
V_1	0.47	-0.47
V_2	0.54	-0.50
V_3	0.48	-0.11
V_4	0.33	0.28
V_5	0.41	0.19
V_6	0.57	0.38
V_7	0.34	0.59
V_8	0.41	
V_9	0.77	

Varimax rotation		
Variable	F_1	F_2
V_1	0.67	
V_2	0.73	
V_3	0.25	0.43
V_4	0.43	
V_5	0.42	0.18
V_6	0.66	0.16
V_7	0.66	-0.15
V_8	0.13	0.38
V_9	0.15	0.55

Table 2.3: Factor loading of varimax and none rotation

Chi-squared decomposition and odds ratio

Pearson's chi-squared test and 95 % confidence interval for odds ratio are used to assess the associations between the determinant variables and the outcome of this study. The formulas based on contingency tables (McNeil, 1998b) are as follows (*X* is a determinant of interest, *Y* is the outcome).

A. 2 × 2 table

X is the determinant and *Y* is the outcome. The odds ratio is a measure of the strength of an association between two binary variables, i.e., both the outcome and the determinant are dichotomous (McNeil, 1998a, 1998b). For example: *X* is gender (coded as female and male) and *Y* is bullying other (coded as not bullied and bullied). To illustrate the definition of the odds ratio, a two-by-two table is constructed as follows.

Gender	Bullying behaviour		Total
	Not bullied	Bullied	
Female	*a*	*b*	*a*+*b*
Male	*c*	*d*	*c*+*d*
Total	*a*+*c*	*b*+*d*	*n* = *a*+*b*+*c*+*d*

The estimate the odds ratio is

$$OR = \frac{a \times d}{b \times c} \tag{2.4}$$

One method of testing the null hypothesis of no association between the determinant and the outcome is to use the z-statistic $z = \ln(OR) / SE$, where *SE* is the standard error of the natural logarithm of the odds ratio (McNeil, 1996). An asymptotic formula for this standard error is given by

$$SE(\ln OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{2.5}$$

A 95% confidence interval for the population odds ratio is thus

$$OR \times \exp\left(\pm 1.96\, SE\, [\ln OR]\right) \tag{2.6}$$

Pearson's chi-square statistic is defined as

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)} \tag{2.7}$$

The p-value is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic.

B. 3 × 2 tables

In this study, some of variables are three categorical. We use *3 × 2* tables to compare them. For example: *X* is age group (coded as 8 yrs or less, 9-10 yrs, and 11 yrs or more) and *Y* is bullying other (coded as not bullied and bullied).

Age group	Bullying behaviour		Total
	Not bullied	Bullied	
8 yrs or less	a_{11}	a_{12}	$a_{11}+a_{12}$
9-10 yrs	a_{21}	a_{22}	$a_{21}+a_{22}$
11 yrs or more	a_{31}	a_{32}	$a_{31}+a_{32}$
Total	$a_{11}+a_{21}+a_{31}$	$a_{12}+a_{22}+a_{32}$	$n = a_{11}+a_{21}+a_{31}+a_{12}+a_{22}+a_{32}$

The estimate of the odds ratio associated with 3 categories and 2 categories is obtained by collapsing the table into a two-by-two table with pivotal cell a_{ij}, that is,

$$OR_{ij} = \frac{a_{ij} d_{ij}}{b_{ij} c_{ij}} \tag{2.8}$$

where $b_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij}$, $c_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij}$, $d_{ij} = n - a_{ij} - b_{ij} - c_{ij}$, $n = \sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij}$.

The standard error of the natural logarithm of the odds ratio is given by the same formula as for the two-by-two table. In general, the association is comprised of *3×2* odds ratios, but only (3−1) (2−1) of them are independent.

Since the odds ratio in this case is obtained from a two-by-two table, Equation (2.5) gives the standard error, that is,

$$SE(\ln OR_{ij}) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}} \tag{2.9}$$

and an asymptotically valid 95% confidence interval is given by Formula (2.6).

Pearson's chi-squared statistic for independence (i.e., no association) is defined as

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}} \tag{2.10}$$

where $\hat{a}_{ij}$ is the expected value of a_{ij} assuming the null hypothesis of independence is true, that is

$$\hat{a}_{ij} = \frac{1}{n}\sum_{k=1}^{c} a_{ik} \sum_{l=1}^{r} a_{lj} \tag{2.11}$$

When the null hypothesis of the independence is true, the right-hand side of Equation (2.10) has a chi-squared distribution with (3–1) (2–1) degree of freedom.

Logistic regression

Multiple logistic regression analysis is used for modelling the association between several determinant variables and bullying behaviour. Logistic regression is a method of analysis that gives a particularly simple presentation for the logarithm of the odds ratio describing the association of a binary outcome with factors, and when fitted to data involving a dichotomous outcome and multiple determinants, it automatically provides estimates of odds ratios and confidence intervals for specific combinations of the risk factor (McNeil, 1996). For a set of predictor variables $x_1, x_2, ..., x_p$ and a

binary outcome *Y* the logistic regression model takes the following form:

$$\ln(\frac{p}{1-p}) = \alpha + \sum_{i=1}^{p} \beta_i x_i \qquad (2.12)$$

where *p* denotes the probability of occurrence of the specified outcome; in this studies the outcome (*Y*) is bullying other (coded as '0 = not bullied' and '1 = had bullied'). The probability of the 'had bullied' category (*Y* = 1) can be expressed as

$$P[Y=1] = \frac{\exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\alpha + \sum_{i=1}^{p} \beta_i x_i)} \qquad (2.13)$$

Using the logistic regression model for the data arising from a two-by-two table, we suppose x_i = 1 or 0, that is, the values of determinant *X* are taken to be 1 (exposure) and 0 (no exposure). Thus the logistic regression model can be written as

$$\ln\{\frac{P(Y=1/X=1)}{1-P(Y=1/X=1)}\} = \alpha + \beta \qquad (2.14)$$

$$\ln\{\frac{P(Y=1/X=0)}{1-P(Y=1/X=0)}\} = \alpha \qquad (2.15)$$

The equations (2.14) and (2.15) are the (natural) logarithms of the odds for the outcome given the exposure (*x* = 1) and non-exposure (*x* = 0), respectively. After exponentiation each equation, the odds for the exposed and non-exposed groups can be written as $\exp(\alpha + \beta)$ and $\exp(\alpha)$, respectively. The odds ratio is therefore obtained from the simple formula

$$OR = \frac{\exp(\alpha + \beta)}{\exp(\alpha)} = \exp(\beta) \qquad (2.16)$$