Appendix 4

Quality of the data: an analysis of inter-observer reliability

This appendix describes work whose purpose is to test what degree of confidence can be placed in the data presented in this dissertation, and if possible to put a figure on that degree of confidence. Information about the methodology used in the study, the hours of observation and interviews, the approach to sampling, the rationale for the use of community members for data collection, and the training of observers, can be found in Chapter 2. This appendix focuses on a comparison of the data collected by different observers.

1. Rationale

The main question addressing the issue of the quality of the data collected was whether it might contain any systematic error arising from non-random differences between observers. Observation sessions were chosen and

matched to observers at random[1], and all observers worked at all site types. Training in procedures was designed to reduce as much as possible the variability between different observers and interviewers. However, it was expected that observers would vary, in their ability to concentrate, or to hear many different languages, or in their stamina, or in their ability to record many details simultaneously.

Most data were collected by individual unsupervised observers. However, some observations were monitored by a research assistant (RA) who independently recorded data at the same event, and sometimes more than one observer was present, in which case all observers independently recorded data. It is this data from multiply-recorded observations and interviews that forms the basis for the investigation into observer variability. The investigation is thus focused entirely on the consistency of data recorded by different observers. It does not examine the reliability of observers against any benchmark, for example one provided by the researcher's own observations; nor does it examine the reliability of individual observers across time, for example comparing an observer's work at the beginning of the data collection period with work at the end.

Inter-rater reliability is discussed by McNamara (1996) in the context of assessing second language performance, where raters make subjective judgements in attempting to match aspects of performance against a criterion scale. 'Judgements that are worthwhile will inevitably be complex and involve acts of interpretation on the part of the rater, and thus be subject to disagreement' (*op cit*: 117). Decisions made by observers in the current study were designed to be reduced to the mechanical, yet there remain some

1 There is some evidence of systematic error in the allocation of sites to observers, with some observers apparently swapping observation assignments in order to get assignments nearer their homes in a widely dispersed refugee camp.

parallels between observers 'mechanically' recording language behaviour and raters assessing language performance. Just as raters systematically vary in their rating of particular aspects of language performance, being lenient in some respects and severe in others (*op cit*: 122ff), so it was expected that observers might systematically over-record, or under-record, the use of particular languages, perhaps for reasons unconscious to them but which had the potential to produce a constant distortion in their work. Multi-faceted Rasch analysis (*op cit*) is an iterative procedure designed to detect and correct for systematic variation in multiple aspects of language assessment, and is a tool that could be used in the present investigation. The researcher, however, judged it to be beyond the scope of this dissertation and instead adopted a simpler, static approach.

The approach to inter-observer reliability taken here was as follows: if an independent researcher R were to be given the data which each pair of observers O_i and O_j recorded at their jointly observed events, would R conclude that O_i and O_j were observing the same events? and if so, what degree of statistical significance could R attach to such a claim?

For observations based on time slices, the procedure adopted was as follows. For the data from each observer pair O_{ij}, Pearson product moment correlation coefficients r_{ij} were calculated for each category of observation. Categories varied according to the type of observation, but they could all be brought under four generic headings, *Leaders' oral output, Participants' oral output, Leaders' writing and written materials* and *Overheard and overseen*[2]; and within

2 Oral output includes any singing or chanting; and 'overseen' includes individual writing by students. Although some of this was public official writing by students on the classroom blackboard, it is believed that most was official, individual writing by students in notebooks. Since only a small fraction, if any, of this is actually seen by an observer, and that which is seen is written by students located close to the observer, it has been called here 'overseen' and regarded as the same as overheard asides and other 'unofficial' fragments captured by the observer. Although it is confusing to conflate the two types of

each of these were the same 6 groups of languages – *Karenni*, *Burmese*, *English*, *Karen*, *Shan* and *Other*.

Treating each category separately gives a possible maximum number of correlations for each observer pair O_{ij} of 24. Call these $\{r_{ij}\}$. There are 5 observers plus one of the RAs who conducted the monitoring and participated in other comparison exercises, giving a total of 15 observer-pair comparison sets $\{r_{ij}\}_{1,\ldots,15}$.

The correlations were assembled into tables, one for *total language occurrences* recorded in the data sheets, and another for *primary language occurrences*. Total language occurrences are all of an observer's marks for a particular language L in a particular category. Primary language occurrences are a subset of an observer's marks for L for that category, and include marks that indicate that L was the only language used during a particular time-slice, or that, if other languages were used, L was underlined by the observer as being the one most used during that time-slice. This 'double marking' was done because it was felt that languages used in secondary or supporting roles might be overreported on the forms unless there was an attempt to distinguish most used from other languages used during a time-slice.

Each table of inter-observer correlations contains one row for each observer pair, and along the top of the table are the different languages under the different observation categories. There are 15 observer-pair rows in the table, and in each row there are 4 observation categories each containing 6 languages, giving a total of 15*24 = 360 cells. Some cells were excluded from consideration or set to 0 or 1 for technical reasons, as described in the next section. For example, if O_i made no marks in a particular category but O_j did,

writing, one public and official and visible, the other private and official but mostly unseen, this is not thought to be seriously distorting, as most writing in class by students was of the second kind.

r_{ij} is undefined. If the number of marks made by O_j was small, the difference between O_i and O_j was regarded as random variation or 'noise' and r_{ij} disregarded, but if the number of marks made by O_j was large, r_{ij} was retained and set to 0. 23 of the 360 cells were affected in this way. Finally, each r_{ij} was examined for significance, at both the 5% level and the 1% level, and this examination provided the basis for a conclusion about the degree of reliability of the observers.

The approach for observations at shops and clinic exit interviews was the same, with procedures modified to take account of the slightly different data categories.

2. Data available for investigating inter-observer reliability

Of 404 hours of observation[3], about 33 hours were monitored or otherwise multiply observed, and of 299 health clinic exit interviews accepted for analysis, 29 were recorded by more than one observer as part of a comparison exercise. All observers and one RA were involved in this multiply-recorded data.

The number of time-slices multiply observed varied from pair to pair; the minimum was 59 slices of 3 minutes each, or about 3 hours, and the maximum was 351 slices, or 17.5 hours; the data came from schools, meetings and acts of worship. Multiple data from shops was less, allowing only a comparison between each observer and the monitoring RA, amounting to 30 minutes each, with a mean of 9 shop transactions for each observer. At clinics, a comparison exercise, in which two teams of 3 people parallel-recorded a total of 29 interviews, was the basis of a small exercise measuring

3 a figure that includes monitoring and other multiply-observed events. See Chapter 2, Table 8.

the consistency of members of the two teams against other members of the same team[4]. Although the comparison data from shops and clinics was meagre, the observers repeatedly described time-slice observations as much more demanding than observing in shops or conducting interviews, and it was therefore possible to argue that if consistency could be demonstrated in the case of time-slice observations, it was likely to hold in the other cases too. Tables 9 and 10 show details of the comparison data collected at different sites, and the numbers of time slices available for comparing each observer pair O_{ij}.

Table 9: Multiply-recorded data used for analysis of inter-observer reliability

site/event	*details of multiple recording*	*quantity of data collected*
3 meetings 8 school periods 2 acts of worship	observer monitored by RA for 30 mins	146 slices* doubly recorded
7 shops	observer monitored by RA for 30 mins	45 transactions doubly recorded
8 meetings	simultaneously observed by 2 observers; in one case by 3 observers	403 slices doubly recorded and a further 43 slices triply recorded
4 school periods	comparison exercise; about 3 hrs simultaneously observed by 5 obs and 1 RA	59 slices quintuply recorded
2 health clinics	comparison exercise; teams of 3; team members independently recorded interview answers for 3 hrs at each site	15 interviews triply recorded at one site, 14 triply recorded at the other

*1 slice = 3 mins

4 but not against members of the other team.

Table 10: Time slices* available for comparisons between observers

observer	*b*	*c*	*d*	*e*	*f*
a	93	113	104	82	83
b		98	98	59	59
c			351	114	59
d				59	59
e					114

*1 slice = 3 mins

3. Time slices and chunks

The approach taken to investigating inter-observer reliability was to compare the detailed observation records for each observer pair O_{ij} at simultaneously observed events. However, the comparison was not done at the level of time-slices, because it was expected that there would be some difficulty in mapping O_i's record against O_j's. Each observer had his own clock, and there was some evidence that observers´ clocks may not have been perfectly synchronised when doing simultaneous observations; there is also some evidence that a few slices were not 3 minutes, but 2, or 4. While in the broadest picture these differences do not matter, it does follow that in any one case a slice-by-slice comparison of two observers may not compare like with like. In some cases, sliding two observation records against each other, forwards or backwards by 1 slice, has produced a better fit. However, in 99 sample comparisons, each of about 45 minutes length, fit could be improved by sliding in only half the cases, and the overall gain in fit for the entire sample, searching for best fit in each case, was less than 1%[5].

A 3-minute time slice is rather like a pixel; just as two screens presenting identical pictures may not require individual pixels to correspond exactly, so

5 As measured by the mean of the differences, between pairs of observers, in the proportions of use assigned to different languages. A separate mean was calculated for each language in each observation category.

good fit between observers may involve a degree of variance at the time slice level. The approach taken here has been to consolidate slices into larger units in order to smooth out some of the random low-level variability. The *chunk* has been chosen for this, where 1 chunk equals 5 slices. Thus a chunk corresponds to 15 minutes of an event.

A chunk size of 5 slices has been chosen because the number of chunks will then be large enough to permit statistical analysis of chunks; at the same time each chunk, containing 5 time slices, will contain between 0 and 5 recorded occurrences of any language in any observation category - in other words the data will be interval data. Tables 11 and 12 contain comparison data for two observers O_i and O_j across 110 slices of common observation, equal to 5.5 hours, first as slices (Table 11), then chunks (Table 12).

Table 11: Examples of observer comparison data as time slices

slice	*Oi*	*Oj*	*slice*	*Oi*	*Oj*	*slice*	*Oi*	*Oj*	*slice*	*Oi*	*Oj*	*slice*	*Oi*	*Oj*	*slice*	*Oi*	*Oj*
1	1	1	22	1	1	43	1	1	64	1	1	85	1	1	106	1	1
2	0	0	23	0	1	44	1	1	65	0	0	86	1	1	107	1	1
3	0	0	24	1	0	45	1	1	66	0	0	87	0	0	108	1	1
4	0	0	25	1	1	46	0	1	67	0	0	88	0	0	109	0	0
5	1	0	26	1	1	47	1	1	68	0	0	89	0	0	110	0	1
6	0	0	27	1	1	48	1	1	69	0	0	90	1	1			
7	1	1	28	1	1	49	0	1	70	1	1	91	1	1			
8	0	0	29	1	1	50	0	1	71	0	0	92	1	1			
9	0	0	30	1	1	51	0	1	72	1	1	93	1	1			
10	0	0	31	1	1	52	1	0	73	1	1	94	1	1			
11	0	0	32	0	1	53	0	1	74	1	1	95	0	0			
12	0	1	33	1	0	54	0	0	75	1	1	96	0	0			
13	0	0	34	1	1	55	0	0	76	1	1	97	1	1			
14	0	0	35	0	1	56	0	1	77	0	0	98	0	0			
15	1	1	36	1	1	57	0	1	78	0	0	99	0	0			
16	1	1	37	1	1	58	1	1	79	1	1	100	1	1			
17	0	1	38	1	1	59	1	1	80	1	1	101	1	1			
18	1	0	39	1	1	60	0	0	81	0	0	102	1	1			
19	1	1	40	1	1	61	1	1	82	1	1	103	1	1			
20	1	1	41	1	1	62	1	1	83	1	1	104	1	1			
21	1	1	42	1	1	63	1	1	84	1	1	105	1	1			

Example comparison between two observers *Oi* and *Oj* across 110 slices of common observations. Each pair of numbers represents one time slice for one language in one observation category for observer *Oi* (left) and observer *Oj* (right), in this case leader's oral output in Burmese. A '1' indicates the presence of *Burmese* in the leader's oral output column of the observation form.

Source: comparison data for Lee Reh and Neh Law

Table 12: Example of observer comparison data as chunks

chunk	*i*	*j*
1	2	1
2	1	1
3	1	2
4	4	4
5	4	4
6	5	5
7	3	4
8	5	5
9	5	5
10	2	5
11	1	2
12	2	4
13	4	4
14	1	1
15	4	4
16	3	3
17	4	4
18	2	2
19	4	4
20	2	2
21	5	5
22	3	4
total marks	*67*	*75*

Comparison data from Table 11 represented as 22 chunks, where each chunk = 5 time slices of 3 minutes each. Accumulating the 1s and 0s of time-slices into chunks smooths random variations and permits the use of correlation coefficients. Pearson's r for the two sets of scores is 0.821.

Source: comparison data for Lee Reh and Neh Law.

An ordered list of chunks containing data for one language in one observation category for one pair of observers, as shown Tables 11 and 12, is called here a *comparison dataset*.

Comparisons were made between each of 6 observers (5 observers plus the RA who carried out monitoring observations). All available time slices simultaneously observed by each pair O_{ij} were concatenated into a single continuous ´ event´, which was converted into chunks, with any residual slices at the end truncated to produce a data set each of whose chunks contains exactly five slices.

The number of chunks available for comparison between observer pairs varied from 11 to 70, and can be calculated for any pair by dividing the figures given in Table 10 by 5 and truncating the result.

4. Critical and non-critical observation categories

The first two observation categories, *Leaders' oral output* and *Participants' oral output* were designated 'critical', since these recorded the public speech of people at events, and it was expected that there would be little variation between observers. On the other hand, there was some doubt about how much concordance could be expected in the third and fourth categories, *Leaders' writing and written materials* and *Overheard and overseen*. Concerning the latter, it was always expected that observers placed differently at events would access rather different overheard or overseen language use[6]. Concerning the third category, some doubts arose in observers' minds about whether marks should record only acts of writing or also include the continued presence of previously written materials, for example on a whiteboard for perhaps half an hour after having been written up. Some observers modified the agreed recording scheme, introducing ditto marks to indicate the continued presence of written materials; others did not, and simply marked the points at which materials were written or first presented. By the time the issue came up for discussion it was too late in the data collection phase to introduce a correction to procedures.

5. Small and empty comparison datasets

In some cases comparison datasets are empty or nearly so. If a teacher did no

6 see footnote 2.

writing at all in a teaching period, the section of the observer form called *Teacher's writing* remains empty. In teaching periods observed by two observers, if neither O_i nor O_j has made marks on their forms for any language in that category, the comparison datasets for that category are empty. In this case Pearson's r is undefined as its calculation involves a division by zero, but intuitively the empty datasets correspond to a concordance between O_i and O_j that there was no teacher writing in any language for that period; in such cases r has been set to 1 to correspond to intuition. All these 1s contribute to, but do not by themselves determine, the degree of concordance between O_i and O_j, and the significance that can be attached to it.

In other cases one observer O_i has made no marks at all while O_j has made a few marks. In this case the chunks for O_i contain only zeros, and again Pearson's correlation is undefined. The way such comparison datasets have been dealt with in the analysis depends on the total number of marks they contain. If it is less than 5, for O_i and O_j combined, the dataset has been excluded from the analysis as being 'noise', or beneath a floor level. However, if the total number of marks is 5 or greater, r has been set to zero and it has been retained in the analysis. The concept of a floor level for the data is explored a little further below.

In yet other cases, both O_i and O_j have made marks for a particular category, and in these cases r is defined and $0<r<1$. As with the small undefined datasets described above, if the total number of marks is less than 5, these datasets also have been excluded from the analysis as corresponding to 'noise'.

Considering all non-empty comparison datasets across all observers, the mean number of marks in each dataset, both observers combined, is 55 for

total language occurrences (SD 29.4) and 47 for primary language occurrences (SD 21.2). The choice of 5 as the 'floor level' for inclusion of data should be considered in the light of this mean number of marks. The floor level chosen is important, because setting a dataset's undefined r to 0 or 1 as a result of a single mark by an observer is likely to have a great effect on the question of whether the correlations between observer pairs reaches significance. On the other hand, leaving r for such datasets undefined leaves open the question of what such datasets mean. To exclude from the analysis correlations from small datasets, whether their correlations are defined or undefined, protects the analysis from small random differences between observers whose effects may be unreasonably magnified when included.

To put the numbers of small datasets into the context of the comparison data as a whole, the table of correlations between all observer pairs, for total language occurrences, contains 360 possible correlations between 15 observer-pairs across 4 observation categories each containing the same 6 languages[7]. Of these 360 possible correlations, the following have been included in the analysis:

- 186 empty cells corresponding to the languages in categories where neither O_i not O_j entered data; these have undefined r because both sides are zero but have been set to 1 for the analysis and are blank in the tables reproduced below;

- 151 correlations with $0<r<1$; these correspond to languages in categories where both O_i and O_j entered data and the total number of marks in the comparison dataset was greater than 5;

- two datasets which have more than 5 marks but for which r is undefined because one side is zero, and for which r has been set to 0.

7 ie 15*4*6.

The following correlations have been excluded from the analysis:

- 23 small datasets containing fewer than 5 marks, of which 14 have undefined r, 7 have r=1 and 2 have 0<r<1.

Examples of small or empty comparison datasets are shown in Table 13.

Table 13: Examples of small or empty comparison datasets

observers	*Oi*	*Oj*	*Oi*	*Oj*	*Oi*	*Oj*
chunks	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	1	1	0	0	0	0
	1	1	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	0
	0	0	0	0	0	2
	0	0	0	0	0	0
	0	0	0	0	0	1
	0	0	0	0	0	1
					0	0
					0	0
					0	0
total marks:	***2***	***2***	***0***	***0***	***0***	***4***
correlation:	**r = 1**		**r undefined**		**r underfined**	

Small dataset showing concordance between observers Oi and Oj (left), empty dataset showing concordance but with r undefined (middle), and small dataset with one side empty and r undefined (right).

6. The correlations

The correlations obtained are shown in Tables 14-17.

Table 14 contains correlations for total language occurrences recorded by pairs of observers at all events they co-observed. Also shown are small excluded datasets. The columns on the left show the number of chunks for each observer-pair correlation, and the degrees of freedom df (= chunks *minus* 2). The critical correlation is shown for each df at the 5% level of

significance, and those correlations in the table that fail to be significant at this level are highlighted in grey.

It can be seen that in the critical categories - Leaders´ oral output, Participants ´ oral output - all correlations are significant at the 5% level, but that in the other categories - Leaders´ public writing & materials, Overseen and overheard - 10 of the correlations fail to be significant. An independent researcher, therefore, asking the question[8] *Are pairs of observers recording the same events or not?* would conclude that in some cases the correlations are strong enough at the 5% level to reject the null hypothesis, but that in others they are not; taken together, therefore, the correlations are not strong enough and the null hypothesis cannot be rejected. If, however, the observation categories can be treated independently, and some arguments have been presented above[9] why this is so, a researcher would conclude that correlations in the critical categories were significant at the 5% level, but those in the non-critical categories were not.

Table 15 contains correlations for primary language occurrences recorded by pairs of observers, showing significance at the 5% level. The picture is similar to that for total language occurrences, and the conclusion is the same.

Tables 16 and 17 show the same correlations as Tables 14 and 15, but with significance shown at the 1% level. It can be seen that correlations in neither the critical nor the non-critical observation categories reach significance at this level.

8 see section *Rationale*, above.
9 see section *Critical and non-critical categories*, above.

Table 14: Correlations between total language occurrences recorded by pairs of observers Oi, Oj co-observing the same events† 5% significance level

					Leaders´ oral output				Participants´ oral output						Leaders´ public writing				Overseen and overheard					
chunks	**df**	**critical r†**	**Oi**	**Oj**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
70	68	0.2500	a	f	0.979	0.946	S&=1		0.858	0.979				S&U	0.356	0.824	0.778		0.687	0.428	1.000	0.307	**0.056**	0.437
22	20	0.4227	c	d	0.962	0.821	S&=1		0.883	0.954						0.995	0.994		0.801	0.696	0.963			0.756
22	20	0.4227	a	c	0.957	0.950	S&=1		0.842	0.907						0.996	0.994		0.845	0.525	0.963	**0**	S&U	**0**
22	20	0.4227	e	a	0.876	0.777	0.885		0.842	0.685					**0.333**	0.961	0.788		0.846	0.881	0.620		S&=1	S
22	20	0.4227	f	e	0.965	0.845	S&U	0.983	0.767	0.957		1.000			0.987	1.000	0.993	1.000	0.955	0.963	0.920		S	S&U
19	17	0.4555	b	f	0.735	0.879	S&U		0.882	0.924					S&U	0.711	0.491		0.822	0.476	0.919			0.792
19	17	0.4555	b	a	0.857	0.815	S&U		0.939	0.898						0.707	0.491		0.911	**0.444**	0.919	S&U	S&U	**-0.031**
18	16	0.4683	d	e	0.981	0.896	S&U	1.000	0.891	0.844		1.000				0.995	0.991			0.994	0.919			
18	16	0.4683	b	e	0.917	0.951	1.000		0.786	0.942						0.711	0.643		0.777	0.647	0.985		S&U	0.686
16	14	0.4973	e	c	0.957	0.958	0.984		0.742	0.877	1.000		S&U		0.993	0.994	0.990		1.000	0.920	0.990			
11	9	0.6021	f	c	0.938	0.694	S&=1		0.693	0.742						1.000	1.000			1.000	1.000			
11	9	0.6021	b	c	0.841	0.749	S&U		0.768	0.689						0.663	**0.458**			**0.558**	0.914			
11	9	0.6021	b	d	0.728	0.869	S&U		0.709	0.833						0.660	**0.563**			**0.549**	0.991			
11	9	0.6021	a	d	0.938	0.914	S&=1		0.777	0.909						1.000	1.000			1.000	1.000			
11	9	0.6021	f	d	0.896	0.884	S&=1		0.698	0.819						0.993	0.993			0.990	0.960			

† monitored events, multiply-observed events and observer comparison exercise; 5% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.458	correlations which do not reach significance

Table 15: Correlations between primary* language occurrences recorded by pairs of observers Oi, Oj co-observing the same events† 5% significance level

					Leaders´ oral output				Participants´ oral output				Leaders´ public writing				Overseen and overheard					
chunks	df	critical r†	Oi	Oj	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
70	68	0.2500	a	f	0.823	0.936			0.834	0.993			**0**	0.837	1.000		0.579	0.498	1.000	S	S&U	0.341
22	20	0.4227	a	c	0.922	0.938			0.809	0.851				0.996	0.994		0.764	0.776	0.963	**0**		
22	20	0.4227	c	d	0.877	0.909			0.841	0.778				0.995	0.994		0.529	0.743	0.963			**0.206**
22	20	0.4227	e	a	0.900	0.787	S&U		0.757	0.856			S&U	0.962	0.992		1.000	0.948	0.620			
22	20	0.4227	f	e	0.768	0.974		0.983	0.793	0.938		1.000	1.000	1.000	0.992	1.000	0.957	0.992	0.920			
19	17	0.4555	b	a	0.767	0.794			0.841	0.987				0.707	0.491		0.855	0.501	0.919			
19	17	0.4555	b	f	0.693	0.826			0.892	0.987			S&U	0.711	0.491		0.625	0.465	0.919			S&U
18	16	0.4683	b	e	0.911	0.699			0.844	0.957			S&U	0.684	0.529		0.937	0.487	1.000		S&U	
18	16	0.4683	d	e	0.847	0.899		1.000	0.801	0.698		1.000		0.995	0.991			0.994	0.919			
16	14	0.4973	e	c	0.937	0.887	0.978		0.742	0.886	0.978		0.993	0.994	0.990		1.000	0.992	0.990			
11	9	0.6021	a	d	0.953	0.972			0.963	0.803				1.000	1.000			1.000	1.000			
11	9	0.6021	b	c	0.762	0.608			0.668	0.919				0.660	**0.563**			**0.549**	0.991			
11	9	0.6021	b	d	0.820	0.684			0.904	0.741				0.663	**0.458**			**0.558**	0.914			
11	9	0.6021	f	c	0.873	0.932			0.653	0.894				0.993	0.993			0.990	0.960			
11	9	0.6021	f	d	0.967	0.950			0.966	0.803				1.000	1.000			1.000	1.000			

* languages used alone in a particular category in a time slice, or used with others but marked as the most used

† monitored events, multiply observed events and observer comparison exercise; 5% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.563	correlations which do not reach significance

Table 16: Correlations between total language occurrences recorded by pairs of observers Oi, Oj co-observing the same events† 1% significance level

					Leaders´ oral output				Participants´ oral output						Leaders´ public writing				Overseen and overheard					
chunks	df	critical r†	Oi	Oj	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
70	68	0.3248	a	f	0.979	0.946	S&=1		0.858	0.979				S&U	0.356	0.824	0.778		0.687	0.428	1.000	**0.307**	**0.056**	0.437
22	20	0.5368	a	c	0.957	0.950	S&=1		0.842	0.907						0.996	0.994		0.845	**0.525**	0.963	**0**	S&U	**0**
22	20	0.5368	c	d	0.962	0.821	S&=1		0.883	0.954						0.995	0.994		0.801	0.696	0.963			0.756
22	20	0.5368	e	a	0.876	0.777	0.885		0.842	0.685					**0.333**	0.961	0.788		0.846	0.881	0.620		S&=1	S
22	20	0.5368	f	e	0.965	0.845	S&U	0.983	0.767	0.957		1.000			0.987	1.000	0.993	1.000	0.955	0.963	0.920		S	S&U
19	17	0.5751	b	a	0.735	0.879	S&U		0.882	0.924					S&U	0.711	**0.491**		0.822	**0.476**	0.919			0.792
19	17	0.5751	b	f	0.857	0.815	S&U		0.939	0.898						0.707	**0.491**		0.911	**0.444**	0.919	S&U	S&U	**-0.031**
18	16	0.5897	b	e	0.917	0.951	1.000		0.786	0.942						0.711	0.643		0.777	0.647	0.985		S&U	0.686
18	16	0.5897	d	e	0.981	0.896	S&U	1.000	0.891	0.844		1.000				0.995	0.991			0.994	0.919			
16	14	0.6226	e	c	0.957	0.958	0.984		0.742	0.877	1.000		S&U		0.993	0.994	0.990		1.000	0.920	0.990			
11	9	0.7348	a	d	0.938	0.914	S&=1		0.777	0.909						1.000	1.000			1.000	1.000			
11	9	0.7348	b	c	0.841	0.749	S&U		0.768	**0.689**						**0.663**	**0.458**			**0.558**	0.914			
11	9	0.7348	b	d	**0.728**	0.869	S&U		**0.709**	0.833						**0.660**	**0.563**			**0.549**	0.991			
11	9	0.7348	f	c	0.938	**0.694**	S&=1		**0.693**	0.742						1.000	1.000			1.000	1.000			
11	9	0.7348	f	d	0.896	0.884	S&=1		**0.698**	0.819						0.993	0.993			0.990	0.960			

† monitored events, multiply observed events and observer comparison exercise; 1% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.491	correlations which do not reach significance

Table 17: Correlations between primary* language occurrences recorded by pairs of observers Oi, Oj co-observing the same events† 1% significance level

					Leaders´ oral output				Participants´ oral output				Leaders´ public writing				Overseen and overheard					
chunks	df	critical r†	Oi	Oj	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
70	68	0.3248	a	f	0.823	0.936			0.834	0.993			**0**	0.837	1.000		0.579	0.498	1.000	S	S&U	0.341
22	20	0.5368	a	c	0.922	0.938			0.809	0.851				0.996	0.994		0.764	0.776	0.963	**0**		
22	20	0.5368	c	d	0.877	0.909			0.841	0.778				0.995	0.994		**0.529**	0.743	0.963			**0.206**
22	20	0.5368	e	a	0.900	0.787	S&U		0.757	0.856			S&U	0.962	0.992		1.000	0.948	0.620			
22	20	0.5368	f	e	0.768	0.974		0.983	0.793	0.938		1.000	1.000	1.000	0.992	1.000	0.957	0.992	0.920			
19	17	0.5751	b	a	0.693	0.826			0.892	0.987			S&U	0.711	**0.491**		0.625	**0.465**	0.919			S&U
19	17	0.5751	b	f	0.767	0.794			0.841	0.987				0.707	**0.491**		0.855	**0.501**	0.919			
18	16	0.5897	b	e	0.911	0.699			0.844	0.957			S&U	0.684	**0.529**		0.937	**0.487**	1.000		S&U	
18	16	0.5897	d	e	0.847	0.899		1.000	0.801	0.698		1.000		0.995	0.991			0.994	0.919			
16	14	0.6226	e	c	0.937	0.887	0.978		0.742	0.886	0.978		0.993	0.994	0.990		1.000	0.992	0.990			
11	9	0.7348	a	d	0.953	0.972			0.963	0.803				1.000	1.000			1.000	1.000			
11	9	0.7348	b	c	0.762	**0.608**			**0.668**	0.919				**0.660**	**0.563**			**0.549**	0.991			
11	9	0.7348	b	d	0.820	**0.684**			0.904	0.741				**0.663**	**0.458**			**0.558**	0.914			
11	9	0.7348	f	c	0.967	0.950			0.966	0.803				1.000	1.000			1.000	1.000			
11	9	0.7348	f	d	0.873	0.932			**0.653**	0.894				0.993	0.993			0.990	0.960			

* languages used alone in a particular category in a time slice, or used with others but marked as the most used

† monitored events, multiply observed events and observer comparison exercise; 1% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.684	correlations which do not reach significance

Non-significant correlations were investigated to see if they were distributed across all observers. It can be seen from Table 14 that 6 out of 10 correlations for total language occurrences that fail to reach significance at the 5% level are from comparisons involving a single observer, *b*. A similar concentration can be seen in the correlations for primary language occurrences in Table 15.

An investigation of observer *b´*s rate of recording marks compared with other observers, taken from the comparison exercise at which all observers parallel-observed 4 teaching periods, shows all observers except *b* clustered closely around the mean (see Table 18), with *b* showing very low recording rates for teacher writing and student writing - probably an indication of a general problem about how to interpret instructions about recording writing, as described earlier[10]. However, a similar comparison excluding writing indicates no problem with observer *b*, and neither does a comparison of marking rates across all school observations whether parallel-observed or not.

10 see section *Critical and non-critical categories*, above.

Table 18: Quantity of marks made by observers

Comparison exercise

marks made per 15-min chunk; total chunks observed = 12

Observer	Teacher speech	Student speech	Teacher writing	Student writing	total/chunk
a	5.14	3.92	2.64	1.69	13.38
b	5.14	3.58	1.22	0.88	10.81
c	5.27	4.46	2.77	1.76	14.26
d	5.81	5.34	2.64	1.69	15.47
e	5.41	4.66	2.57	1.62	14.26
f	5.34	4.19	2.64	1.69	13.85
mean	5.35	4.36	2.41	1.55	13.67

All school observations

marks made per 15-min chunk, teacher & student speech combined

	chunks obsvd	total/chunk
a	152	8.78
b	117	9.4
c	185	14.52
d	173	13.13
e	19*	9.89
f	135	8.5
mean		10.87

* research assistant´s school monitoring activities only

A similar, though smaller, cluster of non-significant correlations is associated with observer *f*, and an investigation of marking rates for *f* shows no obvious problems, although it is interesting to note that *b* and *f* did fewest observations in schools, and also show the lowest rates of marking across all school observations. Additionally, *b* observed school classes in only 3 different sections of the camp, compared to a mean for all observers of 6.6 camp sections and a maximum of 9 sections. The reasons for this were not investigated, but it is assumed that the system of randomisation of work locations broke down to some extent[11]. Finally, observer *b* showed some carelessness in filling in details of schools, teachers and subjects, so that, for example, 14 of 46 school periods observed by *b* contained no details of the

11 see Chapter 2, section *Approach to sampling.*

school, and 16 showed no subject details.

The effects of discarding observer *b´*s data are shown in Table 19, for total language occurrences, and Table 20, for primary language occurrences. It can be seen that, for total language occurrences, three quarters of the non-significant correlations are removed, plus more than half of the small undefined datasets. The picture for primary language occurrences is similar[12].

12 An identical exercise for observer *f* shows a similar though smaller improvement in overall data quality.

Table 19: The effect of discarding data from observer *b:* correlations between total language occurrences recorded by pairs of observers Oi, Oj co-observing the same events†
5% significance level

					Leaders´ oral output				Participants´ oral output						Leaders´ public writing				Overseen and overheard					
chunks	df	critical r†	Oi	Oj	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
Keep the following correlations																								
22	20	0.4227	a	c	0.957	0.950	S&=1		0.842	0.907						0.996	0.994		0.845	0.525	0.963	**0**	S&U	**0**
22	20	0.4227	c	d	0.962	0.821	S&=1		0.883	0.954						0.995	0.994		0.801	0.696	0.963			0.756
22	20	0.4227	e	a	0.876	0.777	0.885		0.842	0.685					**0.333**	0.961	0.788		0.846	0.881	0.620		S&=1	S
18	16	0.4683	d	e	0.981	0.896	S&U	1.000	0.891	0.844		1.000				0.995	0.991			0.994	0.919			
16	14	0.4973	e	c	0.957	0.958	0.984		0.742	0.877	1.000		S&U		0.993	0.994	0.990		1.000	0.920	0.990			
11	9	0.6021	a	d	0.938	0.914	S&=1		0.777	0.909						1.000	1.000			1.000	1.000			
22	20	0.4227	f	e	0.965	0.845	S&U	0.983	0.767	0.957		1.000			0.987	1.000	0.993	1.000	0.955	0.963	0.920		S	S&U
70	68	0.2500	a	f	0.979	0.946	S&=1		0.858	0.979				S&U	0.356	0.824	0.778		0.687	0.428	1.000	0.307	**0.056**	0.437
11	9	0.6021	f	d	0.938	0.694	S&=1		0.693	0.742						1.000	1.000			1.000	1.000			
11	9	0.6021	f	c	0.896	0.884	S&=1		0.698	0.819						0.993	0.993			0.990	0.960			
and discard the following:																								
19	17	0.4555	b	a	0.857	0.815	S&U		0.939	0.898						0.707	0.491		0.911	**0.444**	0.919	S&U	S&U	**-0.031**
19	17	0.4555	b	f	0.735	0.879	S&U		0.882	0.924					S&U	0.711	0.491		0.822	0.476	0.919			0.792
18	16	0.4683	b	e	0.917	0.951	1.000		0.786	0.942						0.711	0.643		0.777	0.647	0.985		S&U	0.686
11	9	0.6021	b	d	0.841	0.749	S&U		0.768	0.689						0.663	**0.458**			**0.558**	0.914			
11	9	0.6021	b	c	0.728	0.869	S&U		0.709	0.833						0.660	**0.563**			**0.549**	0.991			

† monitored events, multiply observed events and observer comparison exercise; 5% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.
Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.563	correlations which do not reach significance

Table 20: The effect of discarding data from observer *b*: correlations between primary* language occurrences recorded by pairs of observers Oi, Oj co-observing the same events† 5% significance level

					Leaders´ oral output				Participants´ oral output				Leaders´ public writing				Overseen and overheard					
chunks	df	critical r†	Oi	Oj	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**KNI**	**B**	**E**	**KR**	**S**	**other**
Keep the following correlations																						
22	20	0.4227	a	c	0.922	0.938			0.809	0.851				0.996	0.994		0.764	0.776	0.963	**0**		
22	20	0.4227	c	d	0.877	0.909			0.841	0.778				0.995	0.994		0.529	0.743	0.963			**0.206**
22	20	0.4227	e	a	0.900	0.787	S&U		0.757	0.856			S&U	0.962	0.992		1.000	0.948	0.620			
18	16	0.4683	d	e	0.847	0.899		1.000	0.801	0.698		1.000		0.995	0.991			0.994	0.919			
16	14	0.4973	e	c	0.937	0.887	0.978		0.742	0.886	0.978		0.993	0.994	0.990		1.000	0.992	0.990			
11	9	0.6021	a	d	0.953	0.972			0.963	0.803				1.000	1.000			1.000	1.000			
22	20	0.4227	f	e	0.768	0.974		0.983	0.793	0.938		1.000	1.000	1.000	0.992	1.000	0.957	0.992	0.920			
70	68	0.2500	a	f	0.823	0.936			0.834	0.993			**0**	0.837	1.000		0.579	0.498	1.000	S	S&U	0.341
11	9	0.6021	f	c	0.873	0.932			0.653	0.894				0.993	0.993			0.990	0.960			
11	9	0.6021	f	d	0.967	0.950			0.966	0.803				1.000	1.000			1.000	1.000			
and discard the following:																						
19	17	0.4555	b	a	0.767	0.794			0.841	0.987				0.707	0.491		0.855	0.501	0.919			
19	17	0.4555	b	f	0.693	0.826			0.892	0.987			S&U	0.711	0.491		0.625	0.465	0.919			S&U
18	16	0.4683	b	e	0.911	0.699			0.844	0.957			S&U	0.684	0.529		0.937	0.487	1.000		S&U	
11	9	0.6021	b	c	0.762	0.608			0.668	0.919				0.660	**0.563**			**0.549**	0.991			
11	9	0.6021	b	d	0.820	0.684			0.904	0.741				0.663	**0.458**			**0.558**	0.914			

* languages used alone in a particular category in a time slice, or used with others but marked as the primary language used

† monitored events, multiply observed events and observer comparison exercise; 5% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

S&U	excluded small datasets with one side empty and r undefined
S&=1	excluded small datasets with r = 1
S	excluded small datasets with 0<r<1
0	non-small datasets with one side empty are arbitrarily defined as having r = 0
0.563	correlations which do not reach significance are highlighted in red.

7. Exclusion of observer *b*'s data

On the basis of the above analysis, and bearing in mind the relatively large amount of data collected using the time-slice approach, it was decided to exclude the data from observer *b* in relation to schools, meetings and acts of worship, but to retain that for observer *f*. Observer *b*´s data was retained for shops and health clinic interviews[13]. Although the investigation did not suggest any particular reason for inadequate data collection, it does appear to confirm observers´ reports that time-slice observations placed considerable mental demands on observers.

The comparison data that has been retained, which can be regarded as a sample of the data collected by observers when alone, would occur by chance only 5% of the time for the critical categories, and an unknown percentage but more than 5% of the time by chance for the non-critical categories. For the comparison data that has been retained, the mean of the correlations for the critical categories is 0.897, and for the non-critical categories 0.841. These figures are for total language occurrences. For primary language occurrences the corresponding figures are 0.877 and 0.881.

8. Inter-observer reliability at shops and clinics

The approach and procedure for comparing observers´ records for shop transactions and interviews outside health clinics was the same as that described above for time slice observations. The main differences were that the tasks were described by observers as being easier than time slice observations, that the amount of available data for comparison was small, and that the number of observation categories for use in the comparisons was

13 see below for further details.

also small.

For shops, the only comparison data available was monitoring data, in which one of the RAs parallel-observed with 4 of the 5 observers in 5 different shops for periods of between 40 minutes and 1 hour, collecting data on between 6 and 18 transactions for each of the 4 observers.

For each observer, the numbers in each observation category, from the start of the monitoring exercise to the end, were strung together to produce a sequence of numbers, and this was correlated with the monitor´s equivalent sequence – transaction start time, transaction length, number of customers, customer types, and occurrences for each language. In the case of customer type and languages, each string consists of 0s and 1s, *ie* for each element of the string only two values are possible; thus these are limiting cases of interval data[14].

Table 21 shows the correlations obtained from the monitoring exercises, together with the number of transactions, the degree of freedom df (= transactions *minus* 2), and the critical value of r at the 1% significance level. It can be seen that all the correlations are significant. In other words, a hypothetical researcher seeking to establish whether or not the monitor and the observer were recording the same transactions would be able to claim that they were, also that the result was significant at the 1% level. Note that this exercise establishes concordance only between the RA and 4 of the observers. Note also that observer *b*, whose time-slice data was discarded[15], was not monitored observing shops, and therefore an interesting comparison is not possible.

14 It is possible that Pearson's r distorts at the limit.
15 see above.

Table 21: Correlations between monitor Oi and observer Oj co-observing transactions in shops: 1% significance level†

Trans-actions	df	critical r †	Oi	Oj	Start time	Transctn length	Customer, friend, business C	F	B	Language used in transaction KNI	B	S	other
8	6	0.8343	e	f	1.000	0.983	1.000	1.000	1.000	1.000	1.000		1.000
20	18	0.5614	e	c	1.000	0.879	1.000	0.909	1.000	0.909	1.000	1.000	1.000
11	9	0.7348	e	a	1.000	0.900	0.904	1.000	1.000	1.000	1.000		1.000
6	4	0.9172	e	d	1.000	0.998	1.000	1.000	1.000	1.000	1.000		

† 1% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.
Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.
1.000 Observer and monitor had a question mark here

All correlations are significant at the 1% level.

Table 22 shows a similar exercise for multiply-recorded health clinic interviews. In this case there are 6 small datasets having undefined r because one of the observer pairs´ data was empty; here, as earlier, they have been set to zero[16]. These zero correlations are the only ones in the table not to reach significance. They all occur in one of the two interviewing groups, and therefore we can conclude that our hypothetical researcher could claim that one group were interviewing the same respondents, with significance at the 1% level, but that no equivalent claim, even at the 5% significance level, could be made about the second group. Interestingly, this second group contains observer *b*, and although he is associated with 3 of the 6 small undefined datasets in the exercise, observer *c* is associated with all 6 of them. It should also be noted that observer *f*, to some degree problematic in the analysis of time-slice data, is not here the locus of any problem.

16 see section *Small and empty comparison datasets*, above.

Table 22: Correlations between pairs of interviewers Oi, Oj co-interviewing the same respondents; 1% significance level†

interview	df	critical r †	Oi	Oj	inter-views	People pres at consultn	Langs used in con		First lang of respondent					Other languages known by respondent								Pref poster lang		resp illiterate	Words not undstd	Words trans-lated	Did anyone assist?
							KNI	B	KNI	B	KR	S	Kayan	KNI	B	E	KR	S	Kayan	Manaw	Thai	KNI	B				
Group 1																											
14	12	0.6614	e	a	14	1.000	1.000		1.000						1.000	1.000						1.000	1.000	1.000	1.000	1.000	1.000
14	12	0.6614	e	a	14	1.000	1.000		1.000						0.886	1.000						0.911	1.000	0.869	1.000	1.000	1.000
14	12	0.6614	e	f	14	1.000	1.000		1.000						0.886	1.000						0.911	1.000	0.869	1.000	1.000	1.000
Group 2																											
15	13	0.6411	d	b	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000		1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
15	13	0.6411	d	c	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.869	0.913	1.000	1.000		1.000	**0.000**	1.000	1.000	1.000	**0.000**	**0.000**	1.000	1.000
15	13	0.6411	c	b	15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.869	0.913	1.000	1.000		1.000	**0.000**	1.000	1.000	1.000	**0.000**	**0.000**	1.000	1.000

† 1% significance level, 2-tailed test; Fisher and Yates, reproduced in Burns (1990):205.

Empty columns have been omitted. Blank cells here mean blank cells in both observers´ observation forms.

0.000 datasets with undefined correlation because 1 side all zero, here defined as 0; dataset sizes varied from a total of 1 to 3 items.

All correlations reach significance at the 1% level except those for small datasets, which have been set at zero here.

The overall outcome of this exercise depends on the treatment of the 6 small datasets with undefined r.

9. Further questions about clinic exit interviews

After the clinics data was processed, the results raised several questions about the validity of the questionnaire design.

The starting point for validity issues about clinic exit interviews is that of the 299 respondents whose data was processed, when asked (1) whether language difficulties came up during the consultation, and if so (2) what they were and (3) if anyone helped, all but 3 respondents replied *No* to all questions, with 2 of the 3 others responding that they did not remember, and the third referring to problems with words for a particular physical complaint.

The researcher´s expectation was that there might be language difficulties at clinic consultations, and the near-zero response to the question has raised suspicions that the translation of the questions into Burmese may be deficient, or that respondents did not fully understand what was being asked, or that the concept of ´ difficulties´ may be culture-relative. In relation to this last point, it may be that difficulties felt by a person but not raised openly are not counted as problems in Karenni society, and it may be that Karenni respondents are indicating this. Their responses may not adequately indicate to what degree they failed to understand the physician´s diagnosis or advice.

The second area of suspicion concerns the responses for the question about how long respondents spent in the consultation. This question was not important for the survey, and was intended as a warm-up question leading to the core questions about language. Of the 299 respondents whose data was used, 298 replied with an exact number of minutes between 4 and 10[17], with a mean of 6.4 and SD 2.5. It has to be admitted that the time question was not

17 the 299th responded with '30 minutes'.

thought through in depth in the design phase, but the precision and completeness of the responses, and the contrast with the blanket *Nos* in the question about difficulties raised suspicions: how many respondents had watches? Given the frequent untimeliness of Karennis[18], why would respondents know to the minute how long they had spent in an interaction? Were the observers supplying precision to answers which were really using a different set of time exponents? At the time of writing, none of the above questions has been explored.

The third issue is that 33 of the 299 respondents were recorded as illiterate when they were asked to state their preferred language(s) for text in health education posters at the clinics. The issue of literacy did not play a part in the preparation of the questionnaire, although previous education data collection in camps has indicated quite high illiteracy rates: a 2000 survey of Karen camps indicates that 43% of women and 26% of men are illiterate (ZOA Refugee Care, 2000), and a Karenni northern camp[19] household survey indicates that 64% of household representatives have had no education at all (Consortium-Thailand, 2001). In the current survey, the question of illiteracy arose when respondents were asked, first, if they had seen the health education posters in the clinics - almost all replied that they had; and second, which language(s) they preferred for written texts in the posters; at this point some respondents reported their illiteracy. However, the way in which the data were elicited means that the figure reported is unlikely to be reliable, and many more respondents may have simply said that they saw the posters and then expressed a language preference for them.

The rate of illiteracy recorded here was 11%, but assuming the clinic exit interview sample was a sample of the camp population at large, the figure is

18 the researcher´s western perception.
19 two separate camps at the time of the survey.

certainly too low, and the actual illiteracy rate is likely to be at least as high as that reported in the 2000 Karen camps survey[20]. It is not clear whether stigma attaches to illiteracy in Karenni society, as it does in western societies.

None of the above issues casts doubts on the veridicality of the language data collected in the survey. Although the issue of language status is one on which educated people express their opinions frequently in Karenni society, it is not, to the researcher´s knowledge, an especially emotive issue to talk about one´ s mother tongue or to talk about the languages one knows. If there are some suspicions that perhaps observers filled in some gaps, as indicated above[21], there is no indication that they introduced bias into responses about first or other languages known.

20 historically Karenni literacy has been lower than Karen.
21 the time question outside clinics