Variation within individual teacher administrators during national monitoring assessment interviews in social studies.

Eleanor M. Hawe
School of Teaching, Learning & Development,
Faculty of Education,
University of Auckland
NEW ZEALAND.
e.hawe@auckland.ac.nz

Isabel R. Browne
School of Critical Studies in Education,
Faculty of Education,
University of Auckland
NEW ZEALAND.
i.browne@auckland.ac.nz

Abstract
In New Zealand, the National Education Monitoring Project (NEMP) is responsible for the national assessment of primary school students' achievement across the curriculum. Standardised tasks with associated interviews are one of the assessment approaches used to assess achievement. This paper addresses the reliability of one-to-one social studies interviews, in particular the variation within individual Teacher Administrators (TAs) during the 2005 round of monitoring. An observation schedule was used to gather data across ten categories as 12 randomly selected TAs carried out three administrations for each of three selected tasks. Levels of internal variation for each TA across these categories are presented along with categories where the greatest and least levels of internal variation were exhibited. It is suggested that singly and/or in combination, specific task related features and the way(s) in which each TA perceived and constructed their role affected the conduct of interviews. Overall, the levels of variation observed and the nature of the variations pose threats to reliability and ultimately to the validity of claims regarding students' achievement. This is particularly concerning when the assessment information and related claims are used as the basis for reporting national patterns of educational achievement and making recommendations to stakeholders.

National monitoring; assessment interviews; intra-administrator consistency.

Introduction: National monitoring in New Zealand

The National Monitoring Education Project (NEMP) in New Zealand has carried out annual national assessments of Year 4 and Year 8 students' achievement, skills and attitudes, over four yearly cycles, in each of the seven essential learning areas of the curriculum (Ministry of Education, 1993) since 1995. The two overarching purposes of national monitoring are to:

- meet public accountability and information requirements by identifying and reporting patterns and trends in educational performance and;
- provide high quality, detailed information which policy makers, curriculum planners and educators can use to debate and review educational practices and resourcing (Flockton & Crooks, 2002).

New Zealand's approach to national monitoring has a number of features that make it quite different from large-scale assessment programmes in other countries. Rather than assessing all primary school students, or all students at specific years, a light sampling (3-5%) is taken at Year 4 (midway through primary education) and Year 8 (at the end of primary education) using a variety of assessment approaches, for example: one-to-one video taped assessment interviews with associated tasks; independent pencil and paper tasks; team tasks; and independent performance tasks set out at a series of 'stations' (Flockton & Crooks, 2002). In addition, teachers are employed to administer and at a later stage to mark the tasks. The administration of one-to-one assessment interviews from the third cycle of social studies monitoring is the focus of this paper.

The one-to-one assessment interview

Twenty-eight of the 45 social studies tasks administered in 2005 took the form of one-to-one assessment interviews where the administrator reads the set requirements of each task to the

student and asks a standard set of questions, confirming both before and during the task that the student understands what they are to do. Students' responses are communicated either orally, by demonstration, in writing, through computer files, and/or through other physical artefacts (Flockton, 1999). All one-to-one interviews are recorded on videotape to obtain a detailed picture of what students and administrators do and say, and to enable analysis and evaluation of students' responses at a later time (Flockton & Crooks, 2002).

Each year NEMP engages approximately 100 teachers to administer the assessment tasks. These teachers spend a week undergoing an intensive training programme where they are briefed on their role, become familiar with the tasks for that cycle and are instructed in the use of technology for delivering the tasks and recording information. The main aim of this week is to ensure that the Teacher Administrators (TAs) are "trained to conduct the assessment of children with accuracy and in a standardised way" (Gilmore, 1999, p.6). Once training is finished two TAs are assigned to administer the tasks to 60 children in at least five different schools (12 per school) over a five-week period. Despite the focused and intensive nature of the training programme it has been suggested that some TAs have not been consistent in the way they administer tasks (Bowyer & Meaney, 2007; Browne & Hawe, 2005; Gilmore, 1999).

Threats to reliability and validity

There are a number of factors that can affect the reliability of the information yielded by an assessment and the validity of any claims. It has long been recognised that student performance "can be greatly influenced by the procedures followed in presenting and administering the tasks" (Crooks, Kane & Cohen, 1996, p.268) and that small variations in administrative procedures and administrators' actions and interactions can translate into persistent, systematic errors in results (Baker & O'Neill, 1994). Failure to control these factors can restrict the generalisability of results (Crooks, Kane & Cohen, 1996; Shavelson & Baxter, 1991).

Variations between administrators and across administrations have been reported in relation to the implementation of a compulsory national assessment programme in England during the early 1990s (Gipps, Brown, McCallum, & McAlister, 1995; James & Conner, 1993), but much of the research literature regarding consistency in large-scale assessment has focused on the interpretation and marking of student performance (for example Baird, Greatorex & Bell, 2004; Brown, 1999; Shavelson & Baxter, 1991). These studies are important, but secondary to what is of primary concern in assessment interviews, that is the extent to which the administration of the task and associated interview is standardised both between administrators and within a single administrator. This paper reports on the nature of variations within individual TAs, in their administrations of one-to-one NEMP social studies assessment interviews.

Method

As the study was a NEMP probe project (see Gilmore, Lovett & van Hasselt, 2003), the two researchers were given access to the video-taped one-to-one assessment interviews in social studies from the 2005 cycle of national monitoring. The project was carried out in two phases.

***Phase one** – selection of one-to-one assessment interview tasks*

The first phase focused on selection of three one-to-one assessment interview tasks from the 2005 cycle (see Browne & Hawe, 2009). The tasks selected were Powhiri, Homes and Up and down (Crooks, Flockton & Meaney, 2006). Powhiri assessed students' knowledge of cultural customs and traditions with reference to a welcoming ceremony on a marae [1]. The TA asks the student if he/she has ever been on a marae and if so to explain what it is. If they do not know, they are told what it is. The student is then asked to order six photographs depicting aspects of

[1] A powhiri is a formal ceremony held to welcome visitors to a marae. A marae is a Maori (New Zealand's indigenous people) meeting place that encompasses a meeting-house, dining area, cooking facilities and a sacred area in front of the meeting-house. It is a place where important cultural ceremonies such as welcoming guests and farewelling the dead are carried out.

a welcome onto a marae and to use the photographs as he/she explains specific protocols and practices that form part of a welcoming ceremony. At the conclusion of the task the TA records the sequence of the photographs. The second task, Homes, aims to assess "understanding [of] differences between environments" (Crooks, Flockton & Meaney, 2006, p.35). A photograph of an Ethiopian family and their home environment is used to stimulate thinking and responses about ways in which living in Ethiopia is different and the same as living in New Zealand and how living in New Zealand is different and the same as living in Ethiopia. The purpose of the third task, Up and down, is to ascertain knowledge of factors influencing the price of commodities and an understanding of how these factors influence prices (Crooks, Flockton & Meaney, 2006). A photograph of a petrol station displaying the price of petrol is shown to the student and he/she is asked to itemise some of the causes for changes in the price of petrol, identify where New Zealand gets its oil from and explain how the price of petrol going up can cause the price of commodities such as bread to go up. Each of these tasks offers opportunities for extended interactions as TAs use the interviews to lever out students' understandings.

Developing and trialling the observation schedule

A two-part structured observation schedule was developed to capture and analyse TA administration of the interview-based tasks (see Browne & Hawe, 2009). Part A was task specific with the left hand side of the schedule containing standardised statements, questions and procedures for highlighting or ticking, and the right hand side having space for a semi-structured 'running record' as the researchers observed the TA and student during an interview. Part B, identical for each of the three tasks, summarised information from Part A in ten categories. The first two categories addressed an administrator's adherence to standardised statements and set procedures; the third category captured information about the use of wait-time. The remaining seven categories addressed administrator use of: general, non-specific verbal approval/encouragement; non-verbal approval/encouragement; verbal prompts; verbal probes; evaluative statements regarding an aspect of a student's response; additional administrative related statements/questions and; statements in response to a student query (see Browne & Hawe, 2009). For each of these seven categories, TA 'moves' were summarised on a four-point scale according to the number of instances observed: none; few (1 to 5); some (6 to 10); and many (more than 10 – actual number recorded) (see Appendix B). As part of the iterative process of development and refinement, the schedules were trialled with the two researchers independently observing and rating twenty-four interviews (eight per task). Levels of agreement were estimated for each category across half of the observations. Phase two did not begin until the two researchers had reached agreement for at least 80 percent of their ratings for every category across these observations.

Phase two

The observation schedules were used in phase two to record and analyse interviews from 12 randomly selected TAs (1-12). Three administrations of each of the three tasks (Powhiri, Homes, Up and down) were observed for each TA. This gave nine observations per TA. The two researchers each observed and analysed half of the interviews (randomly selected). To ensure they had a common understanding of the categories, they observed and recorded information for the same interview from time to time. Throughout data collection, percentages of agreement for these common observations were at or exceeded 80 percent. Once data collection was complete, information on Part B of the schedule was used to determine overall levels of variation within each TA in relation to the ten categories (see Browne & Hawe, 2009).

Overall levels of intra-administrator variation across all tasks

Table 1 presents a general picture of each TA's overall levels of variation across the nine administrations with each cell representing a single observed category. It shows for example, that TA5 exhibited a large level of internal variation in one category [dark grey], a moderate level of internal variation in four categories [trellis pattern], a small level of internal variation in four categories [vertical pattern] and no internal variation in one category [light grey].

Table 1 in here (Appendix A)

Five (42%) of the twelve administrators (TAs 5, 6, 7, 9 and 10) displayed large levels of internal variation in one category. In addition three (25%) of the twelve (TAs 3, 5 and 6) showed moderate and/or large levels of internal variation in at least five of the ten categories. Relatively speaking, TAs 3, 5 and 6 were the most variable, or the least internally consistent of the administrators, across the nine interviews observed. In contrast, four (33%) of the twelve (TAs 4, 9, 11 and 12) exhibited either a small level of intra-administrator variation or no variation in at least eight of the ten observed categories and a moderate and/or large level of variation in the remaining one or two categories. These four administrators were the least internally variable, or the most internally consistent, across the nine interviews observed. The remaining five (42%) administrators (TAs 1, 2, 7, 8 and 10), fell between these two extremes, with each showing either a small level of variation or no variation in at least six of the ten categories, and moderate to large variation in three or four categories. To some extent therefore all TAs displayed a large or moderate level of inconsistency in at least one aspect of their administrative practice and all exhibited moderate to small levels of variation in at least six and in some cases as many as eight or nine aspects of their administration.

While Table 1 indicates which TAs displayed the most and least levels of internal variation across the nine tasks, it does not identify the actual nature of these variations. Table 2 provides this information – it summarises overall levels of internal variation (nil, small, moderate, large) for each administrator (TA1-12) in relation to each of the ten observation categories.

Table 2 in here (Appendix B)

Areas of greatest internal variation

Moderate and/or large levels of intra-administrator variation were apparent in four categories: general non-verbal encouragement/approval, where ten of the twelve (83%) displayed either a large or moderate level of inconsistency; general, non-specific verbal encouragement /approval, where nine (75%) TAs demonstrated either a large or moderate degree of variation; verbal prompts, where seven (58%) TAs demonstrated either a large or moderate degree of variation and; in the 'following of standardised task procedures' where five (42%) TAs were moderately variable in their practice.

General, non-verbal encouragement/approval; general, non-specific verbal encouragement / approval; verbal prompts

The level of internal variation observed in each of these three categories was task specific - with particular reference to Homes and Up and down. The Homes task and its associated questions encouraged students to "tell … as many things … as you can think of" and "tell … as many ideas as you can think of …" in relation to the focus questions (Crooks, Flockton & Meaney, 2006, p.35). This resulted in TA6 for instance using numerous prompts such as "Some other things?", "What else?", "What other things?" and "What else would they need to learn …?" during each interview. In addition, each student response was reinforced with general verbal encouragers such as 'mmm', 'uh-huh', and 'okay' and/or non-verbal encouragers in the form of nods, smiles and the like. Like her counterparts, TA6 urged each student to identify as many objects in the photograph as possible. The nature of the task and the way in which the majority of TAs interpreted it, meant that once a student had provided an initial response, they were prompted and encouraged to provide more and more ideas. This task lent itself to, and was interpreted as requiring the generation of multiple responses and the reinforcement of these responses, often with little regard to the quality of the responses.

In contrast, Up and down called for an explanation regarding petrol price fluctuations and the effects of these fluctuations on the price of other commodities. The following interaction was typical for Up and down:

TA6: "The price of petrol goes up and down. What are some of the things that cause the price to change?"
Student: "People not buying, um getting as much fuel /
TA6: "Mmm"

Student: so they put it down and if they're getting too much they might put it a bit higher because they're running out."
TA6: "Okay, thank you, anything else?"
Student: "No."

While the set questions, like those in Homes, invited students to offer more than one explanation, the challenging conceptual nature of the task (Crooks, Flockton & Meaney, 2006) was such that many students offered no more than a single, short response to each question and seemed unable to provide any further ideas even after a perfunctory prompt or encourager. In addition, the lack of clarity in a number of the students' responses resulted in the TAs moving on, as in the above example, rather than seeking clarification.

A task-by-task comparison in the use of verbal prompts and non-verbal and verbal encouragement/approval for Administrators 1, 2, 3, 6, 7, 8 and 10 highlights the task specific nature of the variation. For each of these seven administrators, the three students completing Homes received far more prompts, and verbal and non-verbal forms of encouragement/approval than the students who completed Up and down. TA1 for instance was observed giving 16, 20 and 12 instances of general verbal encouragement/approval during the Homes task to her three students respectively, while during Up and down the three students each received less than five such encouragers, and during Powhiri the three students respectively received six, seven and nine general verbal encouragers.

Following standardised procedures

Close analysis of the data revealed a further category where TAs were internally inconsistent and again this variation was task specific. All TAs followed the set procedures for Homes and Up and down. With reference to Powhiri however, while four (33%) of the TAs (TAs 2, 4, 9, 12) followed the two set procedures according to NEMP requirements, three (25%) TAs (1, 10, 11) altered the second procedure for one or two of the three interviews and five (42%) TAs (3, 5, 6, 7 and 8) altered the second procedure in each of the three interviews. The second procedure related to the recording of the order of the photographs: this was supposed to be completed at the end of the task, following the student's explanation about the welcoming ceremony. In instances where this was not followed, the TAs recorded the order of the photographs either prior to the student's explanation or as he/she was explaining their understanding of the ceremony. TAs 3, 5, 6, 7 and 8 recorded the order either prior to or during the student's explanation while TAs 1 and 11 'got it right' once or twice and on the other occasion(s) recorded either before or during the student's explanation. TA 10 however was inconsistent across the three occasions observed - in one interview the order was recorded earlier than set down, in another interview it was recorded correctly at the end of the explanation and in the third interview there was no evidence of any recording.

Areas of least internal variation

No internal variation or only small levels of internal variation were apparent in two categories: following set statements/questions and use of verbal probes.

Following set statements / questions

Intra-administration variation was negligible with reference to the following of standardised statements/questions. All TAs read the set questions and/or statements with reasonable accuracy on all occasions. Any changes noted were slight and they did not affect the overall intent of the statement/question. Thus administrators consistently followed the set script.

Use of verbal probes

There was no internal variation observed in the use of verbal probes for five (42%) of the twelve TAs. Administrators 2, 4, 8, 9 and 11 were internally consistent as they did not probe any responses from their students across the nine interviews observed. Of the remaining seven (58%) administrators, each probed the responses of between one and four of their nine students on one occasion and never more than twice in a single interview. Probes included statements such as the following where students were asked to explain further or clarify what they were saying with a view to obtaining a more elaborated response:

"Can you explain what you mean by a complaint a bit more?" (TA6);

> "When you say learn how things are done in this country can you be a bit more specific?" (TA10).

Overall, TAs seemed reluctant to delve deeper into students' responses (see Hawe & Browne, 2010, paper under review) hence the lack of internal variation.

Discussion

During their training week, TAs were briefed on the nature of their NEMP role. Bowyer and Meaney (2007) reported that a number of the 2005 cohort of TAs (the same cohort that administered tasks in the current project) identified the ability to keep to the standardised questions, statements and procedures as one of the most important skills learnt during this time. More specifically, TAs acknowledged that as a result of the training week they were aware and mindful of 'upsetting' or 'damaging' the results through any deviation from standardised protocols (Bowyer & Meaney, 2007). While all TAs were internally consistent in presenting set statements and questions correctly, they were less internally consistent in the following of set procedures. The internal variation observed in this category was however task specific. All TAs followed procedures during Up and down and Homes, but seventy-five percent altered the second procedure during Powhiri. Although the latter task was relatively more complex in its procedural requirements than Up and down and Homes, it was not overly complicated. Recording the order of the photographs earlier than set out, either before and/or during a student's explanation, may have been due to oversight. Alternatively TAs may have believed they were saving time by recording the order when they did. Failure to follow the protocol may also be due to factors observed in other studies such as administrator fatigue, disinterest and/or over-familiarity with the task (Broadfoot, 1995; James & Connor, 1993).

The task-specific variation within individual TAs in the following of set procedures is considered significant on two counts. Firstly, in a number of the Powhiri interviews it was noticeable that the students were aware the TA was recording the order during their explanation rather than giving them their full attention. Divided attention and expressions of disinterest during an assessment interview not only have a de-motivating effect on students (Flockton, 1999) they can divert the administrator from gathering a more elaborate or complete explanation from them. In these instances it seemed that TAs considered recording the order to be as important if not more important than helping the student give their best possible explanation. Secondly, there were students who realised during their explanation that they wanted to change the order of the photographs. While two TAs picked up on this and time was then spent establishing the 'final' order and re-recording it, students in general seemed reluctant to call attention to this, as they were aware the order had already been recorded. It is argued that the failure of TAs to follow set procedures during Powhiri prematurely 'closed down' (Smith & Higgins, 2006) the interaction and task, and possibly disadvantaged a number of students. Changes to standardised practices impact on the way students respond to a task with small variations having the potential to turn into regular and persistent difficulties in results (Baker & O'Neill, 1994; Eley & Caygill, 2002). In this case, the changes worked against NEMP's aim "to maximise and allow students the best possible opportunities to show what they can do in response to the tasks they are given" (Flockton, 1999, p.7).

Over half of the TAs exhibited either moderate or large levels of internal variation in their use of general, non-specific verbal encouragement / approval, general non-verbal encouragement / approval and verbal prompts. This variation was also task specific. The nature of the questions in Homes resulted in TAs constructing the task as requiring little more than the identification of as many (similar or different) features present in or absent from the photograph as they could rather than an exploration of students' "understanding [of] differences between environments" (Crooks, Flockton & Meaney, 2006, p.35). In the context of national monitoring, the level of variation in individual TA practice across the three tasks is concerning insofar as students were given considerably more opportunities to respond during Homes than during either Powhiri or Up and down. Students who are repeatedly encouraged and prompted may provide eight or nine possible answers to a question of which five or six are valid. They would very likely score higher on this task than on one where they are prompted once or twice and provide one valid response. While it would be unrealistic to expect TAs to use virtually

the same number of encouragers and prompts across tasks and interviews, they do need to be aware of how even small adjustments in their practice can affect students. Furthermore, NEMP needs to be more cognisant of how the wording of the questions can impact on TA practice. Notwithstanding the conceptual difficulty of Up and down, it is hardly surprising to find that the results for Homes show superior levels of student achievement in comparison to those recorded for Powhiri and Up and down (Crooks, Flockton & Meaney, 2006). The magnitude of the task related variation within individual TAs in this aspect of their practice, coupled with the way in which they interpreted Homes and the challenging nature of Up and down, are sufficient to raise questions about the reliability of the assessment process and the subsequent validity of the interpretations for these tasks.

Finally, it is ironic that the high levels of internal consistency in TAs' probe related practice pose a threat to the validity of NEMP's claims (see also Hawe & Browne, 2010, paper under review). A surprising outcome from the project was the consistent failure of TAs to probe a student's responses to ascertain if there was any understanding of critical underlying social studies concepts. The probes observed took the form of uptake statements (Nystrand, Gamoran, Kachur & Prendergast, 1997) where TAs reacted in an almost automatic manner to a student's response. There were no examples of the more substantive kinds of probes that encourage students to engage in higher levels of thinking – in this instance to explore and discuss social studies ideas, concepts and generalisations. It is clearly signalled in the curriculum that the aim of social studies will not be met if the focus is solely on collecting factual information and acquiring knowledge – these should be used to develop understandings about society (Ministry of Education, 1997). Perhaps the TAs were hamstrung by the failure of NEMP to provide standardised, substantive probes for use with any of the three tasks (see Crooks, Flockton & Meaney, 2006). Alternatively, TAs may have failed to recognise, during the interview, the significance of what their students were saying. With reference to Powhiri, for example, TAs may have had little more than a general knowledge of marae protocol and Maori culture so they were not in a strong position to notice and/or recognise the significance of a student's response, let alone respond to it with an appropriate probe. Thus the superficial level of students' knowledge and understanding reported in relation to this task (Crooks, Flockton & Meaney, 2006) may be a reflection, to some extent, of TAs' limited knowledge and understanding in the area rather than any lack on the behalf of the students. In relation to Up and down, the challenging nature of the task for the students and the limited nature of their responses meant that TAs had little to draw on even if they wanted to probe. Furthermore, the nature of the questions that accompanied Homes, in association with the way TAs' constructed the task, did little to encourage any probing of responses. A further, more inclusive explanation that could account for the wholesale lack of probing lies in the nature of teachers' everyday interaction and discourse styles. More than twenty years ago Newmann (1988) observed that teachers in classrooms rarely probed students' responses. Since this time numerous studies have noted teachers' over-reliance on directive approaches to teaching at the expense of discourse that provides opportunities for students to explore and elaborate on ideas and demonstrate understandings (Burns & Myhill, 2004; Smith & Higgins, 2006; Smith, Hardman, Wall & Mroz, 2004). Given that teachers typically ask few questions that require students to apply, analyse, synthesise and/or evaluate information, it would be unusual if they suddenly and spontaneously did this in their role as a NEMP administrator.

In conclusion, the factors discussed above highlight important issues about the robustness of the assessment process, the reliability of information and the validity of subsequent claims. More specifically, they raise questions about whether the assessment process and information gathered are sufficiently robust and defensible to meet the demands for public accountability and to provide high quality, detailed information which policy makers, curriculum planners and educators can use to inform the debate and review of educational resourcing and practices (Flockton & Crooks, 2002).

References

Baird, J., Greatorex, J., & Bell, J.F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: principles, policy & practice 11(*3), 331-348.

Baker, E.L., & O'Neill, H.F. (1994). Performance assessment and equity: A view from the USA. *Assessment in Education: principles, policy & practice, 1*(1), 11-26.

Bowyer, L., & Meaney, T. (2007). NEMP probe study report. The effects of teacher perceptions on adopting a new role. Retrieved on 30 June 2008 from http://nemp.otago.ac.nz/_probes.htm

Broadfoot, P. (1995). Performance assessment in perspective: international trends and current English experience. In H. Torrance (Ed). *Evaluating authentic assessment* (pp.9-43). Buckingham: Open University Press.

Brown, N. (1999). NEMP marking procedures and consistency. Paper presented at the combined conference of the New Zealand Association for Research in Education and the Australian Association for Research in Education, 29 November – 2 December, Melbourne, Australia.

Browne, I., & Hawe, E. (2009). Variations in the administration of NEMP one-to-one assessment tasks (interviews) in social studies. A NEMP probe project. Auckland: Uniservices Ltd.

Browne, I., & Hawe, E. (2005). Social studies: Assessment year 8 students' knowledge and understanding about New Zealand society. A National Educational Monitoring Probe Project. Retrieved on 30 June 2008 from http://nemp.otago.ac.nz/_probes.htm

Burns, C., & Myhill, D. (2004). Interactive or inactive? A consideration of the nature of interaction in whole class teaching. *Cambridge Journal of Education* 34, no. 1: 35-49.

Crooks, T., Flockton, L., & Meaney, T. (2006). *Social studies Assessment results 2005.* Dunedin: Educational Assessment Research Unit, University of Otago.

Crooks, T.J., Kane, M.T., & Cohen, A.S. (1996). Threats to the valid use of assessments. *Assessment in Education: principles, policy and practice 3*(3), 265-285.

Eley, L., & Caygill, R. (2002). One test suits all? An examination of differing assessment task formats. *New Zealand Journal of Educational Studies 37*(1), 27-38.

Flockton, L. (1999). *School-wide assessment. National education monitoring project.* Wellington: New Zealand Council for Educational Research.

Flockton, L., & Crooks, T. (2002). *Social Studies. Assessment results 2001*. Dunedin: Educational Assessment Research Unit, University of Otago.

Gilmore, A. (1999). The NEMP experience. Professional development of teachers through the National Education Monitoring Project. Retrieved 30 June 2008 from http://nemp.otago.ac.nz/probes.htm

Gilmore, A., Lovett, S., & van Hasselt, C. (2003). *NEMP probe study findings.* Dunedin: Educational Assessment Research Unit.

Gipps, C., Brown, M, McCallum, B. & McAlister, S. (1995). *Intuition or evidence?* Buckingham: Open University Press.

Hawe, E.M., & Browne, I.R. (2010, paper under review). National monitoring in social studies: the reliability of assessment interviews.

James, M., & Conner, C. (1993). Are reliability and validity achievable in national curriculum assessment? Some observations on moderation at key stage 1. *Curriculum Journal, 4*(1), 5-19.

Ministry of Education. (1997). *Social studies in the New Zealand curriculum.* Wellington: Learning Media.

Ministry of Education. (1993). *The New Zealand curriculum framework.* Wellington: Ministry of Education.

Newmann, F.M. (1988). A test of higher-order thinking in social studies: Persuasive writing on constitutional issue using NAEP approach. *Social Education, 54*(4), 369-373.

Nystrand, M., A. Gamoran, R. Kachur, and C. Prendergast. (1997). *Opening dialogue: understanding the dynamics of language and learning in the English classroom.* New York: Teachers College Press.

Shavelson, R.J., & Baxter, G.P., with Pine, J. (1991). Performance assessment in science. *Applied Measurement in Education, 4*(4), 347-362.

Smith, F., Hardman, F., Wall, K., & Mroz, M. (2004). Interactive whole class teaching in the national literacy and numeracy strategies. *British Educational Research Journal* 30, no. 3: 395-411.

Smith, H., & Higgins, S. (2006). Opening classroom interaction: the importance of feedback. *Cambridge Journal of Education 36*(4), 485-502.

APPENDIX A

Table 1 Levels Of Intra-administrator Variation For Each Of The Ten Observation Categories

TA1	TA2	TA3	TA4	TA5	TA6	TA7	TA8	TA9	TA10	TA11	TA12

KEY: TA = Teacher Administrator;
Light grey = no variation; Vertical = small variation; Trellis = moderate variation; Dark grey = large variation

APPENDIX B

Table 2 Overall Levels Of Variation For Each Administrator, Across All Administrations, According To Observation Categories.

	Observation categories									
	Proced fll'wd	**Statem't fll'wd**	**Wait time**	**Verbal app'vl/ encgmt**	**Non vbl app'vl / encgmt**	**Verbal prompt**	**Verbal probe**	**Eval'tve statem't**	**Task admin statem't**	**Resp to stud query**
TA										
1	Small	Small	Nil	Mod	Mod	Mod	Small	Small	Small	Small
2	Nil	Nil	Small	Mod	Mod	Mod	Nil	Small	Mod	Small
3	Mod	Nil	Small	Mod	Mod	Mod	Small	Small	Mod	Small
4	Nil	Small	Small	Small	Small	Mod	Nil	Small	Mod	Small
5	Mod	Nil	Small	Mod	Large	Mod	Small	Small	Mod	Small
6	Mod	Nil	Mod	Mod	Small	Large	Small	Mod	Mod	Small
7	Mod	Nil	Small	Mod	Large	Small	Small	Small	Mod	Small
8	Mod	Nil	Small	Small	Mod	Small	Nil	Mod	Small	Small
9	Nil	Small	Small	Mod	Large	Small	Nil	Nil	Small	Small
10	Small	Nil	Mod	Mod	Mod	Large	Small	Small	Small	Small
11	Small	Nil	Small	Mod	Mod	Small	Nil	Small	Small	Small
12	Nil	Small	Small	Small	Mod	Small	Small	Small	Small	Small

KEY: TA = Teacher Administrator Mod = Moderate

Levels of variation:

Nil variation - in the case of the yes/no categories[2] the administrator followed all set procedures and statements/questions; in relation to the remaining categories[3] the administrator's 'moves' are all located at the same point on the observation schedule eg: the TA gave between six and ten (some) verbal prompts in each of the nine administrations observed;

A small level of variation - between two and three of the administrations did not follow the set procedures and statements/questions; in relation to the remaining categories, the TA's 'moves' are located at two adjacent points on the observation schedule eg: the administrator made between one and five (few) evaluative statements during each of four administrations and between seven and ten (some) such statements during each of the remaining five administrations;

A moderate level of variation - between four and eight of the administrations did not follow the set procedures and statements/questions; in relation to the remaining categories, the TA's 'moves' are located at three adjacent points on the observation schedule eg: no prompts were observed during two administrations, between one and five (few) prompts were noted in four of the administrations and between six and ten (some) in three of the administrations;

A large level of variation - between nine and twelve administrations did not follow the set statements / questions; in relation to the remaining categories, the TA's 'moves' are spread across all (none-few-some-many) points on the observation schedule.

[2] For 'following set procedures' and 'following set statements/questions', the point of reference is variation from the set or standardised procedures and statements / questions.

[3] For the remaining categories, observations are recorded according to: no observed instances; few (1-5) instances; some (6-10) instances; many (more than 10) instances.