Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 15, Number 1, January 2010 ISSN 1531-7714

Mixing interview and questionnaire methods: Practical problems in aligning data

Lois R. Harris, *The University of Auckland*
Gavin T. L. Brown, *The Hong Kong Institute of Education*

Structured questionnaires and semi-structured interviews are often used in mixed method studies to generate confirmatory results despite differences in methods of data collection, analysis, and interpretation. A review of 19 questionnaire-interview comparison studies found that consensus and consistency statistics were generally weak between methods. Problems in aligning data from the two different methods are illustrated in a questionnaire-interview study of teacher conceptions of assessment. Poor alignment appeared attributable to: differences in data collection procedures, the complexity and instability of the construct being investigated, difficulties in making data comparable, lack of variability in participant responses, greater sensitivity to context and seemingly emotive responses within the interview, possible misinterpretation of some questionnaire prompts, and greater control of content exposure in the questionnaire. Results indicated that if 'confirmatory' results are being sought, researchers must create tightly aligned and structured instruments; present the construct in a simple, concrete, and highly contextualised manner; collect the two types of data with a minimal time gap; and estimate agreement between methods using consistency statistics. However, the cost of confirmation through strong alignment may lead to the loss of rich complementary data obtained through allowing each method to be analysed in its own right.

Introduction

Questionnaires and interviews are often used together in mixed method studies investigating educational assessment (e.g., Brookhart & Durkin, 2003; Lai & Waltman, 2008). While questionnaires can provide evidence of patterns amongst large populations, qualitative interview data often gather more in-depth insights on participant attitudes, thoughts, and actions (Kendall, 2008).

This article focuses on comparing structured questionnaires with semi-structured interviews, although other types of questionnaires and interviews are examined in the literature review. In a structured questionnaire, participants respond to prompts by selecting from predetermined answers (e.g., Likert scales, multiple choice responses); these data are typically analysed quantitatively. In a semi-structured interview, interviewers begin with a small set of open-ended questions, but spend considerable time probing participant responses, encouraging them to provide detail and clarification; these data are generally analysed qualitatively.

A REVIEW OF LITERATURE

Methodological differences between questionnaires and interviews

Although questionnaire and interview data are frequently reported together, studies within educational contexts seldom examine the level of similarity between people's questionnaire and interview responses to determine when and if comparisons between these data sets are appropriate. Mixed methods researchers deny that the paradigmatic differences between ways of viewing the world make qualitative and quantitative methods incompatible (e.g., Day, Sammons, & Gu, 2008; Ryan & Bernard, 2000; Smith, 2006). However, the protocols for data collection and analysis developed for questionnaires and interviews may have evolved from differing ways of viewing the world making it possible that method effects exist.

In the research methods literature, questionnaires and interviews are seen as having differing and possibly complementary strengths and weaknesses. While questionnaires are usually viewed as a more objective research tool that can produce generalisable results because of large sample sizes, results can be threatened by many factors including: faulty questionnaire design; sampling and non-response errors; biased questionnaire design and wording; respondent unreliability, ignorance, misunderstanding, reticence, or bias; errors in coding, processing, and statistical analysis; and faulty interpretation of results (Oppenheim, 1992). Additionally, questionnaire research can be seen as over-reliant on instruments and, thus, disconnected from everyday life, with measurement processes creating a spurious or artificial sense of accuracy (Bryman, 2008).

Neither are interviews neutral tools; here data are based on personal interactions which lead to negotiated and contextually based results (Fontana & Frey 2000; Silverman, 2000, 2006). While interviews provide contexts where participants can ask for clarification, elaborate on ideas, and explain perspectives in their own words, the interviewer can use questioning to lead or manipulate interviewee responses. Due to the interpersonal nature of the interview context, participants may be more likely to respond in ways they deem socially desirable (Richman, Keisler, Weisband, & Drasgow, 1999; Yin, 2009). Hence, interview data are always contrived and, at best, partial and incomplete understandings of a participant's point of view (Lankshear & Knobel, 2004). Additionally, since most qualitative studies have relatively small sample sizes, the results can be difficult to replicate or generalise (Bryman, 2008).

Further differences between the two methods can occur through the coding and analysis of the data. For example, while quantitative data are numeric and, hence, more objective, considerable researcher interpretation comes into decisions about excluding specific participants and/or items from the data set, the statistical processes employed to generate results, and the interpretation of results (Oppenheim, 1992). With qualitative data, researchers generally utilise a process of inductive coding, which can be easily influenced by researcher subjectivities (Bryman, 2008). It can also be difficult to judge how well proposed qualitative categorisations actually suit the data as normally only small excerpts are presented for scrutiny in manuscripts.

Issues relating to self-reporting also plague both methods. Studies have found that people can simultaneously hold conflicting conceptions and beliefs (Marton & Pong, 2005; Pajares, 1992) which may cause them to respond in seemingly contradictory or inconsistent ways. Additionally, poor or incomplete memory of events, external influences, and lack of time to fully recall information may lead to purposefully or accidentally inaccurate recall and responding (Brewer, Hallman, Fielder, & Kipen, 2004). Also, some people may respond based on what they believe is socially desirable rather than what they think is true; research disagrees about which research modes are most influenced in this way (Richman et al., 1999). Oei and Zwart (1986) suggested that participants actually respond differently to questionnaire and interview prompts, claiming that face-to-face interviews trigger strong affective responses while questionnaires permit a wide range of responses, of, perhaps, a more cognitively dispassionate nature. However, despite the weaknesses of both questionnaires and interviews, these are important means of obtaining direct responses from participants about their understandings, conceptions, beliefs, and attitudes; hence, these methods cannot and should not be discarded.

Studies comparing questionnaire and interview results

When examining studies comparing these two methods, there are many ways to measure their level of similarity. Consensus of scores is deemed to be present if the exact agreement between methods is 70% or better (Stemler, 2004). However, when scales involve few rating points (i.e., three or less), then it is highly probable that this agreement will be due to chance. Many studies use Cohen's (1960) kappa statistic (κ) to indicate the degree to which observed rates of agreement are attributable to chance processes. Negative kappa values indicate that the observed levels of agreement are less than would occur purely by chance, zero kappa means consensus rates are equal to chance, values up to .40 are deemed slightly better than chance, values of .41 to .60 are moderately better than chance, while kappa values greater than .60 are stronger than chance. Some studies also use consistency statistics (i.e., Pearson's *r* or Cronbach's alpha) to indicate the degree to which, regardless of the actual value awarded, the two methods give higher or lower scores to each participant.

Generally, coefficients greater than .70 indicate adequate reliability (i.e., consistency between methods accounts for 49% or more of variance) (Stemler, 2004).

It is worth noting that consistency coefficients will be depressed when the rating scale is short (e.g., only three points) or if there is little discrimination in a set of ratings (i.e., very few differed from each other). In circumstances when there is much agreement, these coefficients will under-report inter-method agreement. Another factor in depressing consistency and consensus correlation statistics is the nature of distributions (Orwin, 1994); high inter-rater agreement rates combined with little variation across ratings can result in kappa and correlation statistics being close to zero or even negative.

The majority of work done comparing questionnaire and interview methods has been in medicine and psychology, generally for the purpose of examining the validity of particular screening or diagnostic instrument (Reams & Twale, 2008; Valentinis et al., 2009). However, occasional studies appear in other discipline areas such as education (Villanova, 1984), information technology (Richman et al., 1999), and religious studies (Holm, 1982; Hunsberger & Ennis, 1982). The results of such validity studies are mixed (see Appendix Table 1). The studies reviewed in this table were found by searching a wide range of multidisciplinary data bases (i.e., ERIC, Google Scholar, Ovid, ProQuest, PsycINFO, PubMed, SAGE, ScienceDirect, Scopus, SpringerLink, Web of Science (ISI)) using keywords such as "compare" and "test validity" along with "questionnaire" and "interview". Once a study was located where the comparability of participant responses to both methods had been evaluated, relevant studies it cited were also collected and reviewed. Additionally, more recent studies were found that had cited the selected manuscript.

While the majority of studies reviewed in Appendix Table 1 show positive levels of agreement between methods, this may be in part because researchers are unlikely to publish results which would be seen as discrediting their instruments' validity and reliability. When differentiating between studies showing high and low levels of agreement, variables relating to the study design appear important. First, there is the level of stability that can be expected from the construct being measured; an attribute which is highly transient or contingent is unlikely to elicit consistent responses. Second is the degree of alignment between the questionnaire and interview; it would be expected that interviews which utilised prompts extremely similar to questionnaire items would achieve more consistent results. Third is the emotional and mental stability of the population under study; participants who lack stability in their own lives may tend to respond inconsistently to research instruments (Manassis, Owens, Adam, West, & Sheldon-Keller, 1999; Rojahn, Warren, & Ohringer, 1994).

It is useful to examine several studies in depth to discuss how these variables appear to affect levels of agreement. For example, Bergmann, Jacobs, Hoffmann, and Boeing's (2004) study found generally high levels of agreement when examining people's personal medical histories through a mailed self-administered questionnaire, and, approximately 2 years later, a computer guided face-to-face interview. During both the questionnaire and interview, patients identified whether they had been formally diagnosed with the specific health problems listed and cited their age at diagnosis. In conditions of serious, blatant diseases, patients had very similar levels of agreement between methods (i.e., $\varkappa$>.83), whereas the more transient or less perceptible the illness, the more moderate the level of agreement (i.e., .39<$\varkappa$<.77). They found that participants were more likely to identify less serious or transient health conditions during an interview; it is purely speculative whether these were omitted on the questionnaire due to forgetfulness or because they were deemed unimportant.

Bergmann et al.'s (2004) high levels of agreement were likely due to several factors. First, the study dealt with a concrete construct of particular importance to the surveyed individuals. Second, participants were asked simple, closed questions; both questionnaires and interviews used similar schedules, leading to good alignment between instruments. Finally, this study used a large sample of normal adults, most of whom could be reasonably expected to remember their medical histories.

In contrast, Cutler, Wallace, and Haines's (1988) study of 2,571 British men and women's weekly alcohol consumption achieved low levels of agreement between participants' postal questionnaires and structured interviews with nurses. To judge people's levels of alcohol consumption, participants gave a detailed account of how many drinks they had consumed during the previous week and also responded to items on a quantity frequency scale. Results of the questionnaire and interview were used to classify the severity of the participants' alcoholism. Similarity of classification was used as the measure of method comparison. The HSQ questionnaire's level of agreement with the interview (depending on the group) varied in classification by 8 to 419% for men and 7 to 515% for women. The researchers concluded the questionnaire under-estimated consumption compared to interviews except for male excessive drinkers. While the structured interviews were treated as the gold standard for classifying alcoholism, no rates of actual consumption were available making it impossible to ascertain the accuracy of either response mode.

Like Bergman et al. (2004), this study had well aligned instruments (questionnaire and structured interview). However, the constructs being measured (rates and volumes of alcohol consumption) are inherently difficult to capture accurately. Discrepancy in the results was to be expected as, even within the interview, there were large differences between the quantity of alcohol reported consumed during the previous week and responses to quantity/frequency scales. The classification agreement within the interview between these two questions varied between 28-323% for males and 31-406% for females. Cutler et al. (1988) noted that people generally have difficulty remembering frequencies, without taking into account possible alcohol-caused brain damage and/or alcohol-induced memory lapses. Additionally, some people's drinking patterns and behaviours may be erratic and, hence, difficult, if not impossible, to measure accurately.

While the two previously discussed studies used structured interviews and questionnaires, Williams, Sobieszczyk, and Perez's (2001) study of pregnancy planning was one of very few that paired a questionnaire with a qualitative interview. While a large sample of Filipino men and women completed the questionnaire (*n*=780 men, *n*=1200 women), only 16 women and 10 men were interviewed. These data were coded systematically into three categories in relation to the pregnancy being discussed: intended, mistimed, and unwanted. There was overall agreement of 69%; female participants had 80% and men, 43% agreement. When discussing their results, Williams et al. (2001) noted that when they used dichotomous categorisations (intended/unintended), their agreement percentages improved. While there appears to be reasonable agreement, this study did not present a convincing argument for consistency between method due to the narrow categorisations of data and the lack of measures identifying how much agreement could be due to chance.

The studies reviewed show that conditions for agreement between survey and interview modes of data collection are complex. Highly-structured interview scoring systems aligned to questionnaire scoring (Bergman et al., 2004; Patton & Warning, 1991) seem to generally lead to higher consistency than semi-structured or open-ended interview categorisations (Holm, 1982; Rasmussen, Jensen, & Olesen, 1991), but not in all cases (Valentinis et al., 2009). It also appears that studies working with normal adult populations and those gathering data about relatively stable constructs have higher levels of agreement. Number of participants does not appear to be significant as both small and large samples have low and high cross-method agreement.

While the trends identified in this literature review may provide some guidelines as to how best to achieve the greatest levels of agreement between methods, there are problems associated with these criteria. There is certainly a need to investigate unstable constructs like people's conceptions, values, and beliefs and one's sample population cannot always be normal adults. Also, while structured interviews may give researchers a better chance at achieving 'reliability' between methods, they lack the 'uptake' questions which interviewers can use to inquire about participant reasons for responses (Antaki & Rapley, 1996). One cannot expect to get very 'different' kinds of data through structured interviews and questionnaires, making the entire interview exercise potentially redundant.

Study context

As few studies have examined the comparability between questionnaire and qualitative interview data in educational research, this study sought to extend our understanding of method effects. This study was able to take advantage of a previously validated questionnaire survey focused on teachers' conceptions of assessment. Previous research with New Zealand primary and secondary teachers (Brown, 2007; 2008) found that a survey questionnaire (i.e., Teachers' Conceptions of Assessment version III Abridged—TCoA-IIIA) simplified to four major conceptions of assessment. These were:

- Assessment improves teaching and learning (Improvement).
- Assessment makes students accountable for learning, partly through issuing certificates and credentials (Student Accountability).

- Assessment demonstrates the quality and accountability of schools and teachers (School Accountability).
- Assessment should be rejected because it is invalid, irrelevant, and negatively affects teachers, students, curriculum, and teaching (Irrelevance).

A semi-structured, qualitatively interpreted interview with 26 New Zealand teachers was used to examine the degree of similarity between methods. Given that the two methods were quite different in approach (i.e., surveys are more confirmatory, while interviews are more exploratory) and since there was little pre-planned structure to the interviews (unlike the survey), it was expected that the two different methods would paint quite different pictures. Similarities greater than chance, given the divergence of methods, would provide validation; but, such similarities were thought to be unlikely as the interviews were not tightly aligned to the questionnaire items. Thus, this study explored research questions around the similarities and differences of participant conceptions of assessment across questionnaire and semi-structured interview formats.

METHOD

Data Collection

This study was part of the Measuring Teachers' Assessment Practices (MTAP) project at The University of Auckland. The goal of the MTAP project is to explore the relationships among teachers' conceptions of assessment, teachers' assessment practices, students' conceptions of assessment, and students' academic outcomes. MTAP Study 1 gathered questionnaire and interview data from teachers to examine their conceptions of assessment and ascertain the degree to which methodological artefacts were impacting on results.

Participants included English and/or mathematics teachers of Years 5 to 10 (students nominally aged 9 to 14) from 40 primary, intermediate, and secondary schools in the greater Auckland area. These year levels were selected as both formal and informal assessments are utilised with these age groups, but such assessments are voluntary and school-based; externally moderated qualifications-related assessments do not begin in New Zealand until Year 11 (age 15). English and mathematics teachers were selected as these subjects are compulsory for all students and since curriculum and assessment improvement initiatives in New Zealand have focused on literacy and numeracy domains (see Brown, Irving, & Keegan, 2008 for descriptions).

Instruments

Initially, participating teachers voluntarily completed the 27 item TCoA-IIIA Inventory survey instrument (Brown, 2006) on teachers' conceptions of assessment. Of the 425 questionnaires distributed, 169 were returned (response rate = 40%). The inventory required teachers to respond using a six-point, positively packed, agreement rating scale (Brown, 2004) (i.e., strongly disagree, mostly disagree. slightly agree, moderately agree, mostly agree, and strongly agree). Participants rated statements including the following prompts:

- Assessment provides information on how well schools are doing
- Assessment places students into categories
- Assessment is a way to determine how much students have learned from teaching
- Teachers conduct assessments but make little use of the results

The inventory reduces to nine factors which aggregate into four major conceptions (i.e., assessment measures school quality, assessment grades students, assessment improves teaching and learning, and assessment is irrelevant). The improvement conception is composed of four of the nine factors (i.e., assessment describes abilities, assessment improves teaching, assessment improves student learning, and assessment is valid), while the irrelevance conception is comprised of three of the nine factors (i.e., assessment is bad, assessment is ignored, and assessment is inaccurate). In addition, teachers provided personal demographic details and information about their teaching careers. Further, they selected as many as they wished from a list of 12 assessment practices as the practices they associated with the term 'assessment'.

Of those returning questionnaires, 100 (59%) indicated willingness to be interviewed. The second author selected 26 of these participants for interview, trying to cover the widest range of conceptions profiles possible. The first author conducted the interviews without knowing on what basis each participant had been selected, creating double-blind data collection. The interviews were carried out over a two-week period, some 10 to 12 weeks after the questionnaires were completed. The interview was semi-structured, with the interviewer designing 'uptake' questions based on interviewees' responses. The core interview schedule included questions like the following:

1. Please give me an example of an assessment activity you used recently in your classroom.
2. Describe the purposes of the assessment activity you just described.
3. Can you give me examples of other classroom practices that you would consider to be assessment?
4. What do you think is the best way to assess student learning?
5. So overall, what do you see as the purpose of assessment?

In order to ensure that all participants directly addressed the four conceptions of the TCoA-IIIA, at the end of the semi-structured interview about assessment and its purposes, teachers were asked to indicate the extent to which they agreed or disagreed with four prompts taken directly from the questionnaire, with one prompt relating to each major conception. The prompts were:

1. Assessment keeps schools honest and up to scratch (School Accountability)
2. Assessment determines if a student meets a qualification standard (Student Accountability)
3. Assessment helps students improve their learning (Improvement)
4. Assessment is unfair to students (Irrelevance)

While these prompts were more structured than the previous questions, they were still delivered in a semi-structured way with the interviewer probing responses. All interviews were digitally recorded and transcribed verbatim by the first author. Each participant was assigned a pseudonym for analysis and reporting purposes; these pseudonyms are utilised throughout this paper.

Data Analysis

After data collection, the qualitative and quantitative data were analysed separately. Responses to the four conceptions of the TCoA-IIIA inventory were aggregated into profiles relative to national sample norms and compared to previous New Zealand sample populations (see Brown & Harris, 2009). Statistics for each conception of assessment factor were determined according to the previously established TCoA-IIIA factor patterns. Of 81 possible profiles (i.e., 3^4 profiles from 3 categories by 4 factors), 24 profiles were found among those willing to be interviewed. The 26 teachers selected for interview had 12 different profiles; four profiles had only one person, while eight profiles accounted for 22 people.

To ensure that a robust analysis of the interview data was conducted, the interview data were initially analysed by the first author using the phenomenographic method (Marton, 1981, 1986) to identify the range of conceptions within the data set; these results are available in Harris & Brown (2009). During this analysis, no pre-existing codes were utilised. Seven qualitatively different conceptions of assessment were found which, in general, aligned with the four conceptions of assessment (i.e., improvement, school accountability, student accountability, and irrelevance), the basis of the questionnaire's factor structure.

In order to compare the questionnaire and interview data sets, a second analysis of the interview data was conducted to code for the four conceptions of assessment categories central to the TCoA-IIIA questionnaire. This was done because it was impossible to superimpose the seven categories from the initial analysis of the interview data onto the quantitative questionnaire data set. Categorical analysis was used to identify passages relating to each of the four conceptions (Coffey & Atkinson, 1996). The first author classified each teacher for each conception using three levels of agreement (i.e., disagree, moderately agree, and strongly agree). To achieve this rating, the interviewees' responses to the explicit prompts during the interview were examined and rated. Data from other parts of the interview relevant to each conception were also coded, examined, and used to confirm or modify the initial classification.

In most cases, the participant's answer to the questionnaire prompt matched the first author's holistic judgment of their coded interview data. For example, Tom, a primary teacher, expressed concerns about the dubious reliability of many types of assessment and repeatedly commented on the subjectivity of grades and scores. When responding to the prompt "assessment determines if a student meets a qualification standard", Tom replied:

> *Any assessment, it depends on the day in which it's given, on the, on the environment which the student is coming from. It may be that the student can do the same assessment a week apart and get totally different results. I've also seen assessment, as I've said to you, that actually says a lot more about the person giving the assessment than the person being assessed, so no, I wouldn't agree.* (T1:114)

Tom was internally consistent and classified as 'disagree' for the student accountability conception.

However, in some cases a person's response to the explicit prompt was insufficient to classify his or her level of agreement towards a conception. For example, Alicia, an intermediate teacher, responded to the school accountability prompt by discussing how teachers often use euphemisms to soften negative results when communicating to parents on school reports, a response that was irrelevant to the construct. Elsewhere, Alicia talked extensively about the need for schools to be accountable, so she was rated as strongly agree for this conception based on the strength of other quotations like the one below:

> *Well, we need to be accountable at some point. I mean in reality we need to have assessment. We get paid by the state and we owe it. You can't work and not give evidence of it. It's hard with children because what do we give? … So assessment is a form of guiding your teaching in the classroom, but you also need assessment for reporting purposes. To give evidence. You have to give evidence; everything should be evidence based. The budget we get, how much support we get from the state. I mean we can't just be very subjective, to my mind. I'm very analytical. If we need x amount for this kind of resources, we need to give evidence based on sound assessment or sound standardised assessment to, um, support our argument or our request. So the purpose of assessment, for both reporting and to drive my teaching.* (A1:040)

Thus, the overall interview rating (i.e., disagree, moderately agree, or strongly agree) for each conception was a weighted composite of the explicit response to the conception prompt and the first authors' analysis of teacher's overall pattern of responses.

In order to enable comparison with the interview data, the questionnaire mean scores were converted to the same three levels of agreement used in the interview rating. Values 1.00 to 2.90 (strongly disagree to just shy of slightly agree) were deemed to be disagree; 2.91 to 4.49 (slightly agree to half-way between moderately and mostly agree) were called moderately agree; 4.50 to 6.00 (mostly and strongly agree) were classified strongly agree. Ratings for each conception were compared by examining consensus and consistency values. Validation of questionnaire responses is generally accepted if the percentage of identical agreement is 70% or better (Stemler, 2004) or the kappa coefficient is .60 or better (Cohen, 1960).

There were considerable differences between the two methods in this study that could lead to low levels of agreement. For example, there was a low level of content alignment, the design of the data collection methods was highly divergent, the data classification and reduction techniques were extremely different, the content was highly complex and conceptually abstract, the content was highly sensitive to individual-level contextual factors, and there was a time lapse between data collection points. Hence, it was reasonable to expect that the probability of highly consistent results was low, making data more likely to be complementary than consistent.

RESULTS

Appendix Table 2 provides categorised scores across methods and conceptions for each of the 26 interviewed teachers. There was an overall 57% identical agreement between methods using the three category rating scale. For nine of the twenty-six participants (35%) there were three or four identical agreements across the two methods. Only three cases had only one agreement and none had zero identical.

The identical agreement rate for the Irrelevance factor was 69% with $\varkappa$=-.13, while consistency ratings were r=-.16 and α=-.33. These indicated the two methods tended to be inverse to each other, the rate of consensus being less than would be found by chance. Careful inspection of the distribution of scores showed that only ratings 1 (Disagree) and 2 (Moderate) were used, and that the questionnaire method gave six of the eight differing responses a higher score than the interview method. This skewed distribution explains somewhat the inverse observed relationship. Improvement had an identical agreement rate of 58% ($\varkappa$=-.14), with consistency ratings of r =.15 and α=.26. Note that the interview method assigned none of the responses to rating 1, whereas the questionnaire identified four such responses. This lack of similarity in the distribution of scores reduces both the kappa and consistency ratings. These values paint a picture of weak positive consistency between methods but consensus less than would be expected by chance. School Accountability had 50% identical ratings ($\varkappa$=.13), with consistency ratings of *r* =.03 and α=.05. These indicate that the two methods are fundamentally independent measures since the consistency values were statistically not different to zero. Student Accountability ratings were 50% identical (ϰ=.00) across methods, while consistency ratings were *r* =.25 and α=.40. These indicated moderate positive consistency between methods but consensus no better than by chance.

It appears that there was an inverse relationship between improvement and irrelevance conceptions in that there were no disagree and no strongly agree ratings respectively. This result aligns with the zero correlation between these two scales in the questionnaire. In the interview there was a generally negative position relative to the student accountability conception, a result that was also picked up by the questionnaire. While this result contrasts to previous studies of New Zealand teachers (Brown, 2006, 2007), it would seem that the inter-relationship between some of the constructs across the two sources of data provided a somewhat similar picture.

Nevertheless, in a number of cases, as indicated by the exact agreement rates, participant interview data accurately reflected their responses on the questionnaire. Examples of some of these perfect matches, taken from the 'assessment makes schools accountable' category, are found in Table 1. However, there were cases where the interview data did not appear to align with the person's

Table 1: Examples of participant interview data that matched questionnaire responses to the conception "assessment makes schools accountable".

Strongly agree Chelsea

C1:070 That was in your questionnaire wasn't it? Um it's interesting. I think it does. Yes I do agree with it because um, especially let's speak from the point of view of this being a private school. Parents want to know how well these boys are doing when they go on to secondary school, um, another private school. They might go to the state system. They might go to a private system, but from here and beyond and if we're not up front and honest about our assessment practices and what's going on and we're actually churning out some pretty good kids who aren't just your boy swots, if you get what I mean. They're not just regurgitating the whole pile of facts, but they're actually well rounded learners, which is part of their assessment process, then yes, I do agree with that statement, yup.

Strongly agree Ju-long

J1:118 Yes, I would agree with that. Because um, I mean the assessment has got to be seen to be done. It's got to be, so that the school has got something to compare with the, the national mean. So that you know if we do a pencil and paper test, this is where we all should be.

Strongly agree Oliver

O1:108 Overall I agree, yep, I think parents and the community tend to look at us, tend to look at assessment, you know, that's one way they can get a window on the school. Obviously that's not always the be all and end all, but um, yeah, I guess overall I'd agree that it helps keep track of what is happening, and where thing are heading.

Moderately agree Kuval

K1:094 Honest and up to scratch. Mm. Well to a point, to a point. When you say honest, are they being honest in using the information for the purpose that it was set down? They're probably trying to do that, and not that they want to deceive, but then time, and probably we need to know a little more about how we can use the assessments and what's the other thing you mentioned?

K1:095 Honest and up to scratch.

K1:096 Up to scratch. Well, I think we need to work on that a little more. I don't think it's up to scratch at the moment.

Moderately agree Isabel

I1:118 Honest and up to scratch. I agree with it keeps them up to scratch, because as I was saying before, you know, it fuels my teaching, so it's going to keep you on top of your game if you know where your kids are and where they need to go. What was the first part? Honest?

I1:119 Honest.

I1:120 Um, I guess in a sense honest, but some people can fake assessment and just say, 'Oh, okay, yup Bob's getting it, yeah Jim's getting it' kind of thing and just winging it kind of thing. I know teachers who have done that too. So I'm not necessarily, I wouldn't necessarily agree with keeping it honest, but I would say keeping it up to scratch. Yeah, I would agree with that.

questionnaire responses at all. For example, while Emma agreed strongly to the school accountability conception on the questionnaire, within the interview she consistently argued that assessment was not a valid and reliable measure of school quality. She questioned the reliability of the formal National Certificate of Educational Achievement (NCEA) assessments she was required to use for reporting results of her Year 11 and 12 students at her secondary school:

> Emma: *I think there should be less of a focus on … what the end result is, but how they got there. And the process. I don't know if there's enough of a focus on that. I just don't totally agree with some of the NCEA business.*
>
> Interviewer: *Why is that?*
>
> Emma: *I just think it's too easy to cheat. Far too easy and it's just I find it a bit of a mess, I don't know.*
>
> …
>
> Emma: *I just don't feel that there's an overall structure [for NCEA], like nation wide. There's obviously supposed to be one nation wide, but I think it's too variable between schools and schools can, there's a lot of schools out there that sort of adjust it so the results end up fitting them.* (E1:122-E1:128)

Statements like these show her expressing scepticism that the current NCEA assessment was a good way to measure school quality, yet she had strongly agreed with this conception on the questionnaire. Perhaps when responding to the questionnaire items, she was considering systems other than the NCEA. While there were few cases where the differences between methods were this extreme, this example reinforces the dissimilarity of results attributable to a method effect.

Thus, the modest observed level of agreement is best understood as coincidental. The consensus level is less than would be expected by chance and consistency ratings were inverse for one scale, zero for a second, and weakly positive for two scales. It should be remembered that characteristics of the distributions of the ratings contributed to artificially depressing the values of the comparative statistics. Nonetheless, it is difficult to conclude that the structured questionnaire and the semi-structured qualitative interview classify participant conceptions of assessment in a similar fashion.

DISCUSSION

In this study, the highly structured questionnaire method and the open-ended, semi-structured interview provided only limited and weak evidence of consistency or consensus between methods. This negative result, albeit consistent with some of the previously reviewed literature, should give pause to those using qualitative data to 'support' quantitative results. The results also add weight to arguments put forward by authors like Kendall (2008) that qualitative data should not be used to illustrate quantitative results without first being analysed in their own right using techniques appropriate for the type of data collected.

Comparing questionnaires and interviews proved challenging and these difficulties may explain why only one study was found that had previously attempted to quantify similarities between these methods within an educational research context (Villanova, 1984). First, there was the issue of consistency within the methods. While the questionnaire generally took participants approximately 20 minutes to complete, the interviews usually lasted for about an hour, giving more time to expose the variabilities and inconsistencies within human thinking (Marton & Pong, 2005; Pajares, 1992). This variability made it difficult to classify some participants' attitudes towards assessment.

This was especially the case with the improvement and irrelevance conceptions which, within the questionnaire, are complex hierarchical structures containing four and three sub-factors respectively. The questionnaire mean scores for these two factors averaged out multiple constructs which may separately trigger different attitudes. While most teachers said that assessment was for improvement purposes, there were many that expressed strong distrust of external or standardised assessment measures (or testing in general) as methods which were unreliable or could have negative affective results for students. The improvement factor included items relating to assessment describing student ability, a process often done through testing. Thus, the complexity of this construct made it difficult to rate the teacher's level of agreement towards it. For example, Xavier expressed strong views that assessment's primary purpose should be to improve teaching and learning:

> *If things aren't done in a school for teaching and learning, then there's no point because that's what a school is about. So if assessment's not either to influence and better my teaching practice or to help the child learn, then there's no point.* (X1:094)

However, he also said he was extremely against testing. He described a cartoon he had seen recently, using it to illustrate how most tests are unfair and inaccurate:

> *It's a cartoon where it's got a monkey and a dog, and an elephant, and a seal and then it say;, there's a teacher at a desk that says, "Okay, here's our standardised test. Climb that tree.…" And that's the way I see assessment as well. We've got 26% of our, our largest population group here is Samoan and unless testing can be individualised. So I'm not a fan of standardised testing because we get taught so much that we need to teach different ways because kids learn differently, but we can't assess them. Well, standardised testing is my problem…* (X1:030, X1:032)

This complex interrelationship of factors made reducing teacher thinking to a single value difficult.

A further confound was the lack of measurement precision when reducing teacher thinking to a single, comparable value for each factor within each method. While the questionnaire's design incorporated a continuous scale anchored at six points to determine a participant's score, the interview required a complex holistic judgment to reduce teacher thinking to a three-point scale. The inter-method comparison required further reduction of the questionnaire scale from six to just three points. The reduction of questionnaire and interview data both introduced substantial margins of error. Also, participant responses directly created their questionnaire score, while the scores given to their interview responses were mediated through the analyst's judgments. Thus, there was potentially a greater distance between the participant and the score for the interview data.

An additional measurement issue became apparent when calculating comparative statistics (i.e., Cohen's kappa, Cronbach's alpha, Pearson's *r*) to identify the level of similarity between the methods taking chance into consideration. All of these calculations produced extremely low or negative values despite a reasonable number of exact matches in some classifications (e.g., irrelevance at 69% identical agreement). However, the data distributions had low variation and unequal distributions, both of which are known to depress these statistics (Orwin, 1994). Based on this insight, some of the studies examined in the literature review may have been more robust than analysts had thought. A reanalysis of the data distribution in these studies is beyond the scope of this paper, but seems warranted. This study illustrates the importance of not only calculating comparison statistics, but also inspecting the distribution of codings before concluding agreement was not found.

The particular results of this study also highlight issues related more generally to the differences in these methods. First, within this study, there was evidence that interview data were highly contextualised (Fontana & Frey, 2000), reflecting personal responses given at a particular point and time within a somewhat contrived interaction (Lankshear & Knobel, 2004). Some interview responses seemed highly influenced by the respondent's own context, the school sector being considered (i.e., secondary school versus primary school) or the type of assessment being discussed (i.e., formal test-like practices versus interactive practices). For example, when responding to the school accountability prompt "assessment keeps schools honest and up to scratch", Rebecca, an intermediate teacher, indicated that she would respond differently to the prompt if she were working at a secondary school. Altogether, four teachers rejected the student accountability conception within the interview simply because there were no 'qualifications' at their particular year level. While it is impossible to know whether participants were, in fact, more dispassionate and considered when completing the questionnaire, it is possible that the difference in response patterns between the two methods may have occurred, in part, because the more generic questionnaire format allowed teachers to think of concepts more abstractly rather than grounding them in personal experiences.

These data also highlighted the difference in coverage between the two instruments. While the highly structured questionnaire led participants to address all topics more-or-less equally, within the interview setting participants were freer to speak to or ignore topics as they chose. For example, without prompting, this group of participants rarely talked about assessment as a means of student accountability, instead mainly discussing assessment as improvement. Thus, there were far less data to use when making decisions about their attitudes towards the accountability conception. It is difficult to determine why participants focused so heavily on the improvement conception. While it is likely that this pattern of responding reflects strong beliefs in assessment for improvement, it is also possible that some participants centred discussions around it because of Ministry of Education initiatives that have actively promoted this way of thinking (Ministry of Education, 2001). Hence, these results may be a function of interviewees' desire to respond in a socially desirable way.

This study also showed that some interviewees misunderstood prompts taken from the questionnaire. This raises the possibility that items were also misunderstood in the questionnaire method. For example, multiple respondents took the interview's school accountability prompt to be about school and teacher honesty when reporting rather than, as intended, that assessment is a means of evaluating school quality. In such cases, teachers often responded to the prompt in one of two ways. Three teachers found it insulting because they thought it implied that schools and teachers were dishonest. For example, Quinn said, "*It's a derogatory statement; it assumes that schools would not be honest*" (Q1:236). Twelve participants used it as a chance to talk about how heightened school accountability could or would encourage teachers and school administrators to manipulate assessment in inappropriate ways. These included teaching to the test, excluding children from tests in order to get better results, and selectively and inaccurately reporting data. These teachers' alternative interpretations of this prompt made it difficult to categorise their attitudes and highlighted potential misinterpretations and misunderstandings participants may experience when completing questionnaires.

It appears that the observed similarities between how the teachers responded to the questionnaire and the interview are modest at best and that the two methods did not validate each other. A number of possible explanations for the discrepancy between the methods have been offered. This study has shown the potential hazards of assuming that questionnaire and interview data should be similar simply because they came from the same participants. This study does not invalidate the TCoA-IIIA inventory which has been successfully used with multiple New Zealand and international samples (Brown, 2007; Brown & Lake, 2006; Brown, Kennedy, Fok, Chan, & Yu, 2009), nor does it present the parallel qualitative interview results as flawed or unreliable. Instead, it suggests method effects resulting from instrument design, participant responses, and analytical processes may cause these data to say different things. The results from these two methods (i.e., survey questionnaire and semi-structured, qualitative interview) should be considered not so much as confirmatory or divergent, but rather as complementary (Smith, 2006).

These data raise the question of how researchers can and should deal with two valid yet differing sets of data from the same participants. Kendall (2008) noted that frequently qualitative results in mixed method studies are glossed over, with these data forced into preconceived questionnaire categories, hiding or exacerbating flaws in the original quantitative instrument. Such a use of qualitative data would be abusive because interview data may carry different messages than questionnaire data. Instead, both questionnaire and interview data sets should be analysed separately using methods suitable to each; then results can be compared to see if any common messages resonate from both sets of data. This study and review illustrate the point Smith (2006, p. 465) made: "triangulation attempts to confirm inferences made from the findings of several research methods and approaches. However, triangulation is less a method than a troublesome metaphor". It would be useful for future studies to compare other types of questionnaires (e.g., open-ended questions on questionnaires) and interviews (e.g., focus group interviews) to examine the extent to which these data are confirmatory or complementary. People's focus group interview responses could potentially be even more divergent from their questionnaire answers than this study suggests as within focus groups, other participants' dialogue could influence what they say.

Nonetheless, based on the literature reviewed and research experience, the following recommendations are made to assist researchers in education and social sciences who are trying to maximise the likelihood that their questionnaire and interview data will align:

1. Ensure interview prompts and questionnaire items are structured and highly similar.
2. Separate data collection by only a short period of time.
3. Present the object of interest in a highly concrete and specific way.
4. Anchor participant responses to a common context.
5. Focus on psychological objects that have simple internal structure (i.e., avoid hierarchical, complex structures).
6. Estimate agreement between methods, albeit cautiously in light of data distributions, using consensus and consistency procedures.

While following these guidelines might increase the chances of similarity between the data sets, researchers should realise that this alignment comes at a cost. The main attraction of using mixed method research is that data gained through different methods may complement each other, overcoming weaknesses in individual methods. Pairing structured interviews with structured questionnaires would be unlikely to create this methodological richness (Antaki & Rapley, 1996). The challenge is now for mixed method researchers to demonstrate that triangulation by distinctly different methods can lead to confirmation and explain the circumstances that allow this to occur. This study and review suggest that methodological artefacts prevent such claims and that at best complementary but distinct results will arise.

References

Antaki, C., & Rapley, M. (1996). Questions and answers to psychological assessment schedules: Hidden troubles in 'quality of life' interviews. *Journal of Intellectual Disability Research, 40*(5), 421-437.

Bergmann, M. M., Jacobs, E. J., Hoffmann, K., & Boeing, H. (2004). Agreement of self-reported medical history: Comparison of an in-person interview with a self-administered questionnaire. *European Journal of Epidemiology, 19*, 411-416.

Boniface, D. R., & Burchell, H. (2000). Investigation of validity of closed questions in a survey of British South Asian and White populations. *Ethnicity & Health, 5*(1), 59-65.

Brewer, N. T., Hallman, W. K., Fielder, N., & Kipen, H. M. (2004). Why do people report better health by phone than by mail? *Medical Care, 42*(9), 875-883.

Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement In Education, 16*(1), 27-54.

Brown, G. T. L. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94,* 1015-1024.

Brown, G. T. L. (2006). Teachers' conceptions of assessment: Validation of an abridged instrument. *Psychological Reports, 99,* 166-17.

Brown, G. T. L. (2007 December). *Teachers' Conceptions of Assessment: Comparing Measurement Models for Primary & Secondary Teachers in New Zealand.* Paper presented to the New Zealand Association for Research in Education (NZARE) annual conference, December, 2007, Christchurch, New Zealand.

Brown, G. T. L. (2008). *Conceptions of assessment: Understanding what assessment means to teachers and students.* New York: Nova Science Publishers.

Brown, G. T. L., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of Multidisciplinary Evaluation, 6*(12), 68-91.

Brown, G. T. L., Irving, S. E., & Keegan, P. J. (2007). *An Introduction to Educational Assessment, Measurement, and Evaluation: Improving the Quality of Teacher-Based Assessment.* Auckland, NZ: Pearson Education NZ.

Brown, G. T. L., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy and Practice, 16*(3), 347-363.

Bryman, A. (2008). *Social research methods.* Oxford: Oxford University Press.

Coffey, A., & Atkinson, P. (1996). *Making sense of qualitative data: Complementary research strategies.* London: Sage.

Cohen, J. (1960). A coefficient for agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cutler, S. F., Wallace, P. G., & Haines, A. P. (1988). Assessing alcohol consumption in general practice patients - A comparison between questionnaire and interview (Findings of the Medical Research Council's General Practice Research Framework Study on Lifestyle and Health). *Alcohol & Alcoholism, 23*(6), 441-450.

Daving, Y., Claesson, L., & Sunnerhagen, K. S. (2009). Agreement in activities of daily living performance after stroke in a postal questionnaire and interview of community-living persons. *Acta Neurologica Scandinavica, 119*(6), 390-396.

Day, C., Sammons, P., & Gu, Q. (2008). Combining qualitative and quantitative methods in research on teachers' lives, work, and effectiveness: From integration to synergy. *Educational Researcher, 37*(6), 330-342.

Dodd, S., Williams, L. J., Jacka, F. N., Pasco, J. A., Bjerkeset, O., & Berk, M. (2009). Reliability of the Mood Disorder Questionnaire: comparison with the Structured Clinical Interview for the DSM-IV-TR in a population sample. *Australian and New Zealand Journal of Psychiatry, 43*(6), 526 - 530.

Esler, D., Johnston, F., & Davis, D. T. B. (2008). The validity of a depression screening tool modified for use with Aboriginal and Torres Strait Islander people. *Australian and New Zealand Journal of Public Health, 32*(4), 317-321.

Fontana, A., & Frey, J. H. (2000). The interview: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 645-672). London: Sage.

Harris, L. R., & Brown, G. T. L. (2009). The complexity of teachers' conceptions of assessment: Tensions between the needs of schools and students. *Assessment in Education: Principles, Policy and Practice, 16*(3), 379-395.

Holm, N. G. (1982). Mysticism and Intense Experiences. *Journal for the Scientific Study of Religion, 21*(3), 268.

Hunsberger, B., & Ennis, J. (1982). Experimenter Effects in Studies of Religious Attitudes. *Journal for the Scientific Study of Religion, 21*(2), 131.

Kendall, L. (2008). The conduct of qualitative interview: Research questions, methodological issues, and researching online. In J. Coiro, M. Knobel, C. Lankshear & D. Leu (Eds.), *Handbook of research on new literacies* (pp. 133-149). New York: Lawrence Erlbaum Associates.

Kooiman, C. G., Ouwehand, A. W., & ter Kuile, M. M. (2002). The Sexual and Physical Abuse Questionnaire (SPAQ): A Screening Instrument for Adults to Assess Past and Current Experiences of Abuse. *Child Abuse & Neglect: The International Journal, 26*(9), 939-953.

Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice, 27*(2), 28-45.

Lankshear, C., & Knobel, M. (2004). *A handbook for teacher research*. Berkshire: Open University Press.

Lemery-Chalfant, K., Schreiber, J. E., Schmidt, N. L., Van Hulle, C. A., Essex, M. J., & Goldsmith, H. H. (2007). Assessing Internalizing, Externalizing, and Attention Problems in Young Children: Validation of the MacArthur HBQ. *Journal of the American Academy of Child & Adolescent Psychiatry, 46*(10), 1315-1323.

Manassis, K., Owens, M., Adam, K. S., West, M., & Sheldon-Keller, A. E. (1999). Assessing attachment: Convergent validity of the Adult Attachment Interview and the Parental Bonding Instrument. *Australian and New Zealand Journal of Psychiatry, 33*(4), 559-567.

Marton, F. (1981). Phenomenography - Describing conceptions of the world around us. *Instructional Science, 10*, 177-200.

Marton, F. (1986). Phenomenography- A research approach to investigating different understandings of reality. *Journal of Thought, 21*(3), 28-49.

Marton, F., & Pong, W. Y. (2005). On the unit of description in phenomenography. *Higher Education Research and Development, 24*(4), 335-348.

Ministry of Education. (2001). Developing teachers' assessment literacy. *Curriculum Update, 47*, Available online: http://www.tki.org.nz/r/governance/curric_updates/curr_update47_49_e.php .

Oei, T. I., & Zwart, F. M. (1986). The assessment of life events: Self-administered questionnaire versus interview. *Journal of Affective Disorders, 10*(3), 185-190.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing, and attitude measurement*. New York City: St. Martin's Press.

Orwin, R. G. (1994). Evaluating coding decisions. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 139-162). New York: Russell Sage Foundation.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Patton, D., & Waring, E. M. (1991). Criterion validity of two methods of evaluating marital relationships. *Journal of Sex & Marital Therapy, 17*(1), 22-26.

Rasmussen, B. K., Jensen, R., & Olesen, J. (1991). Questionnaire versus clinical interview in the diagnosis of headache. *Headache, 31*, 290-295.

Reams, P., & Twale, D. (2008). The promise of mixed methods: Discovering conflicting realities in the data. *International Journal of Research and Method in Education, 31*(2), 133-142.

Richman, W., Keisler, S., Weisband, S., & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology, 84*(5), 754-775.

Rojahn, J., Warren, V. J., & Ohringer, S. (1994). A comparison of assessment methods for depression in mental retardation. *Journal of Autism and Developmental Disorders, 24*(3), 305-313.

Ryan, G., & Bernard, H. R. (2000). Data management and analysis methods. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 769-802). London: Sage.

Silverman, D. (2000). Analysing talk and text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (2nd ed., pp. 821-834). London: Sage.

Silverman, D. (2006). *Interpreting qualitative data* (3rd ed.). London: Sage

Smith, M. L. (2006). Multiple methods in education research. In J. Green, G. Camilli & P. Elmore (Eds.), *Handbook of complementary methods in educational research* (pp. 457-475). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), Available online: http://pareonline.net/getvn.asp?v=9&n=4.

Valentinis, L., Valent, F., Mucchiut, M., Barbone, F., Bergonzi, P., & Zanchin, G. (2009). Migraine in Adolescents: Validation of a Screening Questionnaire. *Headache: The Journal of Head and Face Pain, 49*(2), 202-211.

Villanova, R. M. (1984, April). *A Comparison of Interview and Questionnaire Techniques Used in the Connecticut School Effectiveness Project: A Report of Work in Progress*. Paper presented at the Annual Meeting of the American Educational Research Association.

Williams, L., Sobieszczyk, T., & Perez, A. E. (2001). Consistency between survey and interview data concerning pregnancy wantedness in the Philippines. *Studies in family planning, 32*(3), 244-253.

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Los Angeles: Sage Publications.

Appendix

Appendix Table 1- Agreement in studies comparing questionnaire and interview data

Authors	Subject of study	Participants	Study design	Agreement
Correlational analyses: Kappa				
Bergmann et al. (2004)	Personal medical history about former illnesses and their age at diagnosis	7,841 male and female residents of Potsdam	Baseline face-to-face computer guided interview, then, approximately 2 years later, a follow-up self-administered questionnaire.	ϰ=.83-.88 for diseases like diabetes, cancer, etc.; ϰ =.68-.77 for gout, hypertension, etc.; ϰ =.39-.59 for rheumatism, IBS, stomach ulcers, etc.
Kooiman et al. (2002)*	Physical and sexual abuse	134 patients	Questionnaire at home, then, during a course of treatment, a structured interview about physical and sexual abuse.	ϰ=.71 for incidents of sexual abuse; ϰ=.59-.62 for incidents of physical abuse
Valentinis et al. (2009)	Teenage migraine/headache sufferers symptoms	93 students aged 14-19 from north Italian schools	Questionnaire, then, approximately 4 weeks later, an extensive semi-structured interview and physical examination with a neurologist.	ϰ =.66
Daving et al. (2009)	Activities associated with daily living for stroke patients	32 people who had experienced a stroke ten or more years in the past	Self-administered, postal Activities of Daily Living (ADL) questionnaire and then, 1-2 weeks later, a follow-up interview in home or at a medical clinic.	ϰ =-.48-.82 for physical items; ϰ =.36-.80 for independent activity measures; ϰ =.26-.70 for cognitive and social items
Rojahn et al. (1994)*	Depression in patients with mild or moderate mental retardation	38 adults (½ with high and ½ with low depression scores)	Program managers completed a questionnaire about the patients, then repeated it again several weeks later. Patients responded to read-aloud questionnaires followed by an interview with a psychiatrist.	ϰ =.33 agreement on repeated questionnaire; On other measures ϰ <.10
Lemery-Chalfant et al. (2007)	Children's attention problems	Mothers of 814 twin children who were 8 years old	Mothers completed an over-the-phone questionnaire (HBQ) for each twin. Six months later during a home visit, they completed a structured clinical interview (DISC).	ϰ =.27-.42 between DISC and HBQ
Dodd et al. (2009)*	Diagnosing bi-polar disorder in an osteoporosis study	1066 women	Mood Disorder Questionnaire (MDQ) administered, followed by a structural clinical interview.	ϰ =.25; The MDQ only correctly identified 6 patients, missing all cases of bi-polar II (*n*=11) and *n*=7 bi-polar I.
Rasmussen et al. (1991)*	Migraine headaches	713 Danish adults	Questionnaire administered, followed by a clinical semi-structured interview and an examination by a medical practitioner.	ϰ = .24, .30, .43 for different types of headaches
Correlational analyses: Pearson				
Patton & Waring	Marital intimacy in	25 husband	Separately, husbands and wives	*r* = .88

Appendix Table 1- Agreement in studies comparing questionnaire and interview data

Authors	Subject of study	Participants	Study design	Agreement
(1991)*	the context of marital and drug therapy for depression	and wife pairs	completed questionnaires and then a structured interview.	
Villanova (1984)*	School effectiveness	247 school employees	67 item interview, followed by a 100 item questionnaire.	Multi-trait/ multimethod analysis *r* range = .55-.93
Manassis et al. (1999)	Parental attachment	130 emotionally or behaviourally disturbed adolescents	Questionnaire administered, then, 2 weeks later, adolescents took part in a structured interview.	Mean *r* = .26 (*SD*=.22)
Holm (1982)*	Religious beliefs	122 Swedish speaking Finnish adults	An open interview, questionnaire, and two tests administered to participants in alternating order.	Range *r* =.08-.32
Classification Agreement				
Esler et al. (2008)	Depressions in patients with heart disease	34 Torres Strait Islander or Aboriginals	Modified PHQ-9 questionnaire orally administered, then, within 2 days, participants took part in a semi-structured, culturally sensitive, clinical interview.	Agreement levels for classification as depressed ranged 59% to 88%
Williams et al. (2001)	Planned pregnancy	10 men and 16 women (n=26), all married	Survey questionnaire administered, then, 2-3 months later, in-depth interviews conducted.	Consistency between methods: Women = 80% Men = 43%.
Cutler et al. (1988)*	Alcohol consumption	2571 British men and women	Questionnaire administered, followed by a structured interview with a nurse.	Agreement levels for classification as alcoholic varied by 8 to 419% for men and 7 to 515% for women
Miscellaneous				
Boniface & Burchall (2000)	Patient views of inpatient experiences in hospitals	15 white and 14 South Asian hospital patients	Two part interview: ½ open questions, ½ closed questions based on items taken from a questionnaire	29% of closed question variance predicted by open question responses
Richman et al. (1999)	Social desirability distortion in computer, interview, and pen and paper questionnaire methods	61 studies conducted between 1967 and 1997	Meta-analysis using hierarchical regression.	Effect size between computer and pen and paper instruments *M*= .05; Effect size between computerised and face-to-face interviews *M* = -.19

Appendix Table 1- Agreement in studies comparing questionnaire and interview data

Authors	Subject of study	Participants	Study design	Agreement
Hunsberger & Ennis (1982)*	Religious beliefs	126 1st year sociology students	All interviewed by the same person who half the time claimed to be a "minister" and the other half, a "professor". They then completed a questionnaire.	ANOVA found only 1 statistically significant difference (i.e., Social Desirability Scale) questionnaire respondents (*M*=19.6) > interviewees (*M*=14.9).
Brewer et al. (2004)	Social desirability in health assessment	261 American Gulf War Veterans	Questionnaire administered, then, 2 weeks later, they took part in a computer assisted telephone interview.	Interview identified more symptoms (*n*= 51); questionnaire identified more severe symptoms (*n*=86).

Note. *=No time frame given between questionnaire and interview

Appendix Table 2: Interviewee Conceptions of Assessment Profiles

	TCoA-IIIA Conceptions Mean Scores				Interview Conceptions Agreement Rating			
ID	Student accountability	School accountability	Improvement	Irrelevant	Student accountability	School accountability	Improvement	Irrelevant
30 Rebecca	Moderate	Strong	Moderate	Moderate	Moderate	Strong	Strong	Moderate
154 Alicia	Moderate	Strong	Moderate	Moderate	Strong	Strong	Strong	Disagree
16 Lisa	Moderate	Strong	Disagree	Moderate	Disagree	Moderate	Moderate	Moderate
127 Oliver	Moderate	Strong	Moderate	Moderate	Moderate	Strong	Moderate	Moderate
49 Chelsea	Moderate	Strong	Moderate	Moderate	Moderate	Strong	Strong	disagree
13 Wynona	Moderate	Strong	Moderate	Moderate	Strong	Strong	Moderate	Moderate
23 Ursula	Moderate	Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Disagree
34 Ju-long	Moderate	Strong	Moderate	Moderate	Moderate	Strong	Moderate	Moderate
124 Quinn	Moderate	Moderate	Moderate	Moderate	Strong	Moderate	Moderate	Moderate
133 Grace	Moderate	Strong	Strong	Moderate	Moderate	Moderate	Moderate	Moderate
19 Danielle	Strong	Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate

Appendix Table 2: Interviewee Conceptions of Assessment Profiles

	TCoA-IIIA Conceptions Mean Scores				Interview Conceptions Agreement Rating			
ID	Student accountability	School accountability	Improvement	Irrelevant	Student accountability	School accountability	Improvement	Irrelevant
3 Sylvia	Moderate	Strong	Moderate	Disagree	Moderate	Moderate	Moderate	Moderate
20 Pearl	Disagree	Strong	Disagree	Moderate	Moderate	Moderate	Moderate	Moderate
134 Xavier	Disagree	Strong	Disagree	Moderate	Disagree	Disagree	Moderate	Moderate
50 Henry	Disagree	Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
78 Emma	Disagree	Strong	Moderate	Moderate	Moderate	Disagree	Moderate	Moderate
43 Isabel	Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
151 Fred	Strong	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate
41 Madison	Moderate	Moderate	Moderate	Moderate	Moderate	Moderate	Strong	Disagree
141 Yvonne	Moderate	Moderate	Moderate	Moderate	Moderate	Strong	Strong	Moderate
150 Kuval	Disagree	Moderate	Moderate	Moderate	Moderate	Moderate	Strong	Moderate
17 Tom	Disagree	Moderate	Disagree	Moderate	Disagree	Moderate	Moderate	Moderate
155 Vince	Moderate	Moderate	Moderate	Moderate	Moderate	Strong	Moderate	Moderate
40 Nicole	Strong	Moderate	Moderate	Disagree	Moderate	Moderate	Moderate	Moderate
104 Zac	Moderate	Moderate	Moderate	Moderate	Strong	Strong	Moderate	Disagree
99 Bimala	Moderate	Disagree	Moderate	Moderate	Moderate	Moderate	Moderate	Disagree

Note: Yellow highlighting indicates identical agreement between methods.

Acknowledgement

This study was completed in part through the financial support of a University of Auckland Faculty of Education Faculty Research Development Fund grant (#3609273/9215) to the Measuring Teachers' Assessment Practices project. The cooperation and participation of the teachers and schools in the MTAP project is appreciated. An earlier version of this paper was presented at the New Zealand Association for Research in Education (NZARE) Annual Conference November 2008, Palmerston North, New Zealand.

Citation

Harris, Lois R. & Brown, Gavin T.L. (2010). Mixing interview and questionnaire methods: Practical problems in aligning data . *Practical Assessment, Research & Evaluation*, 15(1). Available online: http://pareonline.net/getvn.asp?v=15&n=1.

Corresponding Author

Correspondence concerning this paper should be addressed to Lois R. Harris, School of Teaching, Learning, and Development, University of Auckland, Private Bag 92019, Auckland, 1142, New Zealand, or by email to mailto:lois.harris [at] auckland.ac.nz.