JOEL TAYLOR
Centre for Sustainable Heritage
Bartlett School of Graduate Studies
University College London
London, UK
joel.taylor@ucl.ac.uk

INTRA-SURVEYOR BIAS IN COLLECTION CONDITION SURVEYS

Keywords: condition survey, reliability, bias, intra-surveyor

ABSTRACT

This paper examines intra-surveyor agreement (the extent to which people agree with their own judgments over time) in collection condition surveys specifically in terms of relevance to several conservation practices. The argument is based on the fact that the consequences of poor agreement are inaccurate assessments that may lead to unrecognized problems or misallocation of resources. The experiment in question is concerned with two surveys carried out with an interval of ten weeks by a group of professional conservators. The group carried out two surveys of the same museum objects using a published survey form. Reliability was calculated and it was determined that reliability ranged from very high levels to levels close to chance agreement. Various factors were examined, such as experience, definitions, kinds of objects and the impact of reducing expert judgment to categories and a small range of grades available in assessment forms. It was concluded that reliability cannot be assumed when one surveyor is used, and further, that reliability should be measured in pilot studies regardless of the number of surveyors.

RÉSUMÉ

Cet article examine l'accord interne des évaluateurs (dans quelle mesure les gens sont d'accord avec leur propre jugement au fil du temps) dans les évaluations sur l'état des collections, notamment en termes de pertinence vis-à-vis des pratiques de conservation-restauration. L'argument se fonde sur le constat qu'une faible persévérance dans le jugement a pour conséquence des évaluations inexactes, d'où découlent des problèmes non décelés ou une mauvaise attribution des ressources. L'expérience en question porte sur deux évaluations menées à dix semaines d'in-

INTRODUCTION

Surveying the condition of historic collections is a fundamental part of many conservation management practices, both preventive and interventive. Tick box condition assessments of collections have been used for a variety of reasons, such as examining symptoms of deterioration or inferring treatment needs and causes of deterioration, in addition to determining collections management objectives (Taylor and Stevenson 1999).

Condition surveys categorise different symptoms of damage in order to draw broad conclusions about a collection's state. However, physical examination alone does not reveal all one needs to know about condition (Taylor 2005, Appelbaum 2007). In fact, condition data cannot be used effectively in isolation.

Collection condition surveys require the observation, usually for short periods of time, of a large number of objects to be represented through categories and scores. This limits the detail that can be gathered but allows an overview of large numbers of items. The problems of subjective assessment have been documented by comparing different peoples' assessments of the same objects with the same categories (Taylor 1999, 2009). The implications can be serious for any heritage institution, such as uninformative or misleading data used to make collections management decisions that could lead to increased deterioration, resources spent needlessly on control, or possibly data being questioned by skeptical readers. However, despite criticisms of visual examination for condition (Taylor 1996, 1999, 2005, Henderson 1997), condition surveys are still an important method for the assessment of collections. Even with the development of scientific instrumentation to examine material properties in detail, and predictive methods for informing preventive conservation strategy, visual assessment of deterioration is still necessary to understand collection deterioration (Ashley-Smith 1999, Taylor 2005).

Although variation between different surveyors (i.e. inter-surveyor reliability) in condition assessment is an acknowledged problem (Keene 1991, Taylor 1999), which is sometimes addressed in practice (e.g. Johnsen 1999), variation is also possible when only one person is carrying out a survey. This paper investigates intra-surveyor reliability and, through quantifying reliability, offers some insight into its causes and qualities. What

tervalle par un groupe de restaurateurs professionnels. Le groupe a fait deux évaluations des mêmes objets de musées en utilisant un formulaire d'évaluation publié. Après calcul, il a été déterminé que la fiabilité variait entre des niveaux très élevés et des niveaux proches d'un accord totalement fortuit. Divers facteurs ont été examinés, comme l'expérience, les définitions, les types d'objets et l'effet de la restriction du jugement du spécialiste à des catégories et au faible éventail de notes proposées dans les formulaires d'évaluation. Il a été conclu que la fiabilité ne peut être garantie en présence d'un évaluateur et, par ailleurs, que la fiabilité devrait être mesurée lors des études pilote quel que soit le nombre d'évaluateurs.

RESUMEN

Este artículo analiza el acuerdo entre inspectores (hasta qué punto la gente está de acuerdo con sus propios juicios a lo largo del tiempo) en las inspecciones sobre el estado de conservación de las colecciones, y en particular en términos de relevancia para varias prácticas de conservación. El argumento se basa en el hecho de que las consecuencias de un acuerdo débil son evaluaciones inexactas que pueden dar lugar a problemas que no se reconocieron o a una mala asignación de los recursos. El experimento en cuestión está relacionado con dos inspecciones realizadas en un intervalo de diez semanas por un grupo de conservadores profesionales. El grupo realizó dos inspecciones de los mismos objetos de museo utilizando un formulario ya publicado. Se calculó la fiabilidad y se determinó que iba de niveles muy altos a niveles cercanos a un acuerdo por azar. Se examinaron varios factores, tales como la experiencia, las definiciones, los tipos de objetos y el impacto de reducir el juicio de los expertos a categorías y a un rango pequeño de grados disponibles en los formularios de evaluación. Se llegó a la conclusión de que la fiabilidad no se puede asumir cuando se usa un inspector, y además, que la fiabilidad debería medirse en estudios piloto independientemente del número de inspectores.

is being investigated has implications for any semi-quantitative method of judgment. This includes various preventive conservation activities, such as risk assessment and quantifying loss of value (Waller 2003).

METHOD

For the experiment, 15 conservators completed two surveys of 20 historic objects (sample size discussed with experimental psychologists) in a museum store with an interval of ten weeks. The choice of the ten week period was based on similar experiments carried out in textual analysis where surveyors had time to forget previous responses (Woodward and Franzen 1948), but was too little time for objects to change. The objects weren't chosen by the author, to avoid experimental bias, and represented a mixture of materials and assemblages.

The survey was carried out using the 'Museum of London Working Group' survey form and definitions (Keene 1991). Condition was graded from 1 (good) to 4 (unacceptable). The surveyors were told that the survey was a census survey of objects in the store where the survey took place, and condition grades had the following definitions from Keene (1991):

- 1 (good): good condition, stable
- 2 (fair): disfigured or damaged. No immediate action
- 3 (poor): probably unstable. Needs remedial work
- 4 (unacceptable): actively deteriorating.

The same scoring was applied for the categories used on the form ("Major Structural Damage", "Minor Structural", "Biological", "Chemical", "Disfigurement", "Surface Damage", "Accretions" and "Old Repairs"). Surveyors were given the form with the published definitions of condition and the categories, e.g. "Chemical Damage": acid paper, corrosion, rubber and plastic breakdown, "Surface Damage": flaking, crazing, lifting, abrading.

The museum store was well lit, but with no natural light. The objects were numbered randomly, but assessed in the same order by each surveyor to remove any priming effects. There was no time limit, but times were recorded. No discussion about the objects' condition was allowed.

Surveyors were all trained conservators and had some general experience in conservation (specialists had a general understanding of all materials). Memory effects, such as task familiarity, were examined as well as group reliabilities for the surveys to test the effect of task design on results.

Upon finishing each survey, the conservators were asked a series of questions connected with the task (exit interviews). For example, if they found any objects or categories difficult or if they were confident that the two surveys they carried out had similar responses.

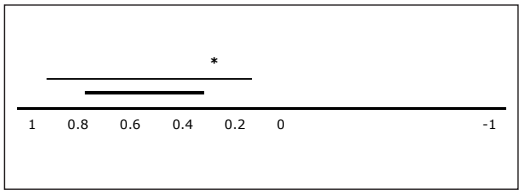

Figure 1
Range of intra-surveyor reliability levels using *alpha*

Recent developments in calculating reliability methods are described by Taylor and Watkinson (2007). Since there is no 'right answer', only subjective judgment, more well-known statistical techniques aren't appropriate. Krippendorff's *alpha* (Krippendorff 2004) is the gold standard for calculating reliability in qualitative judgments, as it accounts not only for when the same grade is given, but for correlation, chance agreement and the number of categories (Neuendorf 2002) and was used for the analysis along with other indices described by Taylor and Watkinson (2007).

RESULTS

Intra-surveyor reliability was calculated for each person from results of the first and second surveys using *alpha* (Table 1). The calculations in Table 1 are the group reliability levels (agreement for all 15 surveyors) for both surveys and the average of each individual's intra-reliability. Group levels were similar for both surveys, but the average intra-surveyor level was significantly higher. The scale of the indices is -1.00 to +1.00; 0.00 is chance agreement; +0.70 is suitable for drawing conclusions, as described in Taylor and Watkinson (2007).

Agreement between 15 people is more difficult to achieve than between two, but the calculations account for sample size. However, there was profound variation in some of the intra-surveyor scores, as illustrated by the wide range and standard deviation (Table 1). One surveyor recorded almost identical responses, whilst another surveyor provided near-chance agreement. It is evident that the judgment of one person is not necessarily less variable than that of different people. Therefore, it cannot be assumed that, over time, one surveyor will show more consistency in assessment than two different people would. Figure 1 illustrates the range that can be found for intra-surveyor agreement (thin line) against one standard deviation of the average (thick line) and the inter-surveyor reliability (star).

Table 1
Inter- and intra-surveyor reliability, with the range and standard deviation for intra-surveyor reliability

Index	Survey 1	Survey 2	Intra Avg.	Range	St. Dev
Alpha	0.328	0.389	0.579	0.141-0.954	0.212
Perreault's Pi	0.448	0.430	0.724	0.516-0.931	0.122
Pearson's r	0.444	0.429	0.626	0.298-0.935	0.190

Checks were made to ensure that recorded agreement was not simply due to surveyors using a smaller range of grades (thus promoting agreement artificially), or objects simply being in good condition (there's nothing to disagree on). Table 2 shows the surveyors (in order of reliability) and the range of grades they used. *Alpha* accounts for the number of categories in its calculation, so no correlation with grade spread existed. There was limited correlation between condition and reliability (objects in good condition showing more agreement because there was less to disagree on), but nothing else. This meant that the effect was real rather than generated by the experiment.

Table 2
Individual intra-surveyor agreement levels, and the spread of grades used in their assessments

Surveyor	Index		Spread of grades	
	Alpha (ordinal)	Pi (nominal)	Survey 1	Survey 2
10	0.954	0.931	1-4	1-4
8	0.913	0.931	1-3	1-3
11	0.739	0.856	1-4	1-3
15	0.725	0.516	1-4	1-4
14	0.631	0.683	1-4	1-4
4	0.624	0.775	2-4	2-4
12	0.603	0.816	2-4	2-4
1	0.597	0.683	1-4	1-3
2	0.594	0.683	1-3	1-3
3	0.578	0.730	1-4	1-4
5	0.524	0.577	1-4	1-3
6	0.428	0.683	1-3	1-3
9	0.383	0.730	2-3	2-4
13	0.244	0.683	1-3	1-3
7	0.141	0.577	2-4	2-3

Experience in condition surveys and conservation

In terms of surveyor proficiency, it is reasonable to expect that assessing condition would improve with experience. Previous studies have revealed that awareness of deterioration factors has more impact than experience by comparing professional conservators, students and a control group (Taylor 1999). However, experience in conservation is different from experience in condition surveys.

Experience in surveying and conservation did have a slight effect, but not a positive one. Surveyors who had some limited experience in surveying (having completed one to five surveys) showed more variation than those who had no prior experience with condition surveys, or than those who had completed more than five surveys. This disparity was possibly due to the inevitable preconceptions held when a surveyor is used to a particular system. Conservators with a lot of experience in surveying (more than five surveys) had no fixed ideas about assessment, nor did those without any experience in surveying. Those that had done a small number of surveys tended to have more fixed ideas, based on exit interviews. This may seem contrary to Appelbaum's (2007) assertion that experience has a significantly positive impact on judging condition, but it is connected to experience with surveying rather than conservation. Experience in conservation (ten or more years) does not have significant impact on survey data, as category sets do not allow for detailed, nuanced judgment to be communicated. The findings demonstrate the extent to which the survey category set gets in the way of representing collection condition. There are not only differences in how people see objects and their physical state, but also differences in how people see the assessment process and categories – a further filter between the reality and the representation of condition.

Internal consistency

One effect of having experience in both conservation and condition surveys was the speed of assessment. There was a reduction in the amount of time taken for the surveys (43 minutes average for Survey Two, 58 minutes for Survey One). Also, task familiarity may improve consistency in the *approach* to assessment, if not reliability. Reliability was not affected by the amount of time taken over the survey, so the process of assessment became more efficient but not more reliable. However, this gave some indication to the kind of internal consistencies that were present. Levels of intra-surveyor correlation (measured with Pearson's r) ranged from statistically significant agreement (0.935; 0.05 significance) to very low (0.249). High correlation indicates that people have a consistent idea of which objects are in 'better' and 'worse' condition than others, even if they have different ideas of what 'better' and 'worse' mean. However, the range of correlation recorded indicated that people approached assessment in quite different ways – a finding corroborated by different studies connected to this experiment (Taylor 2009).

This difference may seem irrational, but is connected to the fact that there are many deterioration mechanisms and kinds of damage (a value-laden concept), and the term condition covers them all. People are simply applying a range of criteria to assess condition due to the complexity of the task and ambiguity of terms. If, for example, a person said they preferred apples to oranges and oranges to bananas they would not be considered irrational to say that they preferred bananas to apples, only that they vary the criteria when judging. This is not recognised in the broad-brush term 'condition', so some people display variation in criteria they use to assess objects.

As discussed, when correlation is high and agreement low, it often means surveyors agree that one object is in better condition than another, but there is a variable idea of what constitutes 'good' or 'bad' condition. This means surveyors correlate well but do not frequently record the same grade both times. This is more common than poor correlation, and is affected by how one sees damage as much as the objects themselves. Condition assessment has two phases – recognizing deterioration, and judging the extent. Conservators are consistent in recognizing deterioration (so often correlate well), but have different ideas of what that means for an object – the value-laden idea of damage and condition (so don't always directly agree). This varies for a number of reasons, such as the role of the conservator (Taylor 2009).

Category sets

A lot of variation could be attributed to surveyors being unfamiliar with the categories (even though the chosen categories were published). Potentially, this is something that could improve with a second survey, but the differences in *alpha* reliability for individual categories were largely negligible in a range between -1.00 and +1.00 (Table 3). Only one category, "Minor Structural Damage", showed difference (the differences in "Biological

Damage" were connected to the index calculation). "Minor Structural Damage" was considered the most problematic category in exit interviews, due to the overlap with "Major Structural Damage".

The ambiguity of categories was a clear cause of disagreement, even with one surveyor, and it did not appear to be something that was reduced by familiarity. Some of the reliability levels recorded, such as "Minor Structural Damage" and "Accretions" were close to chance agreement (Table 3). Ambiguity of categories is clearly a significant factor – not just how different people interpret them, but how they can mean different things to the same person at different times. It also shows that diagnosing causes of deterioration using damage categories, already a challenge (Taylor 2005, 2009), can be very problematic.

Table 3
Inter-surveyor reliability for categories for both surveys with the difference in values

Categories	Survey 1	Survey 2	Difference
Major Structural Damage	0.567	0.570	+0.003
Minor Structural Damage	0.168	0.183	+0.015
Biological Damage	0.124	0.138	+0.014
Chemical Damage	0.548	0.553	+0.005
Surface Damage	0.401	0.405	+0.004
Disfigurement	0.226	0.231	+0.005
Accretions	0.150	0.153	+0.003
Old Repairs	0.660	0.664	+0.004

Average condition grades

Categories may be used in a number of survey applications, but the final judgment of a collection is mostly related to the overall condition grade. The experiment showed that an individual could record two average condition grades that could differ by up to 0.5 on a 1-4 scale (Table 4). The consequence is that a collection (or parts of a collection) could be considered in better or worse condition than they are, leaving undesirable change to go unchecked or costly resources spent on a minor matter.

Calculating the average condition grade alone isn't informative. If variation was random, then the differences would even themselves out, so the difference in average condition score for each survey would be small (e.g. surveyor 4, Figure 4). However, a difference in average condition can illustrate the ways in which 'good' and 'bad' condition manifest, and indicate differences in the frames of reference.

Even if much of the environment is the same (including the same objects), the complex concepts involved in assessing condition can be interpreted in different ways and have a systematic effect on the whole survey, as well as single objects. Most often, the average grade was considered better in the second survey, and since the objects couldn't have improved, the effect could not be connected to the objects.

Table 4

Raw data for five surveyors' first (F) and second (S) surveys, with the difference in average condition and *alpha* value

Surveyor and session		Object																				Grade (S - F)	Alpha level
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
1	F	1	2	1	2	3	2	3	1	4	2	1	1	4	3	1	2	1	4	2	1	-0.5	0.597
	S	1	2	1	1	3	1	3	1	2	3	1	1	2	1	1	2	1	2	1	1		
4	F	2	3	1	1	4	2	2	3	3	2	1	2	3	3	1	3	1	2	2	1	0.1	0.624
	S	2	2	2	2	3	2	3	1	3	3	1	2	3	2	1	2	2	2	1	1		
2	F	1	1	1	2	1	1	2	2	2	2	1	1	3	1	1	2	1	2	1	1	+0.3	0.594
	S	1	2	1	1	2	1	2	3	3	2	1	1	3	2	1	3	1	2	2	1		
13	F	1	2	2	3	3	2	2	1	2	3	1	1	2	3	1	3	1	2	3	1	-0.4	0.244
	S	1	2	1	1	1	1	2	1	2	3	1	1	3	3	1	1	1	2	1	2		
10	F	2	2	1	2	2	2	3	2	4	3	2	1	4	3	1	3	2	3	2	2	-0.1	0.954
	S	2	2	1	2	2	2	3	2	4	3	2	1	3	3	1	3	1	3	2	2		

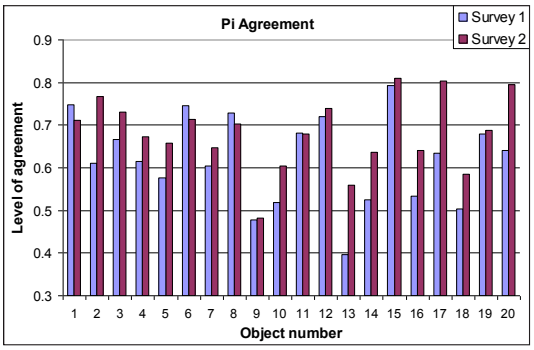

Figure 2

Pi agreement for all the objects for the first and second surveys

Objects

Some individual objects showed variation in reliability between surveys, which may be the result of contextual factors. Comparing inter-reliability levels for the different objects illustrated some further issues. Although the overall reliability was similar (and slightly worse in survey 2), there are objects that did show differences and familiarity is likely to be a part of that. This evidence of task familiarity is interesting, but did not affect the overall reliability in any meaningful way. What it does indicate is that there is potential for reliability to be increased through training. Figure 2 indicates the changes in reliability for individual objects with Perreault and Leigh's *Pi*. *Pi* measures the instances when the exact same grade is recorded, against chance occurrence (Perreault and Leigh 1989). The red line indicates where inter-surveyor agreement was recorded. Almost all the levels are higher for individual objects than for the collection because agreement is much more specific for individual objects.

During exit interviews, conservators were asked if they found any of the objects difficult to survey, and these objects were removed from the individuals' analysis to determine whether disagreement was exacerbated by 'problematic' objects. Reliability values actually decreased by 0.118 for ordinal *alpha* when 'problem objects' were removed, which indicates that confidence in judgment is not a good indicator of reliability. Conservators were also asked how close they thought their two surveys agreed and if particular materials were problematic, neither of which revealed a relationship between level of confidence and reliability. Feeling that one's results are accurate or similarly recorded is not enough to claim reliability. Reliability needs to be measured.

DISCUSSION

Assigning condition is a relativistic task, in that condition is subject to frames of reference that are influenced by contextual factors. People's approaches to assigning condition grades do vary, between each other and

sometimes within the same person. One could argue that this is because of our intelligence, rather than our lack of it. It is indicative of juggling a range of considerations. The internal inconsistency is related to what constitutes 'good' or 'poor' condition at different times and how ambiguous categories can be utilised for different purposes. Often the importance of assessment criteria will vary for each individual case. The lack of correlation in some surveyors' responses suggests that these criteria are not applied consistently. This indicates that the actual approach to condition assessment, as well as judgment of extent, can vary.

If the reliability of every surveyor is not assessed in advance, one cannot generalise about expected reliability. The variation between scores is too great. In a practical situation, it may be possible to measure reliability during a pilot study and find sources of variation, be they people, objects or categories. Improving decision-making and definitions should be the first step. Also, as noted above, reliability indices can help diagnose causes of variation in data (Taylor and Watkinson 2007).

CONCLUSION

In this paper, we have seen that reliability cannot be ensured by using only one conservator, even though a dominant factor in interpretational bias (different people) is removed. Empirical evidence of condition provides essential information for many conservation activities, so understanding this variation is vital in order to reduce its influence.

Individuals' variation can be attributed to a number of things, including the survey form, changes in frames of reference, the type of material, the object's complexity, differences in assessment criteria and personal factors. Different styles of assessment create different kinds of reproducibility to different extents. More intuitive styles lead to more correlation, and more rigid styles to more direct agreement.

Based on the intra-surveyor reliability data, it appears that in practice, pilot studies, training and decision aids are valuable features that should be used even when only one surveyor is carrying out a survey. It also suggests that people are responsive to such techniques. Training and decision aids have been examined (Taylor 2009) and although outside the scope of this paper, will be the focus of future publications.

In the long-term, it may be that collection condition surveys have to change. Reliance on ambiguous categories to document and interpret material change is fraught with difficulty, and cannot be used effectively without other kinds of data (Taylor 2005). Removing the need for observation and holistic judgment is not feasible, but understanding how and why we collect data will ensure that its quality improves.

ACKNOWLEDGMENTS

Much gratitude goes to all the (anonymous) conservators who took part in the survey and to Helen Kingsley and Science Museum staff for selection

and use of objects and space, to David Watkinson, Jonathan Ashley-Smith and Alastair McClelland for advice on the experiment and to Klaus Krippendorff, James Helgeson and Harry Bruhns for statistical advice.

REFERENCES

APPELBAUM, B. 2007. *Conservation treatment methodology*. Oxford: Butterworth-Heinemann.

ASHLEY-SMITH, J. 1999. *Risk assessment for object conservation*. Oxford: Butterworth-Heinemann.

HENDERSON, J. 1997. What use are collection surveys? *Natural Sciences Conservation Group Newsletter* 6: 14.

JOHNSEN, J.S. 1999. Introduction to the surveyor's guide to condition assessment of photographic collections. In *ICOM-CC 12th Triennial Meeting Preprints, Lyon 29 August –3 September 1999*, ed J. Bridgland, 555–560. London: James and James.

KEENE, S. 1991. Audits of care: a framework for collections condition surveys. In *Storage,* eds. M. Norman and V. Todd, 6–16. London: United Kingdom Institute for Conservation.

KEENE, S. 2002. *Managing conservation in museums*, 2nd ed. London: Butterworth-Heinemann.

KRIPPENDORFF, K. 2004. *Content analysis: an introduction to its methodology*, 2nd ed. Beverly Hills: Sage Publications.

NEUENDORF, K.A. 2002. *The content analysis guidebook.* Thousand Oaks: Sage Publications.

PERREAULT, W.D. JR., and L.E. LEIGH. 1989. Reliability of nominal data based on qualitative judgments. *Journal of Marketing Research* 26: 135–148.

TAYLOR, J. 1996. An assessment of condition surveys as objective tools for analysis. Undergraduate dissertation, Cardiff University, UK.

TAYLOR, J. 2005. An integrated approach to risk assessment and condition surveys. *Journal of the American Institute for Conservation* 44:127–141.

TAYLOR, J. 2009. An examination of the validity and reliability of collection condition surveys. Ph.D dissertation, Cardiff University, UK.

TAYLOR, J., and S. STEVENSON. 1999. Investigating subjectivity within collection condition surveys. *Journal for Museums Management and Curatorship* 18: 18–42.

TAYLOR, J., and D. WATKINSON. 2007. Indexing reliability for condition survey data. *The Conservator* 30: 49–62.

WALLER, R.R. 2003. Cultural property risk analysis model: development and application to preventive conservation at the Canadian Museum of Nature. Ph.D. dissertation, University of Göteborg, Sweden.

WOODWARD, J.L., and R. FRANZEN. 1948. A study of coding reliability. *Public Opinion Quarterly* 12: 253–257.