The reliability of trust-level survey scores: a comparison of three different scoring models

Steve Sizmur
Picker Institute Europe

Abstract

Comparing trusts with one another can be useful for performance monitoring and quality improvement. Three different scoring approaches were examined to determine how reliably they differentiated between trusts' aggregated patient experience results. National survey data for a range of questions were analysed using Generalizabilty theory applied to the Picker 'problem score', the partial credit scoring system used for benchmarking by the Care Quality Commission, and a 'bottom box' score like that used in Care Quality Commission Quality and Risk Profiles. Variance estimates obtained from multilevel regression models (both with and without case-mix adjustment) were used to calculate trust-level generalizability coefficients. The problem score and partial credit approached produced similarly high levels of reliability, supporting use of both these methods in comparing trusts' performance and guiding service improvement, while the bottom box approach fared rather less well. The meaning attached to scores needs to be considered in conjunction with reliability when choosing a scoring approach.

Assessment of patient experience at the trust level

Although surveys are completed by individual patients, these individual responses are never reported, for both ethical and practical reasons. Instead, results are aggregated to summarise patient experience at the level of healthcare provider units, usually (but not necessarily) NHS trusts. However, not all respondents make the same evaluation of the care they have received. An important question, then, is how well the reported summary represents the views of many more patients. An important facet of this question is how reliable the result is, and a key aspect of reliability in this context is how well the results differentiate between providers. Comparing trusts with one another is useful in benchmarking trust performance and identifying areas of experience for consolidation or improvement.

Many questions have more than one response option. Interpreting results for several different answer options at the same can be challenging, particularly if comparing to other trusts or to previous results. For this reason, some form of response weighting or scoring is often used to give a single summary result for each question in each trust. We consider here three kinds of scoring:

(i) The Picker 'problem score', used by Picker Institute Europe in its reports to client trusts. For this, any response other than the most positive is scored as a 'problem', and the most positive option as 'no problem'. This reflects the Picker Institute emphasis on

the highest standards in patient experience and its use of data for quality improvement work. Problem scores are reported because any instance where a patient has selected something other than the most positive option is one where there was room for improvement in that patient's experience. It also reflects a common approach in health measurement more generally where higher scores indicate greater health issues. It is the equivalent (but inverse) of a 'top box' score that rates only the most positive response for a question.

(ii) The scoring model applied by the Care Quality Commission and its predecessors for reporting trust benchmark results on questions. This comprises scoring the most positive answer option as 10 (previously 100) and the least positive as 0. Intermediate answer options are scored with intermediate values (for example, the middle of three options would be scored 5). This might be termed a 'partial credit' score. The conceptual value of this approach is that it allows all response options to be taken into account when scoring questions, rather than just the most extreme ones. This can differentiate between organisations with similar proportions of patients selecting an extreme option, but a limitation is that two trusts with the same score might have quite different distributions of responses.

(iii) What might be called 'bottom box' scoring, in which attention is focused on the least positive answer option – a more extreme form of problem score, appropriate to identifying the greatest shortfalls in patient experience. This is the approach used by the Care Quality Commission in its Quality and Risk Profile assessments of compliance to care standards (CQC, 2012), where the rationale is to identify the most serious or the most alarming problems only and where observed risk is highest.

The different scoring options are illustrated here.

Problem score	Partial credit	Bottom box
Did you have confidence and trust in the nurses treating you?	Did you have confidence and trust in the nurses treating you?	Did you have confidence and trust in the nurses treating you?
0 ☐ Yes, always	10 ☐ Yes, always	0 ☐ Yes, always
1 ☐ Yes, sometimes	5 ☐ Yes, sometimes	0 ☐ Yes, sometimes
1 ☐ No	0 ☐ No	1 ☐ No

These different approaches may well result in different assessments of the trust's performance but may also differ in the extent to which they can differentiate between different levels of trust performance.

Confidence intervals and reliability

The use of a sample, rather than all possible patients, introduces uncertainty to results. Most often, this uncertainty is expressed in the form of confidence intervals. Typically, 95% confidence intervals are shown. The interpretation of these is that in 95 out of 100 equivalent samples, the interval will include the 'true' population result (and in five per cent they won't). This is often somewhat loosely translated as 'the range within which the true result is likely to be, with 95% probability'. Confidence intervals are dependent on the scale and variability of the measurement, making it difficult to set a general criterion for an acceptable level of precision. They may well be useful for comparing a trust's

result to a national average (and in modified form for comparing between two trusts), but are of limited value for evaluating how well in general the scores can differentiate trusts from one another. This is the subject of reliability theory.

As commonly understood, reliability refers to the consistency or reproducibility of results. However this is not completely accurate. A simple analogy shows why: a clock that is stopped will be perfectly consistent (it will always show the same time), but it will be useless for telling the actual time because it is unable to differentiate one moment in time from another. Similarly, if all patients gave the same response to a question in every trust, the trust scores would be perfectly consistent, but the calculated reliability would actually be zero. Streiner and Norman (2003) helpfully focus their definition of reliability on how well a measurement differentiates between the objects it is intended to measure. Reliability is therefore the extent to which a measurement consistently differentiates between units that differ on whatever is being measured. Technically, reliability is the correlation between the 'observed' scores (the reported results) and the 'true' average results that would theoretically be obtained in the long run, if it was possible to keep repeating the measurement. Equivalently, it represents the proportion of all variation that is the underlying 'true score' variation.

As it is impossible to obtain true scores, reliability has to be estimated using the data available. Cronbach et al (1972) elaborated a comprehensive framework for doing this: Generalizability Theory. This uses the statistical analysis of variance approach to partition variation in scores according to the different sources of variation. The primary purpose is to distinguish between true variance (that due entirely to differences in the phenomenon of interest) from various sources of 'error' variance or unreliability. The framework and its terminology are complex, but mercifully most of that complexity is irrelevant in the present context.

Generalizability coefficients are based on the intraclass correlation (ICC). This shows the proportion of total variance that is 'between-groups' variance attributable to the objects of measurement (here, trusts), or equivalently the extent to which lower-level units (here, responses) are consistent within those objects of measurement. ICCs have been used directly in assessing the discriminative power of survey data (Boer et al, 2011). However, they are difficult to interpret and no account is taken of the effect on reliability of differing numbers of respondents. Generalizability theory offers a mechanism for doing just this, and also falls within the wider reliability literature in which there are established (if contested) standards for what counts as sufficiently reliable. In a so-called 'D-study', the effect of different sample sizes on the reliability coefficient can be investigated by substituting alternative numbers into a formula.

This was the approach used to investigate the comparative reliability of the three different scoring models.

Data analysis

The data used are from the NHS national Inpatients Survey 2010 for England. An arbitrary selection of questions was made from the survey, representing a range of different areas of experience at different points in the care pathway. The question responses for each respondent were converted into scores, using the three different weighting models, to derive three new sets of variables that were then the subject of analysis.

The data were modelled in MLwiN regression software (Rasbash et al, 2012) using a two-level linear variance components model. This incorporated trusts as random effects at level 2 and respondents within trusts at level 1. The intraclass correlation was calculated from the variance components at the two levels, and in addition a generalizability coefficient was computed, based on an illustrative sample size of 250 patients per provider.

Because responses can vary according to patient demographics, and because this is taken into account in national survey results by applying weighting, a further set of models was run with stratification group as a fixed effect at level 1, providing a case-mix adjustment. Stratification group was a 16-level categorical variable that reflected the standardisation strata used by the Care Quality Commission to weight the inpatient data. For this, patients were cross-classified according to age group (16-35, 36-50, 51-65, 66+), gender and admission route (emergency or elective). The variance partition coefficient (equivalent to an ICC) and the generalizability coefficient were calculated from the adjusted model for a sample of 250 patients, in the same way as for the unadjusted model.

Results

The full results for all questions examined are provided in Appendices A and B. The tables in these appendices show the question variance at trust level (V(T)) and at respondent level within trusts (V(R:T)) under the three different scoring models. The ICC (VPC in the adjusted model) is shown for each scoring method, and alongside it the generalizability coefficient (G) applicable to a sample of 250 patients per trust.

As intended, the case-mix adjustment reduced variance at both the trust and patient levels, sometimes substantially. The effect was generally to decrease the VPC, indicating that trust level variance was reduced more than patient-level variance.

With one exception (Q67: *Did a member of staff tell you about any danger signals you should watch for after you went home?*), the bottom-box scoring approach consistently produced the least reliable results at trust level. For Q67 without case-mix adjustment, this scoring was marginally more reliable than the problem score approach, but the advantage disappeared in the adjusted model.

The performance of the other two scoring models was more similar across the range of questions, with sometimes the partial credit scoring having a small advantage, sometimes (but slightly less often) the problem score, but often little to distinguish between them.

Discussion

A commonly-accepted criterion for reliable differentiation is a coefficient of 0.80. The target sample size for the D-study was set to 250 respondents (fairly low for most patient surveys, except for where questions are intended to be answered by only a subset of patients). With this number of responses, the partial credit and problem score models achieved close to the minimum reliability for a number of questions, while the bottom box scoring generally fell rather short of the criterion. For one question (Q3: *While you were in the A&E Department, how much information about your condition or treatment was given to you?*), none of the scoring models achieved high reliability, though the problem score came close to a more relaxed criterion of 0.70 that is sometimes accepted for assessments that are not 'high stakes'. This underlines the potential for using

generalizability analysis to evaluate question performance and to add to the information used to select questions for inclusion in surveys.

In common with other findings in this area (Boer et al, 2011), case mix adjustment reduced the amount of variation in patient experience that could be attributed to providers. In national surveys, a similar aim is addressed by applying standardisation weights to the patient-level data.

Returning to the main aim of this work, it can be concluded that both the Picker problem score and the partial credit scoring produced similar levels of trust-level reliability and would therefore be similarly capable of discriminating between providers. The bottom box scoring did not generally distinguish between trusts and performed rather less well. This is indicative of wide patient-level variation within trusts in selecting the bottom response option, which in itself might be worthy of further investigation.

That two of the approaches produced similar levels of reliability does not necessarily indicate that they are doing exactly the same job: the 'winners' and 'losers' under these models might be different. Reliability should therefore be considered in conjunction with score meaning. Clearly, the three different scoring approaches say different things about trusts. The problem score emphasises excellence, the bottom box detects failure and the partial credit approach is more even-handed (or accepting of mediocrity, depending on the point of view).

The Picker problem score (or equivalently the 'top box' approach) has fared well in this evaluation of patient experience scoring models, supporting its use in assessing trusts and providing information for service improvement, and is in keeping with the organisation's promotion of excellence.

References

Boer DD, Delnoij D, Rademakers J. (2011). The discriminative power of patient experience surveys. *BMC. Health Services Research*. 11: 332.

Care Quality Commission (2012). *Quality and Risk Profiles. Data Sources: Acute and Specialist NHS trusts*. London: Care Quality Commission.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley.

Rasbash, J., Charlton, C., Browne, W.J., Healy, M. and Cameron, B. (2012) MLwiN Version 2.25. Centre for Multilevel Modelling, University of Bristol.

Streiner, D., & Norman, G. (2003). *Health measurement scales* (4th ed.). Oxford: Oxford University Press.

Appendix A: results without case mix adjustment

While you were in the A&E Department, how much information about your condition or treatment was given to you?						
Q3 partial credit						
Trusts:	V(T) =	7.3	G =	0.63	ICC =	0.007
Respondents(Trusts):	V(R:T) =	1063.7				
Target N:	N =	250				
Q3 bottom box						
Trusts:	V(T) =	2.2	G =	0.39	ICC =	0.003
Respondents(Trusts):	V(R:T) =	878.4				
Target N:	N =	250				
Q3 problem						
Trusts:	V(T) =	17.5	G =	0.69	ICC =	0.009
Respondents(Trusts):	V(R:T) =	1966.1				
Target N:	N =	250				
When you had important questions to ask a doctor, did you get answers that you could understand?						
Q31 partial credit						
Trusts:	V(T) =	13.2	G =	0.79	ICC =	0.015
Respondents(Trusts):	V(R:T) =	857.8				
Target N:	N =	250				
Q31 bottom box						
Trusts:	V(T) =	2.6	G =	0.56	ICC =	0.005
Respondents(Trusts):	V(R:T) =	529.5				
Target N:	N =	250				
Q31 problem						
Trusts:	V(T) =	32.4	G =	0.79	ICC =	0.015
Respondents(Trusts):	V(R:T) =	2158.8				
Target N:	N =	250				
Did you have confidence and trust in the nurses treating you?						
Q36 partial credit						
Trusts:	V(T) =	10.1	G =	0.79	ICC =	0.015
Respondents(Trusts):	V(R:T) =	659.4				
Target N:	N =	250				
Q36 bottom box						
Trusts:	V(T) =	1.2	G =	0.50	ICC =	0.004
Respondents(Trusts):	V(R:T) =	312.9				
Target N:	N =	250				
Q36 problem						
Trusts:	V(T) =	27.2	G =	0.79	ICC =	0.014
Respondents(Trusts):	V(R:T) =	1848.8				
Target N:	N =	250				
Did you find someone on the hospital staff to talk to about your worries and fears?						
Q44 partial credit						
Trusts:	V(T) =	33.8	G =	0.86	ICC =	0.023
Respondents(Trusts):	V(R:T) =	1422.3				
Target N:	N =	250				
Q44 bottom box						
Trusts:	V(T) =	23.8	G =	0.78	ICC =	0.014
Respondents(Trusts):	V(R:T) =	1642.7				
Target N:	N =	250				
Q44 problem						
Trusts:	V(T) =	46.0	G =	0.83	ICC =	0.019
Respondents(Trusts):	V(R:T) =	2374.3				
Target N:	N =	250				

Did a member of staff tell you about any danger signals you should watch for after you went home?						
Q67 partial credit						
Trusts:	V(T) =	47.7	G =	0.86	ICC =	0.024
Respondents(Trusts):	V(R:T) =	1929.7				
Target N:	N =	250				
Q67 bottom box						
Trusts:	V(T) =	48.1	G =	0.84	ICC =	0.020
Respondents(Trusts):	V(R:T) =	2302.0				
Target N:	N =	250				
Q67 problem						
Trusts:	V(T) =	47.7	G =	0.83	ICC =	0.020
Respondents(Trusts):	V(R:T) =	2376.8				
Target N:	N =	250				
Overall, did you feel you were treated with respect and dignity while you were in the hospital?						
Q72 partial credit						
Trusts:	V(T) =	8.5	G =	0.79	ICC =	0.014
Respondents(Trusts):	V(R:T) =	581.1				
Target N:	N =	250				
Q72 bottom box						
Trusts:	V(T) =	0.8	G =	0.41	ICC =	0.003
Respondents(Trusts):	V(R:T) =	281.9				
Target N:	N =	250				
Q72 problem						
Trusts:	V(T) =	24.3	G =	0.79	ICC =	0.015
Respondents(Trusts):	V(R:T) =	1586.4				
Target N:	N =	250				

Appendix B: results with case-mix adjustment

While you were in the A&E Department, how much information about your condition or treatment was given to you?						
Q3 partial credit						
Trusts:	V(T) =	7.2	G =	0.63	VPC =	0.007
Respondents(Trusts):	V(R:T) =	1057.9				
Target N:	N =	250				
Q3 bottom box						
Trusts:	V(T) =	2.0	G =	0.36	VPC =	0.002
Respondents(Trusts):	V(R:T) =	875.2				
Target N:	N =	250				
Q3 problem						
Trusts:	V(T) =	17.1	G =	0.69	VPC =	0.009
Respondents(Trusts):	V(R:T) =	1952.3				
Target N:	N =	250				
When you had important questions to ask a doctor, did you get answers that you could understand						
Q31 partial credit						
Trusts:	V(T) =	7.5	G =	0.69	VPC =	0.009
Respondents(Trusts):	V(R:T) =	828.1				
Target N:	N =	250				
Q31 bottom box						
Trusts:	V(T) =	1.6	G =	0.44	VPC =	0.003
Respondents(Trusts):	V(R:T) =	522.7				
Target N:	N =	250				
Q31 problem						
Trusts:	V(T) =	18.0	G =	0.68	VPC =	0.009
Respondents(Trusts):	V(R:T) =	2088.5				
Target N:	N =	250				
Did you have confidence and trust in the nurses treating you?						
Q36 partial credit						
Trusts:	V(T) =	7.9	G =	0.75	VPC =	0.012
Respondents(Trusts):	V(R:T) =	640.6				
Target N:	N =	250				
Q36 bottom box						
Trusts:	V(T) =	1.0	G =	0.44	VPC =	0.003
Respondents(Trusts):	V(R:T) =	310.2				
Target N:	N =	250				
Q36 problem						
Trusts:	V(T) =	21.2	G =	0.75	VPC =	0.012
Respondents(Trusts):	V(R:T) =	1798.7				
Target N:	N =	250				
Did you find someone on the hospital staff to talk to about your worries and fears?						
Q44 partial credit						
Trusts:	V(T) =	24.3	G =	0.81	VPC =	0.017
Respondents(Trusts):	V(R:T) =	1400.4				
Target N:	N =	250				
Q44 bottom box						
Trusts:	V(T) =	17.3	G =	0.73	VPC =	0.011
Respondents(Trusts):	V(R:T) =	1626.5				
Target N:	N =	250				
Q44 problem						
Trusts:	V(T) =	32.5	G =	0.78	VPC =	0.014
Respondents(Trusts):	V(R:T) =	2344.9				
Target N:	N =	250				

Did a member of staff tell you about any danger signals you should watch for after you went home?						
Q67 partial credit						
Trusts:	V(T) =	26.8	G =	0.78	VPC =	0.014
Respondents(Trusts):	V(R:T) =	1840.5				
Target N:	N =	250				
Q67 bottom box						
Trusts:	V(T) =	26.6	G =	0.75	VPC =	0.012
Respondents(Trusts):	V(R:T) =	2208.7				
Target N:	N =	250				
Q67 problem						
Trusts:	V(T) =	27.2	G =	0.75	VPC =	0.012
Respondents(Trusts):	V(R:T) =	2290.3				
Target N:	N =	250				
Overall, did you feel you were treated with respect and dignity while you were in the hospital?						
Q72 partial credit						
Trusts:	V(T) =	5.6	G =	0.72	VPC =	0.010
Respondents(Trusts):	V(R:T) =	555.9				
Target N:	N =	250				
Q72 bottom box						
Trusts:	V(T) =	0.5	G =	0.31	VPC =	0.002
Respondents(Trusts):	V(R:T) =	278.4				
Target N:	N =	250				
Q72 problem						
Trusts:	V(T) =	16.1	G =	0.73	VPC =	0.010
Respondents(Trusts):	V(R:T) =	1519.4				
Target N:	N =	250				