STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2277–2292
Published online 8 June 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3616

TUTORIAL IN BIOSTATISTICS

Hip psychometrics‡

Peter Baldwin[1,*,†], Joseph Bernstein[2] and Howard Wainer[1]

[1]*National Board of Medical Examiners, Philadelphia, PA, U.S.A.*
[2]*School of Medicine, University of Pennsylvania, Philadelphia, PA, U.S.A.*

SUMMARY

When data are abundant relative to the number of questions asked of them, answers can be formulated using little more than those data. But when data grow more sparse, so too does our tendency to lean on strong models to help us draw inferences. In this research we show how a strong item response model embedded within a fully Bayesian framework allows us to answer two important questions about the reliability and consistency of the clinical diagnosis of hip fractures from very limited data. We also show how the model automatically adjusts diagnoses for biases among the surgeons judging the radiographs. This research illustrates how a Bayesian approach expands the range of problems on which item response models can profitably be used. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: small sample; item response theory; Bayesian estimation; Garden's classifications; hip fractures

1. INTRODUCTION

When data are abundant, weak models are usually sufficient for robust analysis. For example, in a study of a single mis-scored test item, Wainer [1] used the responses from 830 000 examinees to show convincingly that although one item was scored incorrectly, it did not adversely affect the quality of the test's measurement. No formal test-scoring model was required. On the other hand, when data are limited, strong models are required to draw inferences. In the case we describe here, we show that with an extremely small data set, models may be profitably used. Our findings suggest that subtle structures that would otherwise have been invisible without a sample many times larger may be revealed through modeling.

*Correspondence to: Peter Baldwin, National Board of Medical Examiners, Philadelphia, PA, U.S.A.
†E-mail: pbaldwin@nbme.org

‡This work is collaborative in all respects and the order of authors is alphabetical. We would like to express our gratitude to Professor Xiaohui Wang of the University of Virginia for helping to get us to convergence and Eric T. Bradlow of the University of Pennsylvania for extensive comments on the manuscript. We are also delighted to thank the National Board of Medical Examiners for supporting the work of Peter Baldwin and Howard Wainer.

Received 20 October 2008
Accepted 1 April 2009

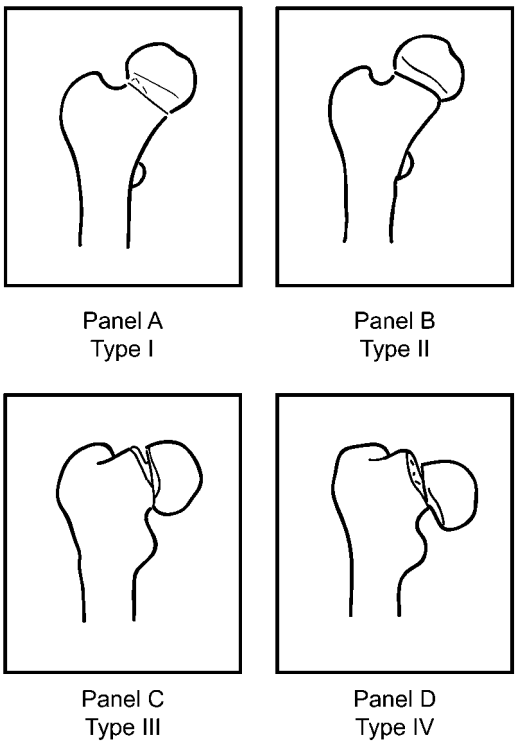

Figure 1. Garden's [3] four hip fracture classifications. Adapted from Figure 59 in Müller *et al.* [4].

In this study, we will show how to answer two difficult clinical questions about the diagnosis of hip fractures from radiographs using a Bayesian variant of Samejima's [2] polytomous item response model when we have but 12 surgeons and 20 radiographs.

The limitations on the size of the data sample add a methodological question to this mix, which relates to the suitability of item response theory to aid data summarization and facilitates inferences for these kinds of data. Such models, long a key element in the scoring of mental tests with moderate to large sample sizes, have not been used for such small data sets before; however, with the development of numerical estimation implementations for Bayesian methods such applications are now feasible.

At the heart of this study are two clinical questions of interest in the diagnosis of hip fractures. In particular

(1) Does Garden's [3] approach of classifying femoral neck fractures into four categories (see Figure 1), which is considered the *de facto* standard, contain distinctions between categories that are too fine to be useful given that there are only two clinical treatment choices?
(2) How consistent are orthopedic surgeons in their diagnoses? Should we expect consistent judgments from individual surgeons? Are Garden's classifications applied consistently by different surgeons?

DOI: 10.1002/sim

2. THE PROBLEM

Hip fractures are common injuries; more than 250 000 annually are treated in the U.S. alone. These fractures can be located in the shaft of the bone or in the neck of the bone connecting the shaft to the head of the femur. Femoral neck fractures vary in their severity. Garden [3] described a four part classification represented in the four panels of Figure 1: a partial crack through the neck (Panel A), a crack that fully extends across the femur but is impacted (Panel B), a complete crack and a slight displacement of the bone (Panel C), and a complete break with extensive displacement (Panel D). These four degrees of severity are classified by categories I, II, III, and IV, respectively.

When the break is displaced, the blood vessels that traverse the neck to nourish the head of the femur are more likely to be disrupted, markedly increasing the risk that the head will become necrotic. Due to this risk, the clinical treatment for fractures of types III and IV is a hip replacement; the broken end of the bone is removed and replaced with an artificial joint. By contrast, if the fracture is not displaced, preservation of the blood supply can be assumed and thus the fracture can be treated with pins to align and stabilize the bone and promote fracture healing. The intuitive assumption that hip replacement is a more major and drastic surgery is correct.

Because the clinical consequences of the diagnosis are profoundly different depending on how the fracture is classified, it is natural to ask how accurately orthopedic surgeons make such diagnoses. Unfortunately, it is not practical to answer the question of accuracy because a patient's true condition is typically not observed directly: the surgical dissection and manipulation performed during hip replacement will displace even those fractures that were not displaced pre-operatively, whereas the pinning procedure is performed through 1 cm slits in the skin using x-ray guidance and thus provides no opportunity to observe the fracture directly. Therefore, *all* fractures that were pre-operatively diagnosed as displaced, correctly or otherwise, will be discovered to be displaced when examined under direct visualization in the operating room and all fractures that were diagnosed as non-displaced will never be subject to direct visualization and independent confirmation.

Diagnostic accuracy, though unknowable in this context, is bound by diagnostic consistency: one cannot make correct judgments without first making consistent judgments. Thus, even without knowledge of patients' true diagnoses, it is still important to know how consistently orthopedic surgeons make these judgments. Consistency can be measured across surgeons (do different surgeons apply Garden's classifications equivalently?) and within surgeon (does the same surgeon reach the same conclusion when seeing the same radiograph a second time?).

For the purposes of our analysis, we treat Garden's classifications as a discrete representation of a continuous characteristic of the injury (its severity). The transformation of a continuous variable into a discrete one, while offering many practical advantages, results in a loss of information (e.g., any differences among radiographs with the same classification are ignored). We will show that some of this lost information may be recovered through modeling that allows us to locate surgeons and radiographs on this underlying severity continuum. The model we use is a stochastic one that yields a probabilistic outcome as well as estimates of the reliability of the categorization process.

3. THE DATA

Fifteen radiographs of broken hips were collected from the clinical practices of a university department of orthopedic surgery. The first five were repeated and added to the end of the list. The resulting 20 radiographs were presented to 12 orthopedic surgeons, who were asked to classify

DOI: 10.1002/sim

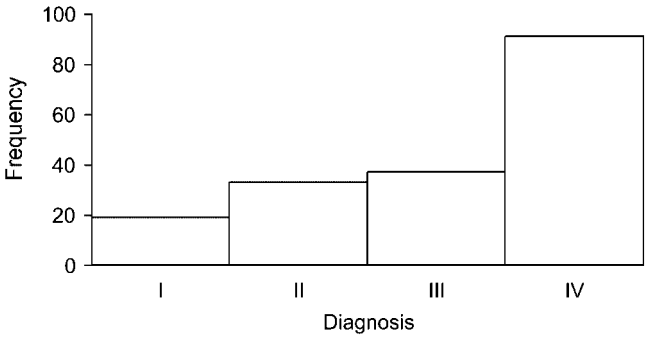

Figure 2. Frequency distribution of each of Garden's four classifications in the data set.

Table I. Raw data of hip fracture diagnoses.

	Case																			
Doctor	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	1*	2*	3*	4*	5*
A	1	4	4	3	4	3	4	4	4	4	1	4	4	4	2	2	4	4	3	4
B	1	4	4	3	4	2	4	4	3	4	2	4	4	4	2	1	4	4	3	4
C	2	4	4	2	3	2	4	4	3	4	1	4	3	4	2	1	3	4	3	4
D	2	4	3	2	4	1	3	3	2	4	1	3	3	4	2	2	3	4	2	3
E	3	4	4	2	4	3	4	4	3	4	2	4	4	4	2	2	4	4	3	4
F	1	4	3	2	3	2	4	4	3	4	1	4	3	3	2	2	4	3	3	3
G	1	3	4	2	2	1	4	4	4	4	1	4	3	3	2	2	3	4	2	4
H	1	4	4	3	3	3	4	4	3	4	2	4	4	4	1	1	4	4	3	4
I	4	3	3	2	4	2	4	4	4	4	1	4	3	4	2	2	4	4	2	3
J	1	4	4	2	3	2	3	4	4	4	1	4	3	4	1	1	3	4	4	4
K	3	4	2	3	4	2	4	4	2	4	2	4	4	4	1	3	4	4	2	4
L	4	4	4	2	4	2	4	4	3	4	1	4	4	4	3	2	4	4	2	4

The * indicates the 2nd administration of a previously viewed radiograph.

the fractures according to the I to IV categories Garden specified. Figure 2 shows a histogram of the 12 surgeons' responses to the first 15 radiographs. Each histobar represents the total number of diagnoses in each category; while the entire range was spanned, IV was the most frequently chosen diagnosis.

The full data set is shown in Table I.

4. THE MODEL FOR POLYTOMOUS ITEMS

The model we utilize for the I–IV ordinal Garden response scale data is an extension of the Samejima [2] ordinal response model in which the entire analysis is embedded in a fully Bayesian

DOI: 10.1002/sim

framework as described by Wainer *et al.* [5]. Specifically, we utilize a probit response model for response category $r=1,\ldots,R$, given by

$$P(Y_{ij}=r|d_j,t_{ij})=\Phi(d_{jr}-t_{ij})-\Phi(d_{jr-1}-t_{ij}) \tag{1}$$

where d_{jr} represents the latent cutoff for stimulus j and observed score r, t_{ij} is the latent linear predictor of response score r and is represented as $t_{ij}=a_j(\theta_i-b_j)$, and $\Phi(x)$ is the normal cumulative distribution function evaluated at x. In essence, the Samejima model posits that the probability of falling in the rth category is the probability that a normally distributed latent variable with mean t_{ij} and variance 1 falls between cutoffs d_r and d_{r-1}. We use the logical and common convention that $d_0=-\infty$, $d_R=\infty$ and $d_1=0$ to help to identify the model.

For the parameters describing the model in (1), we utilize the following priors to nest this within a Bayesian framework:

$$\theta\sim \mathrm{N}(0,1)$$

$$\begin{pmatrix}\log(a_j)\\ b_j\end{pmatrix}=\begin{pmatrix}h_j\\ b_j\end{pmatrix}\sim \mathrm{MVN}\left(\begin{pmatrix}\beta_h\\ \beta_b\end{pmatrix},\begin{pmatrix}\sigma_h^2 & \mathrm{cov}(\beta_h,\beta_b)\\ \mathrm{cov}(\beta_h,\beta_b) & \sigma_b^2\end{pmatrix}\right)=\mathrm{MVN}(\mu,\Sigma)$$

$$d_r\sim \mathrm{Unif}(d_{r-1},d_{r+1})$$

Noninformative hyperpriors were placed on the coefficients,

$$\beta_h\sim \mathrm{MVN}(0,V_a)\quad \text{and}$$

$$\beta_b\sim \mathrm{MVN}(0,V_b)$$

where $|V_a|^{-1}=|V_b|^{-1}$ is set to 0. Further, as is typically done, slightly informative hyperpriors were utilized for Σ to ensure both proper posteriors and to allow the data to drive posterior inferences. They were as follows:

$$\Sigma\sim \text{Inv-Wishart}\left(2,\begin{pmatrix}100 & 0\\ 0 & 100\end{pmatrix}\right)$$

For additional details, see Wang *et al.* [6] and Wainer *et al.* [5].

To aid in the intuition behind the polytomous model, consider the following interpretation. When surgeon C rates radiograph 4, the latent variable that is generated is t_{4C}, the latent linear predictor of score, which is assumed to be normally distributed with mean $a_C(\theta_4-b_C)$ and variance equal to one. This situation is portrayed graphically in Figure 3.

In this figure, we depict the conditional response distribution for a radiograph with severity of -1 encountering surgeon C. The dotted lines indicate the estimated values of the cutoffs d_j. The model predicts that when presented with a radiograph of severity -1, surgeon C will make a diagnosis of I with a probability of 0.05, II with a probability of 0.65, III with a probability of 0.30, and IV with a probability of 0.00. In the actual case of radiograph 4, which had an estimated severity of -1, surgeon C conformed to model expectations and categorized the fracture as a category II.

An item response model like this one can be fit to these data in two different ways. First, the data can be modeled as we have done: treating the radiographs as replications and the surgeons as stimuli. Alternatively, we could have followed traditional methodology and treated surgeons as

DOI: 10.1002/sim

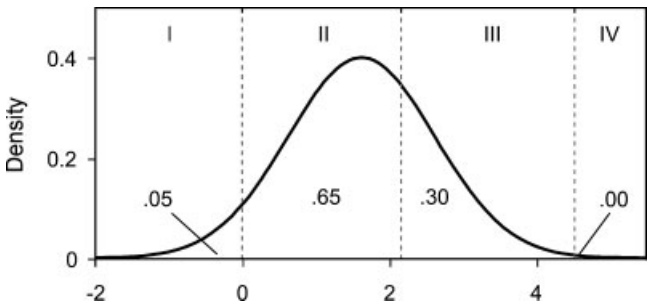

Figure 3. A graphical description of a polytomous test model. The dashed lines represent the estimated latent cutoffs that determine the responses for surgeon C and the curve represents the conditional response distribution for a radiograph with severity $=-1$.

replications and radiographs as stimuli. Our approach is the opposite of traditional practice and thus calls for further explanation.

The reasoning behind our decision falls along two lines: a data rationale and a model rationale. The data rationale reflects the constraints of working with such a small data set. Were we to treat surgeons as replications and radiographs as stimuli, our model would include 80 tracelines (20 radiographs with 4 tracelines per radiograph). However, 32 (40 per cent) of the resultant tracelines would have no observations associated with them at all (i.e., 32 radiograph/diagnosis combinations were not chosen by any surgeons). On the other hand, by treating surgeons as we did, we had only 48 trace lines (12 surgeons with 4 tracelines per surgeon). In this case, there was but a single traceline that lacked observations, a clear improvement.

The model rationale follows from the data rationale. With the exception of radiograph 1, no radiograph was classified into all four of Garden's categories. This helps to explain why 40 per cent of the radiograph/diagnosis combinations were not chosen. Surgeons may not frequently agree, but for these data, their disagreements were not profound. Because we do not expect vast differences in diagnoses for a given radiograph, it makes sense to choose a model that expects diagnoses to vary across radiographs for a given surgeon rather than a model that expects diagnoses to vary across surgeons for a given radiograph.

5. INFERENCES UNDER THE MODEL

To derive inferences under this model, we obtained samples from the marginal posterior distributions of the parameters, given the responses, using Markov chain Monte Carlo (MCMC) techniques [7–10] implemented in the software program SCORIGHT version 3.1 [6]. Details of the MCMC approach are found in Chapter 9 of Wainer *et al.* [5] (including information about access to the software and user's manual).

As standard output from the Markov chain, SCORIGHT produces posterior means and standard deviations for all parameters of the model as well as the posterior draws themselves that can be utilized for further analysis.

For the data analyzed here, SCORIGHT was run using four independent chains, each for an initial burn-in period of 165 000 draws and 20 000 iterations thereafter. Convergence of the MCMC

DOI: 10.1002/sim

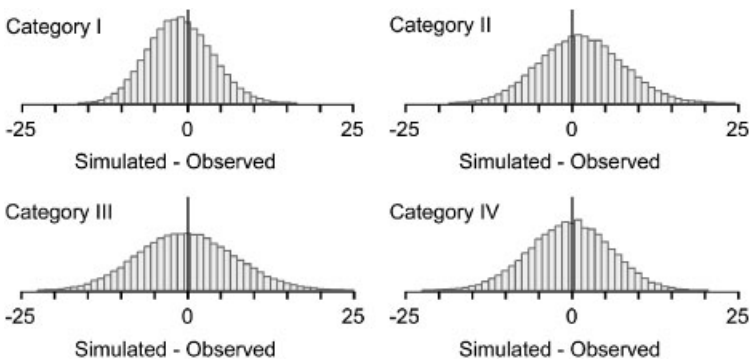

Figure 4. Distribution of discrepancies between the number of model-based simulated diagnoses in each of Garden's categories and the number in the empirical data.

sampler was deemed satisfactory using the F-test diagnostic developed by Gelman and Rubin [11], which is a part of the standard SCORIGHT output. We kept every 20th draw after the initial burn-in so that the draws can be considered independent; standard variance estimates of the estimated parameters therefore are reasonable. Users of MCMC estimation might be surprised by the number of iterations required for the process to converge, since a more typical number is 10 000 [12]. It takes this long because the data set is so small; thus, the movement away from the structure specified by the priors to the final solution is slow. Happily, even though the small data set requires a large number of iterations, its small size also means that each iteration is completed quickly. The total time needed for this analysis (on a personal computer with an Intel 1.86 GHz dual-core processor and 2 GB of RAM) was 75 minutes.

One strategy for assessing model-data fit is to compare model-predicted outcomes with the data. We used the model to predict the number of responses in each of Garden's four categories by randomly sampling the multivariate posterior and simulating a response for each radiograph/orthopedist interaction based on the sampled values. By repeating this sampling/simulation process a large number of times, we obtained a distribution of discrepancies for each diagnostic category: number of simulated diagnoses in each category minus the number of observed diagnoses. Figure 4 shows these distributions for each of Garden's four categories.

If the model-data fit is reasonable, zero (i.e., *no* discrepancy) should be plausible given these distributions. This indeed was observed, with the expected discrepancy in all cases very near-zero. This result suggests that the model is consistent with the observed diagnoses.

6. THE RESULTS

The fitted model generates four trace lines for each orthopedist. Figure 5 shows the posterior mean value of these curves for orthopedist C (taken point-wise). The idea, depicted in equation (1), is that as the severity of the fracture increases, the likelihood of it being classified as a category I fracture declines and its likelihood of being placed into category II increases. As the severity increases

DOI: 10.1002/sim

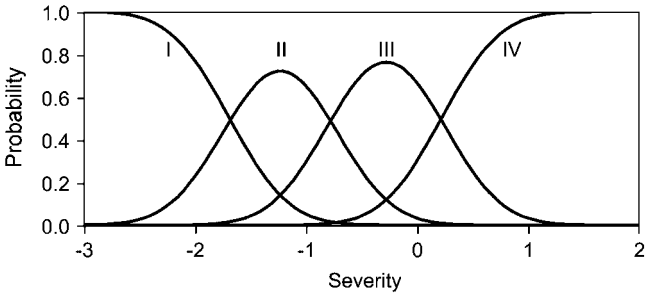

Figure 5. Tracelines for a typical orthopedist: the empirical structure of hip fracture classification for orthopedist C.

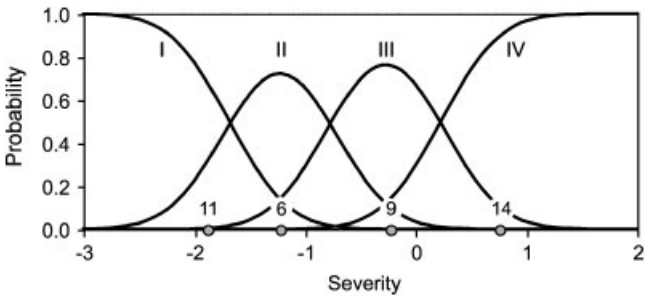

Figure 6. Plot of orthopedist C's tracelines with the posterior mean locations of four radiographs (6, 9, 11, and 14) superimposed.

still further, its likelihood of being classed in category II starts to decline and its likelihood of being placed into category III increases. Finally, if its severity is judged more serious still, the probability of it being classed as a category IV increases. Note that the intersections of adjacent curves are the points at which the expected diagnosis changes. A set of tracelines like these is generated from the model for each orthopedist.

The power of the model is that it describes, in a probabilistic way, what happens when an orthopedist meets a radiograph. Each radiograph occupies a specific position on the severity axis. In Figure 6 we illustrate this with four different radiographs that span the severity range. These are denoted as 6, 9, 11, and 14 and are represented by small circles on the horizontal axis.

Radiograph 11's position indicates that this orthopedist is expected to deem the fracture minor, with a high probability (0.70) of categorizing it as a I. Radiograph 6 is expected to be judged as somewhat more severe, with only a small likelihood of being classed I or III, but a higher likelihood of being a II. Radiograph 9 is expected to be judged still more severe, equally likely to be a II or a IV, but most likely to be judged III. Last, radiograph 14 is the most severe and would be most likely judged a IV with a much smaller chance of being called a III. Incidentally, for these radiographs, orthopedist C did indeed conform to model expectations classifying radiographs 11, 6, 9, and 14 as I, II, III, and IV, respectively.

DOI: 10.1002/sim

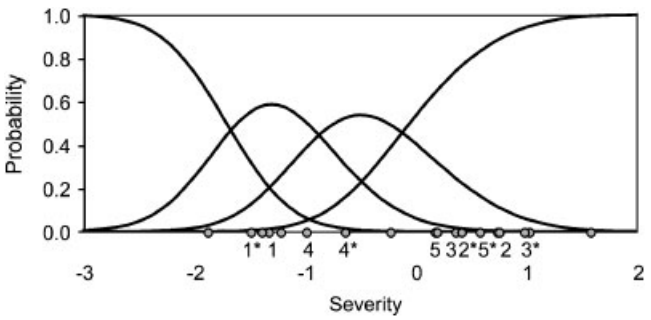

Figure 7. Empirical structure of hip fracture classification averaged over all twelve orthopedic surgeons with the locations of the radiographs superimposed.

In Figure 7 we show the average tracelines—averaged over the 12 orthopedic surgeons—and the 20 radiographs. We note that the radiographs are divided primarily into categories I and II or categories III and IV. There are only two cases where the model predicts considerable disagreement among physicians about the appropriate treatment (i.e., a disagreement as to whether the fracture should be classified as type II or type III). It is important to note that these two cases, labeled '4' and '4*' in Figure 7, were the same radiograph presented on two occasions. Furthermore, the uncertainty in the model-predicted diagnoses accurately reflects the uncertainty in the observed diagnoses: on the first presentation, 8 out of 12 surgeons classified the radiograph as a II and 4 classified it as a III; on the second presentation, 5 classified it as a II, 6 classified it as a III, and one classified it as a IV. For the remaining cases, there may be some confusion between being classified as a I or as a II, and sometimes between categories III or IV, but there is a little disagreement over what the treatment (pinning or surgery) ought to be. Moreover, the two occurrences of those cases that were repeated are always close to one another.[§]

Recalling the clinical question posed above about diagnostic consistency, the stability of the repeated cases' diagnoses shown in Figure 7 suggests that when presented with a radiograph on more than one occasion, these surgeons *collectively* reach similar conclusions. However, it does not follow that for individual surgeons diagnostic consistency should be expected. For example, while we observed in Figures 5 above that orthopedist C had a high probability ($p=0.70$) of classifying radiograph 11 as a I, this is hardly unequivocal: 30 per cent of time this surgeon would classify radiograph 11 differently. Moreover, for radiographs near the intersection of adjacent tracelines, even greater within-surgeon *inconsistency* is expected (after all, where curves intersect, every category has a probability of selection less than 0.50). So, to generalize, we see that within-surgeon consistency is (i) conditional on radiograph severity and (ii) low enough in some cases that multiple diagnoses may be sensible. We next discuss consistency.

[§]The standard errors of the parameters (which we did not present) indicated that all cases were well-estimated except for case 1. This was the first radiograph shown suggesting that perhaps it was needed by the orthopedists for acclimation to the study. When its replication, 1*, was presented it had about the same estimated severity, but with a much smaller standard error. Thus, even for case 1, which is a bit anomalous, the repeated presentation yielded a similar estimate of severity.

DOI: 10.1002/sim

7. ON CONSISTENCY

So far we have described the results from a stochastic model that we fit to a small data set. We showed that a severity continuum can be constructed and radiographs can be located on it. We have characterized surgeons in terms of their probability of making certain diagnoses as a function of radiograph severity. Was this worth the analytic effort? If all we get from the model fitting is what we have described so far, the answer must be equivocal. In fact, we can get similar results by simply reorganizing the data table that we showed previously as Table I. Such a reorganization, following the directions found in chapter 10 of Wainer [13], is shown in Table II. The rows and columns are reordered by the means, the re-administration of the first five radiographs is spaced apart, the row and column means are shown, as are the column variances, and some diagnoses of interest are emphasized in **bold**.

If we compare the location of the radiographs (cases) shown in Figure 7 with their means in Table II, we find complete agreement. All cases seem clearly defined as being judged to need a pin or a replacement with only case 4 (and its replication 4*) as being poorly determined. So, what does the measurement model give us that we couldn't get from carefully studying Table II? To answer this question, we must first show how the formal model handles within- and across-surgeon inconsistency.

We observed above that for some radiographs, inconsistent diagnoses from a single surgeon may be expected. This uncertainty is modeled as within-surgeon inconsistency and is made explicit by the stochastic model.

So, how consistent are surgeons' diagnoses? Recall that Figure 5 presented the tracelines for a typical surgeon, orthopedist C. These tracelines described the relationship between radiograph severity and the probability of each diagnosis. Hypothetically, we could ask ourselves: what is the probability of orthopedist C making the same diagnosis twice if presented with the same radiograph on two occasions? Assuming independence across occasions, this probability is equal to the sum of the squared tracelines, $P = \sum_{r=I}^{IV} P(Y = r|d_C, t_C)^2$, and is shown in Figure 8.

For radiographs of middling severity ($-2 < \theta < 0.5$), orthopedist C is expected to reach the same conclusions about case severity only around 54 per cent of the time when presented with the same radiograph on two occasions (this is represented in Figure 8 by the dashed line). More generally, this example helps to explain why for the five repeated cases, only 58 per cent of all within-surgeon judgments remain unchanged across occasions. Clearly, for many radiographs, the reliability of a single diagnosis is often in doubt.

We defined within-surgeon consistency as the amount of agreement across occasions when a surgeon is presented with the same radiograph. We could take the same approach to across-surgeon consistency and simply define it as the amount of agreement across different doctors when presented with the same radiograph. This definition is not ideal, however, because the amount of agreement across doctors is confounded by the amount of consistency within each doctor. Thus, instead of comparing how surgeons actually behave, it may be more useful to compare how we *expect* them to behave. In this way we don't allow model-predicted within-surgeon inconsistency to confound our inferences about the similarity of surgeons.

Returning again to orthopedist C, imagine some hypothetical surgeon, say, orthopedist X, whose tracelines are identical to orthopedist C's tracelines. What can we say about the consistency of these two surgeons? Clearly, we cannot say that they will make the same judgments (as we just observed, orthopedist C cannot even be relied on to provide consistent diagnoses with himself). However, because they are described by identical tracelines, we can say that their *expected* (rather

DOI: 10.1002/sim

Table II. The original data matrix with rows and columns reordered by mean size and unusual points emphasized.

	Case																			
Doctor	11	15	1	6	4	9	5	13	3	7	2	14	12	8	10	1*	4*	2*	5*	3*	Means
D	1	2	2	1	2	**2**	4	3	3	3	4	4	3	3	4	2	**2**	3	3	4	2.8
G	1	2	1	1	2	4	**2**	3	4	4	3	3	4	4	4	2	**2**	3	4	4	2.9
F	1	2	1	2	2	3	3	3	3	4	4	3	4	4	4	2	3	4	3	3	2.9
J	1	1	1	2	2	4	3	3	4	3	4	4	4	4	4	1	4	3	4	4	3.0
C	1	2	2	2	2	3	3	3	4	4	4	4	4	4	4	1	3	3	4	4	3.1
I	1	2	**4**	2	2	4	4	3	3	4	3	4	4	4	4	2	**2**	4	3	4	3.2
H	2	1	1	**3**	**3**	3	3	4	4	4	4	4	4	4	4	1	3	4	4	4	3.2
K	2	1	**3**	2	**3**	**2**	4	4	**2**	4	4	4	4	4	4	**3**	**2**	4	4	4	3.2
B	2	2	1	2	**3**	3	4	4	4	4	4	4	4	4	4	1	3	4	4	4	3.3
A	1	2	1	**3**	**3**	4	4	4	4	4	4	4	4	4	4	2	3	4	4	4	3.4
L	1	**3**	**4**	2	2	3	4	4	4	4	4	4	4	4	4	2	**2**	4	4	4	3.4
E	2	2	**3**	**3**	2	3	4	4	4	4	4	4	4	4	4	2	3	4	4	4	3.4
Means	1.3	1.8	**2.0**	2.1	2.3	3.2	3.5	3.5	3.6	3.8	3.8	3.8	3.9	3.9	4.0	1.8	2.7	3.7	3.8	3.9	3.1
Variance	0.2	0.3	**1.5**	0.4	0.2	0.5	0.5	0.3	0.4	0.2	0.2	0.2	0.1	0.1	0.0	0.4	0.4	0.2	0.2	0.1	0.0

DOI: 10.1002/sim

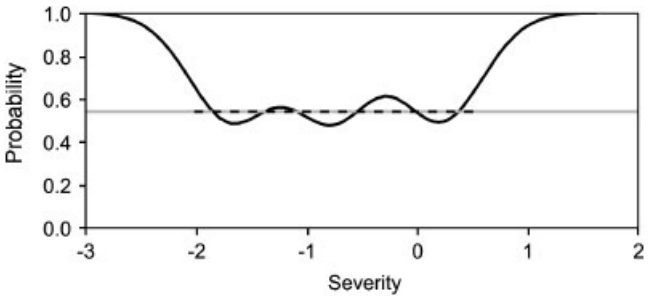

Figure 8. Probability of orthopedist C making the same diagnosis twice when presented with the same radiograph on two occasions, assuming independence across occasions. The dashed line shows the expected probability for severity greater than −2 and less than 0.5.

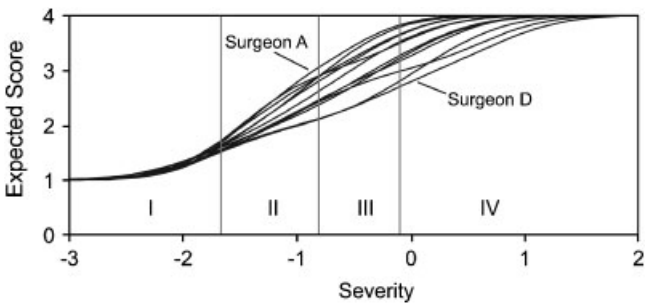

Figure 9. Expected score curves for the 12 orthopedic surgeons plotted with 3 plausible cut-points. The two surgeons who applied Garden's classifications most differently are labeled.

than *actual*) judgments are the same; or, in other words, that they apply Garden's classifications equivalently after accounting for within-surgeon inconsistency.

Instead of comparing the full set of tracelines for each surgeon, we can summarize each surgeon's tracelines with a unique expected score curve.[¶] Figure 9 shows these curves for our data set along with three plausible cut-points, which are represented by the three vertical lines. (In lieu of actual standard setting, each cut-point was defined as the intersection of each adjacent category's mean traceline, averaged across all surgeons. These three intersections can be seen in Figure 7.) Ideally, all surgeons would apply Garden's classifications in the same manner and it would make no difference which orthopedist one consults. If this was so, the expected score curves in Figure 9 would be identical across surgeons excepting estimation error. Since this was not what we observed (e.g., the curves for surgeons A and D), it is natural to ask: where are the differences most pronounced and how important are they?

The 'where' is easily observed. Like within-surgeon consistency, across-surgeon consistency is conditional on radiograph severity. The greatest differences can be seen near the cut-points for categories II–III and III–IV, in the region between −1.25 and 0.50. Expected scores are similar

[¶]These curves are formed by multiplying the score at every level of severity by its probability and summing them: $E(Y|\theta)=\sum r \times P(Y_i=r|\theta)$.

DOI: 10.1002/sim

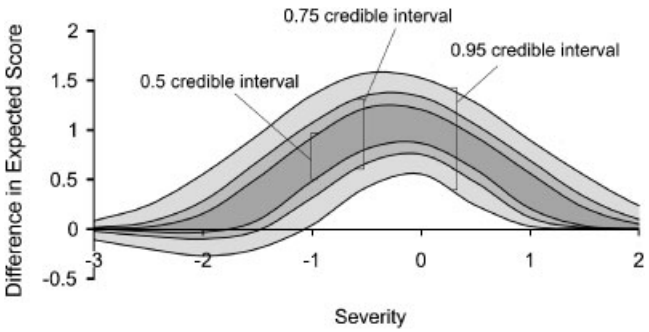

Figure 10. Three credible intervals for the distribution of expected score differences between surgeons A and D plotted as a function of radiograph severity.

at the bottom of this region, become increasingly dissimilar as severity increases until severity ≈ -0.25 and then become gradually more similar for severity > -0.25. At their most variable (≈ -0.25), expected scores vary from 2.5 to 3.7.

We have shown where on the severity continuum to expect across-surgeon inconsistency, but evaluating the importance of these differences is less straightforward. Let us consider the case of two surgeons, A and D, chosen because their expected curves were among the most different (as shown in Figure 9). Because we obtained samples from the marginal posterior distributions of the parameters, we can produce a set of posterior draws for each surgeon's expected curve for any given severity. We then can sample these expected score curve posteriors, which allows us to describe the difference between the surgeons' expected diagnoses probabilistically by creating a probability distribution of differences. If two surgeons applied Garden's classifications consistently with one another, we would expect this distribution of differences to have a mean of zero. When the expected difference is not zero, credible intervals can be constructed showing the probabilities associated with differences of various magnitudes. Figure 10 contains such an analysis showing three such credible intervals (the middle 0.50, 0.75, and 0.95) for surgeons A and D as a function of severity.

Figure 10 suggests that surgeons A and D probably do not apply Garden's classifications consistently for radiographs with severities between -1.0 and 1.0. Here, the greatest differences are expected when classifying radiographs of severity ≈ 0.0; the probability of observing a greater than 0.55 difference in expected scores is 0.975.

8. CONCLUSIONS AND DISCUSSION

In addition to showing that surgeons can be characterized by their probability of making certain diagnoses as a function of the location of radiographs on a severity continuum, we showed that the model provides a useful framework for examining consistency both within and across surgeons. Using this framework, we have concluded that the orthopedists who judged the radiographs were reasonably consistent in their application of Garden's classifications at the ends of the severity scale, but somewhat less consistent for radiographs with medium or upper medium severity. Moreover, we found that individual diagnoses could easily be contaminated by random effects. However, the formal model does more than provide a framework for examining consistency; the model

DOI: 10.1002/sim

automatically adjusts for both within- and across-surgeon inconsistency. Next, we illustrate this powerful feature with an example.

By excluding case 6, re-estimating the model, and rescoring case 6 using the re-estimated surgeon parameters, the model yields the probability of this radiograph falling in each of the four categories as

I	II	III	IV
0.00	0.99	0.00	0.00

These probabilities were computed by first estimating plausible classification cut-points (in the manner described above) and then looking at the proportion of radiograph 6's severity posterior distribution that falls in each resultant category. If we assume, for the sake of this example, that the cut-points are estimated without error, there is little question that case 6 falls in the II category, and so if we had 12 different opinions for this case we would feel confident that pinning the fracture was the best course of treatment.

But let us try an experiment. Although nine doctors diagnosed case 6 as a I or a II, there are three doctors who judged case 6 as a III. If we rank order orthopedists' overall propensity for severe diagnoses,[‖] we see (as we might hope) that the three surgeons who recommended surgery for radiograph 6 (orthopedists A, E, and H) were also the three surgeons most inclined to recommend hip replacement surgery in general. Now, suppose we omit the nine surgeons who classified radiograph 6 as anything other than a III; surgeons A, E, and H are the ones who remain, and if we went no further the patient would have a hip replacement in his immediate future. But if we use the model, it automatically adjusts for the severity of those three judges and yields the probabilities of case 6 falling in each of the four categories as

I	II	III	IV
0.00	0.28	0.70	0.02

This case's location on the severity scale has shifted to the right, but not completely because the model adjusts for the fact that these three judges had a propensity for severe diagnoses.[**] In other

[‖]Each surgeon's expected score curve (shown in Figure 8) may be approximated with a two-parameter logistic function

$$Y = 1 + \frac{3}{1 + e^{-g(x-h)}} \tag{2}$$

where Y is the expected score, ranging from 1 (I) to 4 (IV), for a radiograph with severity x, g characterizes the slope of the logistic function, and h its location (point of infection) on the severity scale. By using their h-parameter estimates, orthopedists can be ranked by their overall propensity for severe diagnoses.

[**]To give a sense of the magnitude of this adjustment, consider the hypothetical case wherein *every* surgeon classified radiograph 6 as III. If we take every possible combination of three surgeons (of which $12!/[3!(12-3)!] = 220$ exist), the resultant median probabilities are

I	II	III	IV
0.00	0.09	0.75	0.13

We see that the result based on surgeons A, E, and H is considerably more ambiguous about surgery than we would expect from a group of three randomly selected surgeons who made the same judgments: the probability of pinning increases from a median of 0.09 to 0.28.

DOI: 10.1002/sim

words, when estimating the severity of a radiograph, the model accounts for differences in the application of Garden's classifications (i.e., across-rater inconsistency). The model further adjusts for *within*-rater inconsistency; there is considerably more diagnostic uncertainty associated with having only three diagnoses instead of 12. With three diagnoses, the random effects of within-rater inconsistency could explain the surgeons' decisions; with 12 ratings, much of the within-rater stochastic variability cancels out. The result is that Case 6 is not a clear hip replacement but falls on the boundary between II and III with the probability of pining adding up to $0.28(0.00+0.28)$. Prudence would suggest that we seek additional opinions when we find a boundary case like this. In this case, those opinions are likely to be the Is and IIs that we had omitted previously. These automatic adjustments are not easily available without an explicit model.

This result highlights our conclusion *vis-à-vis* the first research question that motivated this research: Does Garden's four-category classification scheme contain distinctions between categories that are too fine to be useful? We conclude that the answer to this is 'no,' but for a subtle reason. Certainly we saw that variations in ratings for a specific radiograph tended to be primarily within categories that yield the same treatment choice. But sometimes ratings cross the II/III barrier. If only two categories were to be used, say 'Pin' or 'Replace', and we had a disagreement among raters, we would have no way to judge how firmly the raters held their positions. With Garden's four categories, we can see (as with Case 6) that some ratings of 'I' could overbalance some of 'III' and provide a more precise estimate of the severity of the fracture. The fewer the number of surgeons that rate the case, the more important the extra information gained from this classification scheme becomes.[††]

We are not unaware of the additional possible applications of this methodology. Had the contents of Table I been generated by four ordered categories of mitral valve cardiac murmurs, this paper could have been retitled 'Hearty Psychometrics' and have few other changes. We also are mindful of the possibilities for misuse that present themselves through the combination of a strong model, a set of priors (even if they are diffuse) and very few data points. Yet, the apparent advantages of modeling these data have convinced us to disseminate our findings.

9. AN OBITER DICTUM ON SAMPLE SIZE

Throughout this paper we have focused on the specific problem of evaluating the diagnostic judgment of orthopedic surgeons on a set of radiographs of hip fractures. We concentrated our attention on the statistical model we used and the results it yielded. Left implicit, so far, was what many may perceive as the most remarkable aspect of this research: the sample size. Most experts in item response theory do not believe that this sort of scoring model can be used with a data set this small. At a recent psychometric conference, we surveyed what might be considered 42 of the most eminent psychometricians and statisticians in the world. We asked them to estimate what would be the minimally acceptable sample size to fit a 4-choice polytomous IRT model to 20 items. The results are shown in the box-and-whisker plot below (Figure 11). The middle 50 per cent of the audience felt that the minimal size should be between 500 and 2000, with a median of 1000.

[††]We make no assertions about whether the individual Garden categories are distinct on a *biological* basis. It well may be that some fractures are type I and some type II but that these are merely way points on the underlying continuum of severity and have no particular biological significance.

DOI: 10.1002/sim

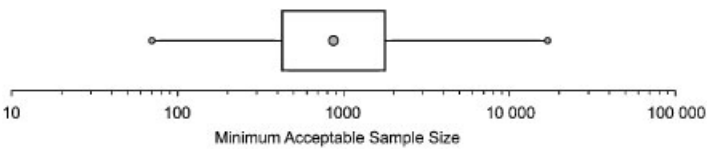

Figure 11. Distribution of responses from 42 eminent psychometricians and statisticians who were asked to estimate the minimum acceptable sample size needed to fit a 4-choice polytomous IRT model to 20 items.

The discrepancy between these survey results and our success with only 12 subjects suggests that the expansion of the range of applicability of these kinds of models is one of the principal results of this research.

REFERENCES

1. Wainer H. Pyramid power: searching for an error in test scoring with 830 000 helpers. *The American Statistician* 1983; **37**:87–91.
2. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monographs* 1969; Whole No. 17.
3. Garden RS. Low-angle fixation in fractures of the femoral neck. *Journal of Bone and Joint Surgery* 1961; **43-B**:647–663.
4. Müller ME, Nazarian S, Koch P, Schatzker J. *The Comprehensive Classification of Fractures of Long Bones*. Springer: Berlin, 1990.
5. Wainer H, Bradlow ET, Wang X. *Testlet Response Theory and its Applications*. Cambridge University Press: New York, 2007.
6. Wang X, Bradlow ET, Wainer H. User's guide for SCORIGHT (version 3.1): a computer program for scoring tests built of testlets. [Research Report]. Educational Testing Service, Princeton, NJ, 2004.
7. Gelfand AE, Smith AFM. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* 1990; **85**:398–409.
8. Patz RJ, Junker B. A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics* 1999; **24**:146–178.
9. Patz RJ, Junker B. Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics* 1999; **24**:342–366.
10. Bradlow ET, Wainer H, Wang X. A Bayesian random effects model for testlets. *Psychometrika* 1999; **64**:153–168.
11. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992; **7**:457–511.
12. Sinharay S. Experiences with MCMC convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics* 2004; **29**:461–488.
13. Wainer H. *Visual Revelations*: *Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot* (2nd edn). Lawrence Erlbaum Associates: Hillsdale, NJ, 2000.

DOI: 10.1002/sim