

Peter Fayers

IRT: The way forward

In many ways the use of IRT methods in health-outcomes assessment is one of the most exciting developments of recent years. In my opinion, its impact is as significant as the introduction of logistic models and logistic regression, which first became so popular in medical research in the 1980s as a replacement for analysis based upon mean values. Indeed, there are a number of obvious parallels between logistic regression and IRT models. Both were initially introduced as a means of handling binary outcomes, and both replaced the analysis of mean values for continuous data with, instead, the estimation of probabilities of a binary response. And I think the analogy can also be extended because logistic regression was at first slow to be adopted by researchers because of the initial lack of user-friendly software. Indeed logistic regression only became popular in the nineteen eighties because facilities were at that time becoming standard options in many of the mainstream statistical packages.

Similarly, the seminal paper on Cox models for survival was published in 1972 – again, one of the most important breakthroughs in medical statistics, and nowadays one of the most widely used of techniques. But I well remember in the 1970s having to use DOS based programs with exceedingly unfriendly user interfaces for these purposes, and Cox models too only became popular when endorsed by the mainstream statistical packages such as SAS, STATA, SPSS and S-Plus.

Of course, that is exactly the stage at which we are at with IRT software – it is not yet available in the mainstream packages, and users must purchase standalone software, some of which has decidedly unfriendly interfaces. There are currently a large variety of different specialist IRT packages, and unfortunately when analyzing the same dataset it is very difficult to obtain identical results from several packages. In many cases one package may converge to a solution one another package failed to do so. For some packages one has to specify such things as “quadrature points” – not something that most users are really interested in. And one may need to play around with various convergence criteria and accelerated approximations using Newton methods, etc. Different packages can only fit particular models, and if one model does not fit adequately the user may need to turn to a different package instead. Only when all these issues are resolved can the researcher make full use of this magic black box. At present one needs to be far too skilled in computational methods and in the black art of data manipulation. There is clearly a need for software improvements.

I suggest a major area for future work is the development of software that is reliable, easy to use, can fit the range of IRT models that are used in PRO research, and has extensive graphical displays. Ideally this software should integrate with one of the major standard statistical packages.

Returning to the analogy with logistic regression, medical statisticians next had to grapple with the problem of data that was neither continuous (when traditional linear regression models could be used) nor binary (where logistic regression was appropriate); that is to say, ordered categorical data. The solution for logistic regression, as with IRT models, was to generalize the models and develop a new generation of software that can deal with ordinal data. However, both for IRT and logistic regression, the models become much more difficult to explain to non-mathematicians and also more difficult to interpret. The fitting of these models requires much larger sample sizes, and they appear to be considerably more fragile than simpler models. At present, investigators often try to fit one model – say, GRM – and if, after juggling the number of quadrature points and other estimation parameters, the model fails to fit then they might try to fit another model – say, GPCM. And, with another data set examining the same outcomes, perhaps GRM fits best after all. It would be far more satisfactory if there were prior justification regarding the choice of appropriate model.

There is need to explore ways of communicating the advantages of the various polytomous models, and to explain in lay terms the clinical significance of the coefficients and the interpretation of the models. There should be guidelines about selection of models.

I have emphasized the need for better and more widely available software, with improved graphical facilities. The current software limitations, combined with the initial enthusiasm for IRT as a panacea in test development, has resulted in a number of publications in which enthusiasm perhaps outweighs prudence. One must read publications critically, to separate the sound papers from the many that are less convincing or even cavalier in their approach to analysis. Thus, I wonder how many users of IRT to systematically check the basic assumptions for their models, and how many users seriously consider the sample size implications? Local independence (conditional independence) and uni-dimensionality are of particular importance for many of the computational algorithms that are used to fit IRT models and estimate the parameters. Both of these assumptions are usually essential, but they may be very dubious for particular PROs, and are frequently ignored in the naïve application of IRT methods. For example, symptoms may be highly correlated with each other because they are treatment or disease related, and in terms of their impact on quality of life they will not be conditionally independent. Thus models that may be perfectly reasonable for the multiple items in, say, a fatigue scale, may be inappropriate for scales that combine a number of symptoms in a summary symptom score. Currently, there is still relatively little information about the robustness of the IRT models and the computer algorithms for fitting them. We simply do not know how sensitive the models are to violation of the various assumptions. Of course, one reason why the assumptions are all too frequently ignored is, again, the lack of adequate software. Most IRT software provides only limited graphical or other diagnostic facilities. And sometimes the output that current software does provide seems to me to be unnecessarily abstruse.

There are three major areas for future work. Firstly, there is a need for theoretical and empirical research into the robustness of the models when applied to PRO data, and their sensitivity to the inherent assumptions. Secondly, there is a need for training courses and tutorial articles to emphasize the testing of assumptions and the use of diagnostics. Thirdly, IRT software should provide a better range of easily-comprehensible diagnostic output, including summary statistics and graphical displays; model-fit, item-misfit, person-misfit, and coefficients for conditional independence would seem to be some of the most basic essentials.

I mentioned sample size. Nowadays, it is well-known that far too many of the early survival studies were too small to be able to detect the clinically relevant treatment differences that are realistic and plausible. Over recent years there has been much publicity about this, and nowadays the main medical journals always demand that pre-study sample size calculations must be reported in detail. I wish the same applied to publications that used IRT methods. At present, very few papers make any comment about the sample size requirements of their investigation. I think far too many of the publications are based on derisively small sample sizes. Many authors gaily fit complex ordinal models to small datasets. But at least in the case of survival analysis we know how to calculate the power of the test. However, in the case of IRT, very few books offer advice beyond making vague cursory comments. For example, Embretson & Reise (generally an excellent book) provide but half a page on this important topic. Based on a simulation study, they suggest that around 500 examinees be used to estimate the GRM. They follow this by pointing out that “simulation studies are useful only if the data matches the simulated data,” and suggest that researchers should also have enough subjects to make standard errors of parameter estimates reasonably small. Other books on IRT also provide crude rules of thumb, based on practical experience. In general most authors recommend hundreds of patients for simple models, and larger numbers for models with larger numbers of parameters. Thus it is hardly surprising when PRO publications fail to report how they evaluated sample size requirements – this is a consequence of a lack of theoretical work about sample size requirements.

Future research on sample size requirements is badly needed. I hope this area will become addressed in the future, with clear guidelines becoming available and perhaps software to facilitate the calculation of sample size requirements. Fortunately much software already provides standard errors (or confidence intervals) for the parameter estimates; these are essential, and should be reported in publications.

Undoubtedly IRT models are one of the most exciting developments in our field, despite being widely abused. Of course, until recently much of the work on IRT came from other fields of application such as education – which is why we still use terms such as "difficulty". I think it is instructive to consider in a little detail just why it was that Rasch models and IRT models first became popular in education. In the first place, in education it is natural to think of binary test items that are marked "right" or "wrong". Secondly, in education one is usually assessing a single dimension such as mathematical

ability. Thirdly, it is usually natural to have a large number of test items measuring these dimensions – school tests usually contain a lot of items. And finally, in educational tests one can usually select items that fit the IRT model, and reject those that do not fit the model well.

Contrast this with patient reported outcomes. Some dimensions such as pain or fatigue may indeed be so important that one wishes to use a multi-item test. But many PROs may be simple symptoms or other complaints. Instead of lengthy multi-item tests assessing a single dimension, clinicians may wish to use a large number of single item questions that assess many dimensions. Such questionnaires are not really suitable for IRT methods – one *can* use IRT, but the methodology is largely unnecessary and unhelpful. Also, instead of binary items, the clinician will commonly want symptoms and other outcomes to be assessed using ordinal scales such as “not at all”, “a little”, “quite a bit”, “very much” etc. And finally, if it is important to measure the particular symptom, there may be limited possibilities for adapting the questions to fit the model. In medical research we usually want to select a model that fits the data, not the other way round. Thus, I think, IRT methods are really most relevant for the sort of clinical scale that contains many items assessing a single dimension – such as questionnaires for pain, fatigue, depression, etc. Of course, one should distinguish between IRT-based CAT methods as opposed to more general computer aided testing. It is perfectly reasonable, without the use of IRT, to use interactive computer programs that expand out relevant symptoms for particular patients. In this case the software would be programmed to simulate a clinical interview so that, for example, a patient receiving treatment for a particular form of cancer would be presented with a list of relevant symptoms for that site and for the treatment they are being given. Each symptom could be graded for severity, and where appropriate questions could be asked about duration or frequency. None of this requires IRT, it merely mimics the process that a clinician might use when interviewing a patient about their problems, and identifies symptoms that are of consequence.

Perhaps there is a need for follow-up conferences to explore and discuss the role of IRT in different types of PROs, and to prepare circumspect guidelines. For some purposes clinicians may prefer symptom checklists for which a number of single-item questions may suffice to assess multiple symptoms; then IRT may have little to offer.

When considering global dimensions such as quality-of-life, there is another major difference between clinical research and educational testing. In educational testing we wish to assess ability in terms of how well the student can perform in the external world, relative to others. We require an objective assessment, and if items are to be weighted the weights should be chosen from a population perspective. In contrast, with health-related quality of life, it is up to the patient to choose his or her own weights, corresponding to preferences, values and utilities. The integration of functioning and symptoms into quality-of-life is essentially an internal and subjective weighting by the individual patient. I do not see how one can try to relate these internal judgments to objective external criteria. We still need a lot

more research into how to assess subjective internal feelings, and how to use – or even whether to use – IRT methods to combine multiple items such as symptoms into symptom scores, or to combine items and dimensions into a global HRQOL score. And I would just comment in passing that recently a number of authors have written papers proclaiming that perhaps visual analogue scales and numerical rating scales are just as good as a multi item scales, especially for such dimensions as overall quality-of-life (e.g. de Boer et al, *Quality of Life Research*, 2004), pain analgesia (e.g. Collins et al., *Pain*, 2001) and mood (e.g. Bernhard et al., *Br. J. Cancer*, 2001). Bernhard et al. have suggested that although the possible loss of reliability by using a single global item may be important when evaluating single patients on a cross-sectional basis, it may be much less relevant for large groups of patients in longitudinal comparisons, as in clinical trials.

When should “global questions” be used, enabling patients to use implicit personal preference weightings to say how their symptoms affect them overall? When should IRT-based methods be used to combine symptoms into summary scores?

Given these above reservations about the application of even simple IRT methods, it may be no surprise that I say I am rather anxious about the recent developments in multidimensional IRT. At one level it seems quite obvious that if two dimensions of quality-of-life are highly correlated, then we should make use of data regarding one time mention when assessing the other dimension. But consider two dimensions, say, emotional function and physical function. We know these are correlated dimensions, a patient with good emotional function will tend to have good physical functioning and vice versa. Therefore, the physical function items do provide extra information about emotional functioning. Multidimensional IRT can make use of this information, in a most efficient manner. But let us consider two patients, reporting equal levels of emotional function. If we measure the physical function of both these patients, and one has better physical function than the other, is it really sensible to argue that therefore one patient has better emotional function too? Because the physical items are correlated with emotional functioning and are hence informative about it, that conclusion would be the implication of multidimensional IRT. I find that none of my clinical colleagues have much belief in such curious reasoning. Instead, they say that they are very interested in those patients where there is an apparent discrepancy between the dimensions of emotional function and physical function, and would wish to keep assessment of the dimensions separate. Like my clinical colleagues, I find it very difficult to get my head around the logic of multidimensional IRT scales.

One might note an analogy here with preference measures. Health economists have long explored methods of eliciting values and preferences, and using these to combine PRO dimensions into summary indexes. Perhaps research into multidimensional IRT should investigate potential links with the theory of preference measures.

The potential of multi-dimensional IRT remains contentious ...

So far I have been talking largely about IRT models for developing measurement scales. Another major application of IRT models is differential item functioning. This is an exciting application that is particularly relevant when developing scoring algorithms to be used across different subgroups, such as different cultures, for different age groups, or for the two genders. It provides an excellent method for cross-calibrating scales, to ensure fair assessment in these subgroups. However, here too I think we need a lot more work about how to interpret and communicate the relevance of the "effect sizes" that are reported when DIF is reported. And this links too with the need for more work on sample size estimation: if the sample is too small, DIF will either not be detected or will not be statistically significant; but on the other hand, if the sample size is large enough, one is almost certain to detect statistically significant DIF – yet the detected DIF may be far too small to be of any clinical relevance. When using IRT (or other methods) to test for DIF, how many users explore the magnitude of the DIF in terms of its clinical relevance? How often is observed DIF large enough to represent a clinically meaningful difference?

In some instances the detection of DIF may be an indication that an item is simply inappropriate and should clearly be dropped – for example, when measuring physical activity an item using the playing of American football would clearly show DIF for gender and is surely a foolish item. But sometimes it is far from clear what to do about DIF. Men *do* report less pain than women, and so we may expect to find DIF in scales that combine pain with other outcomes. But does that mean we should or should not make a gender adjustment for individual patients, and should we let such adjustments influence the treatment of pain?

DIF is particularly important for examining test bias with respect to culture, gender, age and other group differences. But there is need to develop guidelines about the interpretation and communication of DIF, and about whether or how PROs and summary scales should be modified when DIF has been detected. Further research into the methodology and implications of DIF remains important.

Finally, I would just like to express a note of caution about the whole-hearted use of IRT in instrument development. Firstly, IRT does not replace traditional psychometrics, but is merely an additional tool, albeit a powerful one. Traditional methods – and traditional subjective scales – still retain an important place in the assessment of PROs, and in my opinion rightly so. But even more importantly, no amount of IRT or other psychometric analyses can salvage a poorly conceived instrument. When constructing a grand house, it is essential to build on solid foundations – or the whole edifice may collapse. Similarly with PROs – if secure foundations are well thought out and carefully constructed, the resultant instrument is likely to be sound. I am of course referring to face validity and content validity. If these are weak, the final instrument will be of limited value – to use an old cliché, you cannot make a silk purse out of a sow's ear. Furthermore, it should be noted that *quantitative* methods such as IRT are of limited value for confirming or disproving face and content validity, so the developers will be blissfully

unaware of the limitations of their work. But if *qualitative* procedures for ensuring face and content validity were rigorously applied during the initial development, the instrument should be fine irrespective of the complex IRT models that are subsequently applied. There tends to be a chasm between qualitative scientists and those from a quantitative background. Despite being a statistician, I deplore this divide. When developing a new questionnaire, it is crucial to ensure that there is substantial investment in the early-phase qualitative research. Clinical doctors, nurses, psychologists, psychiatrists – and especially patients themselves – are amongst the many whose opinions should be elicited and who should be included in the qualitative studies that precede any numerical psychometric investigations.

IRT provides additional tools to supplement, but not replace, other psychometric methods. None of these tools are supplant the need for careful and rigorous qualitative methods of ensuring face and content validity. A poorly conceived instrument cannot be salvaged by IRT. Guidelines should emphasize good practice in early-phase qualitative methods.

In summary, IRT and CAT two extremely powerful and exciting tools to add to our toolbox when developing questionnaires for patient reported outcomes. It is in my opinion unfortunate that this has at times led to excessive enthusiasm. Perhaps greater circumspection is required. These tools do not replace the more traditional psychometric and statistical techniques, but provide additional and supplementary tools. However, in order to use these new tools effectively there is an urgent need for better software, for diagnostic tools within the software, and for research into the limitations of the models, the estimation of sample size, and the interpretation and communication of IRT analyses. In order to realize the full potential of these exciting methods, theoretical work and software development should be encouraged, while ensuring that academic IRT research is not purely abstract but is linked to and focused on the practical issues that arise in clinical assessment of PROs. This should be supplemented by critical review of the application of IRT, perhaps by a series of workshops, and the preparation of guidelines on usage and reporting.