Methods of Psychological Research Online 2003, Vol. 8, No. 3, pp. 1-22 Department of Psychology
Internet: http://www.mpr-online.de

On Choosing a Model for Measuring

Mark Wilson

University of California, Berkeley

This paper describes an approach to the issue of selecting a measurement model that is based on a comprehensive framework for measurement consisting of four conceptual building blocks: The construct map, the items design, the outcome space, and the measurement model. Starting from this framework of building blocks, the measurement model selected must conform to the constraints imposed by the other three components. Specifically, to preserve the interpretability of the construct map, the models must preserve the order of items throughout the range of person locations, and must do so in a way that is consistent with the interpretational requirements of the map. In the case of item response modeling, this translates into selecting models that have a constant slope—i.e., they must be from the Rasch family. In the conclusion, some next steps in investigating these issues are discussed.

Keywords: *item response theory, construct validity, Rasch model*

The central issue of this paper is the question: What criteria should I use for choosing a measurement model? This is an issue that arises in any field where measurement models are to be used to help guide the development of instrumentation. For instance, in the context of medical outcomes research, this issue has arisen in the fairly recent advent of item response models within that domain (Fisher, 1994; Ludlow & Haley, 1992)—for example, in a supplement to one of the leading journals in that area, *Medical Car*e, several papers were devoted to this development (Cella & Chang, 2000; Hambleton, 2000; Hayes, Morales, & Reise, 2000; McHorney & Cohen, 2000; Ware, Bjorner, & Kosinski, 2000). Within those accounts, several described possible choices of item response model (Cella & Chang, 2000; Hambleton, 2000; Hayes, Morales, & Reise, 2000), and one directly discussed the choice of a measurement model for a particular data set (McHorney & Cohen, 2000). To gain the advantages (Hambleton, 2000; Hayes, Morales,

Correspondence concerning this article should be addressed to Mark Wilson, Graduate School of Education, University of California, Berkeley, Berkeley, CA 94720, USA.
Email: mrwilson@socrates.berkeley.edu

Acknowledgements: The author wishes to thank Everett Smith for helpful comments in revising the manuscript.

& Reise, 2000) of this measurement approach, one must necessarily make a choice of measurement models, on some grounds. This paper is a description of one basis for such a choice.

A necessary precursor to answering this question is to decide what are the characteristics of a "measurement model", so the discussion will start by addressing that issue first. In deciding what should be the characteristics of a measurement model, one comes across two broad strategies: The first gives most prominence to the statistical nature of the measurement model, while the second gives most prominence to the nature of the measurement approach, and the nature of that approach effectively builds constraints into the measurement (statistical) model.

In the sections below, the following are discussed: First, the two approaches mentioned above are discussed and exemplified, and the consequences for choosing a measurement model in a dichotomous context are discussed; and, second, some consequences of this perspective and some possibilities for future work, are discussed in the conclusion.

In order to make this discussion reasonably contained, and also to focus on the crucial issues, a particular scope will be adopted throughout. There are many possible issues that one could focus on in a discussion of model choice—the issue that has been selected for this paper is the issue of how to score the categories of item responses. This issue has been chosen because it is present in almost all measurement applications, but it is also an issue that is fairly specific and straightforward to define. The discussion will be set in the context of what has been termed item response theory (IRT), or sometimes item response modeling, because in that area the issue of response category scoring is one that already has a reasonably long history. Thus we need to define strategies that respond to the question: What should a measurement model do?

There is a particular limitation to the discussion in this paper that should be noted at the outset: This is not a *derivation* (Fischer & Molenaar, 1995) of any particular measurement model—the aim is to help the researcher choose among competing models that are readily available—hence, the discussion is not carried out within a formal framework of mathematical theorem and proof. Instead the discussion starts with a researcher's intuitive idea of what it is that is to be measured, and sets up requirements for a measurement model on that basis. Mathematical derivations inevitably come around to the necessity to decide whether certain mathematical conditions should be insisted upon or not, and these abstractions are often hard to decide upon. The plan

here is a different one—to seek instead a conceptual basis for measurement, and to work out from there. If one disagrees with that basis, then one would, most likely, end up with a different conclusion, ergo, a different model.

Strategy One: Fit the data

Under the first strategy identified in this paper, labeled "fit the data", a measurement model is seen as no more than a statistical model (e.g., a multi-level latent variable statistical model) applied in a particular sort of context—a context where one wishes to infer something about an individual when one observes a set of responses to "items". This purpose is analogous to the function performed by many statistical models, so the fact that one is choosing a measurement model carries no particular strictures for the statistical model. Hence a measurement model needs to do what statistical models usually do: "...the successful model is the one ... most consistent with the observations" (Dwyer, 1983, p. 240). This can be operationalized in a number of ways, by, for example, explaining as much of the variance as possible, maximizing the likelihood, minimizing a prediction squared error, etc. In the majority of statistical modeling contexts there is a set of variables to include in the model that is limited either by the theory under which the study is being conducted, or by the actual availability of data. Given these circumstances, the strategy makes good sense—the modeler must make the most of what is available, and that corresponds to "fit the data". A recent description of this approach concluded: "the purpose of a theory in this approach is to explain the data" (Thissen & Orlando, 2001).

When the focus is moved specifically to the measurement situation described above, this approach will make sense if the items are interpreted as a fixed set of variables that the statistical modeler must make the most of. Note that this is contradictory to the commonly held idea that the items in a specific instrument are no more than a random sample from some universe of items (Guttman, 1944). Nevertheless, there are occasions where this situation might arise: such as where a traditional factory-style of item development process is in place—the items are developed somewhere outside of the range of communication and feedback with the statistical modeler, perhaps by a separate project team that does not have any say in the use of the items once they have been field tested and analyzed. Consequently, the modeler will seek a measurement model that is as forgiving as possible, as it would be uneconomical to lose items because of a "mere" model

choice. Thus the pressure is to use as complex a model as possible—the items are a constant, and the model must fit itself around them.

At this point, a reader might say: "No reasonable statistician would argue that a measure is better if it arose from a more complex model. Similarly, basics indicate that parsimony is a goal." If this were true, it would seem to negate the need for the current paper. However, another reader might say: "In my experience, estimated scores (θ) from a one-parameter [i.e., more parsimonious] model correlate about .98 with scores estimated from a two-parameter model [i.e., more complex]; under such circumstances, how much is really lost by using a two-parameter model for measuring?" In fact, these quotes are from two people who were asked to review this paper (but who will remain anonymous)—clearly there is indeed a considerable range of opinion on this matter from leading experts in the field.

Some specific consequences of this pressure towards complex models are:

(a) measures are considered *better* if they arise from more complex models using more parameters;

(b) you don't need to worry about having a substantive basis for the scores of different items—just have the data tell you what the scores should be.

Of course, this strategy does have its problems, problems that arise independently of the motivation for choosing it. The chase after more complex models can be a long one. In fact, the most complex "model" of the data is the data set itself. Few would accept this as a *useful* model, but it does give one perspective to one of the consequences of an unfettered pursuit of a closer and closer fit. Essentially, there needs to be an end-point to this strategy, otherwise known as the "stopping rule" problem (Gershenfield, 1998).

Strategy Two: Develop items that satisfy your standards for measurement

In the second strategy identified in this paper, a measurement model is seen as a statistical model that is used within a particular measurement framework. That framework determines the role of the measurement model, and will also determine, to a certain extent, its characteristics. There is no universally accepted framework for measurement (indeed some might doubt that such a framework is even necessary). This will be dealt

with by positing a specific framework that has the virtue of being reasonably straightforward to describe, and also capable of serving a wide range of applications.

The framework is based on that described by Wilson (Wilson, 2004; Wilson, in press). It consists of four building blocks that define the act of measuring: (i) a Construct Map; (ii) an Items Design; (iii) an Outcome Space; and (iv) a Measurement Model. In order to make the discussion concrete, a brief description of these building blocks for a particular instrument follows.

The Four Building Blocks

An instrument is always something secondary: First there is an idea or a concept that is the theoretical object of our interest in the respondent—this will be called the "construct". A construct could be a part of a theoretical model of a person's cognition, such as their understanding of a certain set of concepts, or their attitude toward something, or it could be some other outcome variable such as "need for achievement" or a personality variable such as a bipolar diagnosis. It could be a health-related construct such as "Quality of Life", it could be from the domain of educational achievement, or it could be a sociological construct such as stratification or migrant assimilation. The type of construct that will be described in this paper is one that is particularly suitable for a visual representation called a *construct map*. Its most important features (Masters, Adams, & Wilson, 1994) are that:

(a) there is a coherent and substantive definition for the content of the construct;

(b) there is an idea that the construct is composed of an underlying continuum—in other words, the respondents are ordered from greater to less—one can make a comparison among the respondents of more, equal, or less;

(c) this continuum can be "mapped out" in terms of the responses to the items (either of individual items or groups of items).

Note that, in this conception, the ordering of the respondents implies also an ordering of the item response features. The content of a construct is usually delimited by the intended use of the instrument. A construct can be most readily mapped where the construct has a single underlying continuum—and that implies that for the intended use of the instrument the measurer wants to array the respondents from high to low, or strong to weak, in some context. Note that this does not imply that this ordering of the respondents is their only relevant feature. Some would see that measurement can *only* be thought of in such a context (Wright, 1977), and there are good reasons for taking such

a position, but the arguments involved are not necessary to the development in this paper. Here the argument is that this is a good basis for instrument construction—the argument in this paper is not carried through to try and show that such an assumption is required.

One example of a construct that may be mapped in this way is the Physical Functioning subscale (PF-10: Raczek et al., 1998) of the SF-36 health survey (Ware & Gandek, 1998). This instrument is used to assess generic health status, and the PF-10 subscale assesses the physical functioning aspect of that.

For the construct called Physical Functioning (PF-10), an initial idea of the construct can be built up by considering how a person might increase in reported levels of physical functioning as they progressed, say, through to recovery from an injury. A sketch of the construct map for this case is shown in Figure 1. Note that this depicts an idea rather than being a technical representation. Indeed, later this idea will be related to a specific technical representation, but for now, just concentrate on the idea. In the center of the figure is a line, which represents the underlying continuum. Up the page represents *increasing propensity* to report higher levels of physical functioning; down the page, the opposite. On the left side of the continuum is a representation of the locations of several persons who are being measured with the PF-10 instrument. Each is seen as having a location on the continuum, and other people may be above, below, or be equal to that person. Just one person is shown here, but many more could be located on the map.

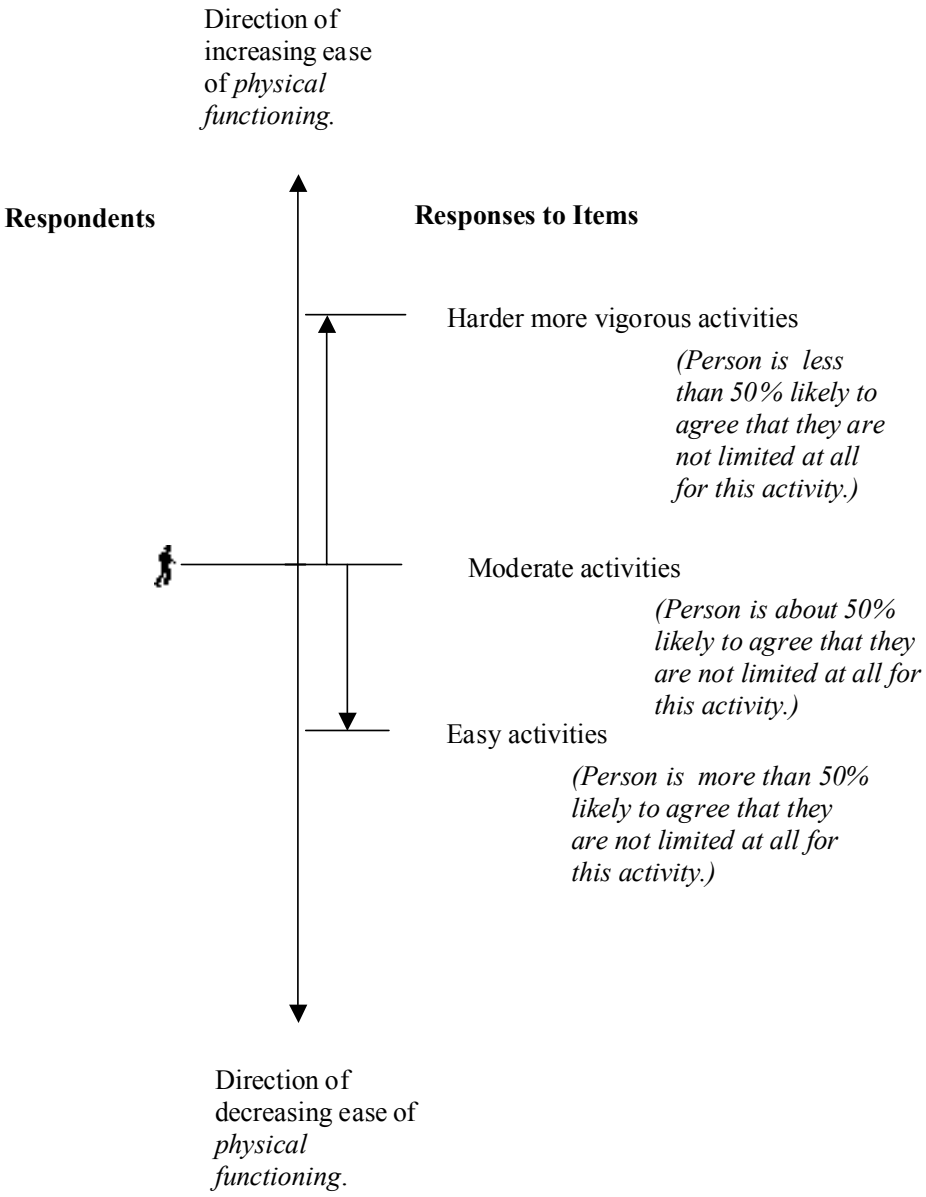

Figure 1. Construct map illustrating the interpretation of different item locations with respect to a single person.

On the right side of the continuum is a description of how these people are tending to respond to the PF-10 items. The description is in terms of short labels for each of the levels of response. Note the sequence of increasing ease of physical functioning is indicated by the order of the generic responses given on the right hand side. This sequence ranges from very much more strenuous activities, such as those represented by the label "Vigorous Activities" (running, lifting heavy objects, etc.), down to activities that take little physical effort for most people, such as bathing and dressing. In between there will be responses that indicate more moderate levels of activity. Respondents at about the same location as a particular activity will tend to respond positively about 50% of the

time when asked about the activity; respondents above the location of an activity will tend to respond positively more often; respondents below that activity will tend to

respond positively less frequently. This can be seen as a probabilistic analogue of a Guttman scale (Guttman, 1944). Thus, the people highest on the PF construct would tend to respond positively to questions that ask about vigorous activities. But for the lowest people on the PF construct, just about the opposite would be true—that person would tend to respond positively only to questions about activities that indicate a very low level of physical functioning. The respondents in between would tend to give answers that were somewhat in between these two extremes. The whole structure is a construct based on the idea of a continuum, and its representation is a construct map. The meaning of these locations on the map will be made quite concrete later.

Next the measurer must think of some way that this theoretical construct could be manifested in a real-world situation. For example, the PF-10 items probably began as informal questions that a health researcher might ask a patient. Typically, there will be more than one real-world manifestation used in the instrument; these parts of the instrument are generically called "items," and the way they are conceptualized will be called the *items design*. Each item will generate a response from the respondent, and this will be classified into a category in the *outcomes space* (Masters & Wilson, 1997). These are the second and third building blocks, respectively. The respondent may produce a "free-response" within a certain mode, such as a short written response, or a performance of some sort. Just as there are many possible ways to create items to match a certain construct, there can be many different outcome spaces that correspond to a given type of item. One common item format is the "forced choice" item, such as the Likert-type item and the multiple choice item from achievement testing. In a forced choice format, the items design and the outcome space are confounded by design. The items design and outcomes space for the PF-10 are illustrated by showing the PF-10 items in Figure 2. This step from block one to block two is a very challenging one, and one where there is little science available to help. Good item developers tend to act more like artists than, say, engineers. And the support provided by measurement is typically in hindsight, rather than at the planning and design stage. An important area for future work is the construction of useful guidelines or templates that item developers can follow to attach items to construct maps.

The final building block in the design of the instrument is the measurement model. The measurement model must provide a principled way to use the information about

respondents and responses that is coded in the outcome space to locate the respondents and the responses on the construct map. Note that in the discussion above, and in what follows, it is assumed that the construct, and hence the instrument, is unidimensional. Discussion of a multidimensional context is beyond the scope of this paper, but the beginnings of such a discussion are available elsewhere (Wilson, in press).

The following items are about activities you might do during a typical day. Does your health now limit you in these activities? If so, how much?

1. Vigorous activities, such as running, lifting heavy objects, participating in strenuous sports (*Vigorous Activities*)
2. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling, or playing golf (*Moderate Activities*)
3. Lifting or carrying groceries (*Lift/Carry*)
4. Climbing several flights of stairs (*Several Stairs*)
5. Climbing one flight of stairs (*One Flight Stairs*)
6. Bending kneeling, or stooping (*Bend/Kneel/Stoop*)
7. Walking more than a mile (*Walk More Mile*)
8. Walking several blocks (*Walk Several Blocks*)
9. Walking one block (*Walk One Block*)
10. Bathing or dressing yourself (*Bathing/Dressing*)

The responses to all questions are:

___Yes, limited a lot; ___Yes, limited a little; ___No, not limited at all.

We will assume that these are dichotomously scored:

__0__Yes, limited a lot; __0__Yes, limited a little; __1__No, not limited at all.

Figure 2. Items designed for the "Physical Functioning" Instrument—PF-10 (note that the phrases in parentheses are labels for the items.).

Consequences for the measurement model

The requirements for a measurement model, the last of these 4 building blocks, are:

1. the measurement model must enable one to interpret the distance between respondent and response on the construct map,

2. the measurement model must enable one to interpret distance between different responses on the construct map, and also the difference between different respondents.

In order to make sense of these requirements, we must say what does "distance" mean? On a geographical map, distance and direction on the map have meaning on the surface of the earth: For example, 1 kilometer north might equal 1 centimeter "up" the map. On a construct map: distance between respondents and responses will indicate the *probability* of making that response. To express this as an equation, assume that the respondent's position is represented by θ and the item response location is represented by δ. Then the probability of the response [Pr(response)] will be given by some function (f) of the difference between the respondent and the response:

$$Pr(\text{response}) = f(\theta - \delta). \tag{1}$$

That can be interpreted thus:

(i) zero distance between a person and a response would mean that that person is likely to endorse statement with a certain probability (say, .50),

(ii) respondent *above* response would indicate a greater probability,

(iii) respondent *below* response would indicate a lesser probability,

and hence we can say that the model must have qualitative features (i) to (iii). However, these qualitative features are not sufficient to preserve the idea of a "map". For that the requirements (ii) and (iii) would need to be more detailed, giving a specific metric form.

Consider now what that means for interpretations of the construct map itself. Figure 1 illustrates the situation for a person in the middle of the construct. For this person, items (i.e., shown on the right hand side) that are at a similar level would be expected to elicit agreement at about a .50 probability, while items that are above, would tend to result in a positive response with a lower probability, and the opposite for those below. We can also consider distances *between* item responses. Figure 3 illustrates the distance between two item responses, considered by people at different points on the scale.

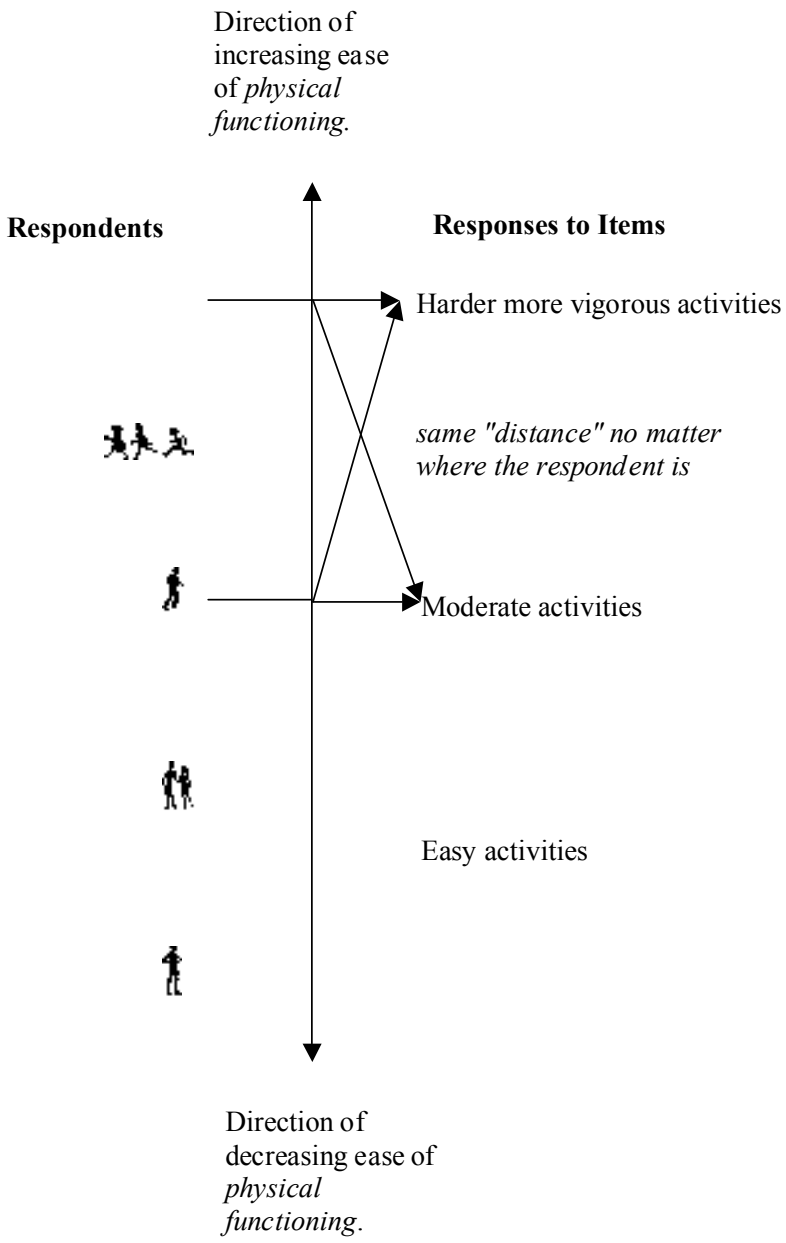

Figure 3. Construct map illustrating the interpretation of different item locations with respect to two persons.

Note, that the distance between "harder and more vigorous activities" and "moderate activities" is the same no matter whether you are looking from the perspective of the people lower on the scale, or those higher on the scale. This is so obvious that it seems odd even to point it out. But it is fundamentally important to the sorts of interpretations one can make of the map: the idea of "location" of an item response with respect to the location of another item response only makes sense if that relative meaning is independent of the location of the respondent involved — i.e., the interpretation of rela-

tive locations needs to be uniform no matter where the respondent is. Another way to put this is that meaning is the same *no matter where you are on the map* (Wright & Stone, 1979; Andrich, 1988). This *invariance* requirement corresponds to the idea that a "centimeter represents a kilometer" wherever you are on a geographical map.

One specific qualitative consequence of this is that the order (on the map) of the item responses must remain the same for *all* respondents, and that the order of the respondents (on the map) must remain the same for *all* item responses. This is equivalent to the requirement of *double monotonicity*, a concept from the topic of nonparametric item response modeling (Mokken, 1997). But recall requirement (2) above—it is stronger, not just order is preserved, but metric properties too. To achieve this in the framework of item response modeling, the item model must have the property that the shape of the item characteristic curve (ICC) is the *same* for all items, such as in Figure 4. The ICC relates the probability of a person's response to the location on the scale (θ). Note that this graph is presented in a way that is rotated 90 degrees from the standard—this is done to make it match the construct map where the θ goes up the page. In this Figure, the items will always have the same order in terms of probability no matter where the respondent is. For example, at θ=0, the lowest respondent location on the map, the items are ordered in difficulty (from highest to lowest) as item 1, item 2 and item 3, respectively. Passing up to θ=1.0, and θ=2.2, one can see that the item difficulty order is identical. At the extremes, these differences will become very small, eventually smaller than any given limit of accuracy—nevertheless, in a mathematical sense, they will always be ordered at any finite point on the scale. One can see that, for the relative item difficulty order to change, it would be required that these ICCs have different shapes.

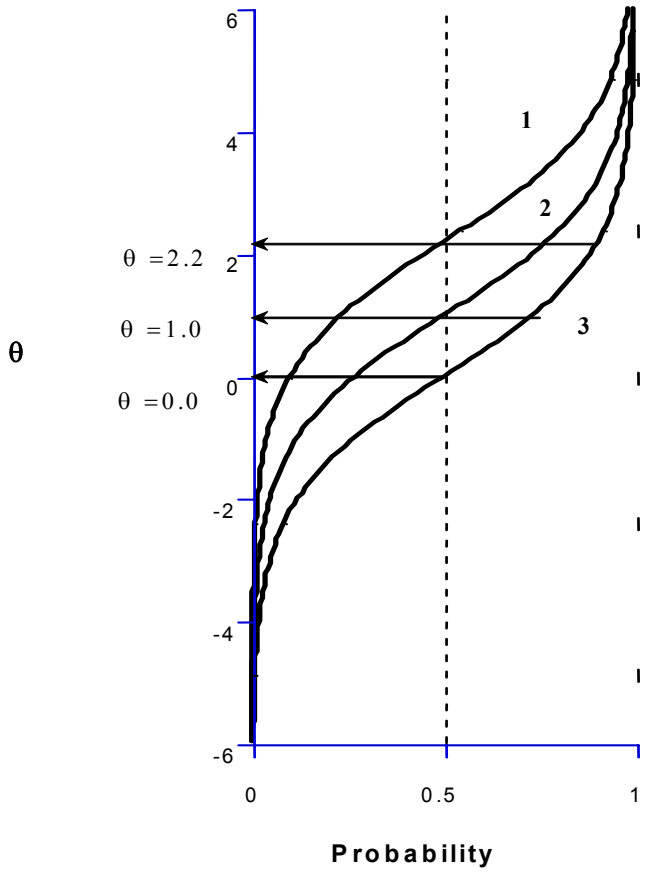

Figure 4. Item characteristic curves—all ICCs have the same shape.

Let us make this choice concrete by moving into the context of parametric item response modeling. Introducing the various item response models is beyond the scope of this article, but introductions have been provided elsewhere (e.g., van der Linden & Hambleton, 1996). In the context of item response models, the conceptual basis of the argument as described above is translated into a different format—it becomes a discussion about parameters within the item response models. Suppose now that the ICCs we are discussing are given by a two-parameter logistic function as in equation (2)

$$P(X_i = 1 \mid \theta, \delta_i, \alpha_i) = \frac{e^{\alpha_i(\theta - \delta_i)}}{1 + e^{\alpha_i(\theta - \delta_i)}} \tag{2}$$

where the θ and δ_i are as above, and the α_i is a slope parameter. In this context, having the same shape for all the ICCs implies that the α_i are all equal, that is, that we are requiring a Rasch model:

$$P(X_i = 1 \mid \theta, \delta_i) = \frac{e^{\theta - \delta_i}}{1 + e^{\theta - \delta_i}} \tag{3}$$

If the α_i are not equal, then the item responses will (at least somewhere) swap their order. This is illustrated in Figure 5. In this Figure, consider three persons at different locations on the scale: At -3.0, the order of the items is (1, 3, 2) (from hardest to easiest); at 0, the order of the items is (1, 2, 3); at 4.0, the order of the items is (2, 1, 3). Note that the argument below, and the following discussion, do not require that the function be the logistic one shown in equations 2 and 3 — any suitably shaped function, such as a normal cumulative distribution function, could equally well be used in the argument. Thus, the argument is quite general, but I shall keep it in the logistic family of functions, as that is what is most commonly used in item response modeling. The consequences of such differences in the shapes of the ICCs is that no equivalent of Figure 5 can be drawn without a very complicated set of rules for interpreting the "locations." To see this, note that for a person at 0, the item order was (1, 2, 3), but as we move to -3.0, the order has changed to (1, 3, 2)—so that items 2 and 3 have switched their order as the person locations have changed. Thus, the relationship between the items is no longer invariant for people at different locations. For a geographical map, this would be like saying that when you move from, say Rome to Naples, the latitudes of other places appear to change—Corsica moves south of Sardinia, say. This would make it impossible to construct a geographical map as we understand it—effectively, "locations" no longer are consistently meaningful. In the same way, the idea of a construct map would no longer be tenable—each possible person location would possibly give a different ordering of the items, and hence the interpretation outlined above would not be useful.

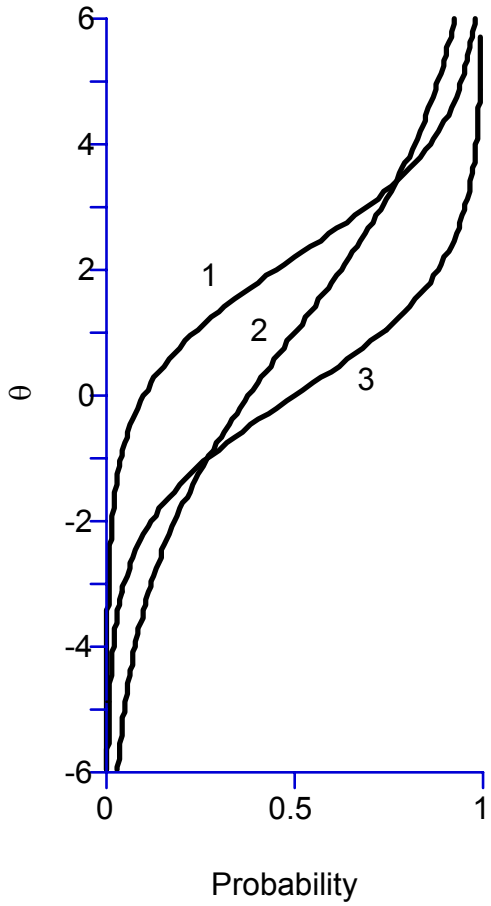

Figure 5. Item characteristic curves—ICCs have different shapes.

Discussion

If an item set really did have characteristics like those in Figure 5, what could one do to overcome the problem of this lack of invariance? Some possible strategies that one could consider are given below.

(a) One could develop a more complex construct interpretation where the underlying theory is consistent with (in fact predicts that) the order of the items should change for people at different points of the construct. If one could not find a set of items that had the construct map property, then this would be the clear next choice. But, if one could find an acceptable set of items with the construct map property, then it would seem to be odd to not use them, and thus preserve the ease of interpretation. This is a rather uncommon approach, mainly because it involves making interpretations that are quite complex, more complex than are usually entertained in psychometric contexts. One example is Yen's interpretation of differences in slopes as evidence of "increasing cognitive complexity" (Yen, 1985). Another is Wilson's extension of Rasch modeling to incorpo-

rate elements of Piagetian stages (the "saltus" model: Wilson, 1989; Wilson & Draney, 1997). In order to make the 2PL model admissible, the invariance principle behind the construct map would need to be modified into something more complex. For example, the requirement that the distance between any two persons looks the same for all items could be replaced by a more complex requirement that the ratio of these distances be a constant for any pair of items (but different across different pairs). This is a requirement that the 2PL model would satisfy. It is not clear how this property would lead to useful interpretations, but indeed it is an invariance property, and thus could be helpful. This approach could be called "make your construct more complex".

(b) One could pretend this issue does not matter and find ways to display something that looks the same. An example of this is given by the usage of the "Response Probability 80" (RP-80) convention (Kolstad et al., 1998). Under this convention, one uses the locations at which a person has a probability of responding of .80. This approach simply ignores the inconsistency with the ordering at other criterion levels of probability. This could be called "hiding behind the statistical complexity".

(c) One could ignore the issue of substantive interpretability of the item parameters. That is, the item parameters and their attendant probability interpretations are ignored in terms of internal construct validity—they are simply allowed to be whatever they turn out to be. This is inconsistent with prominent measurement theories, and also with current testing standards (AERA/APA/NCME, 1999), yet is probably the most common occurrence. This could be called "let the data tell you what the weights should be".

(d) Lastly, one could work to improve the item set so that the ICCs do not cross (e.g., the Rasch model does hold). This can be seen to make sense in terms of the relative rarity of items and constructs, after all, there are many more items than sound constructs. This could be called "make items to fit the construct".

It should be clear that, *given* that one is working in a "construct map" framework, (d) is the only viable alternative. It may be the case that a measurement model is being used without such a framework. Perhaps there is no framework, and one is simply trying to scale all of a certain set of items together, or perhaps one has a different framework. Then the arguments above not necessarily hold. In case (d) above, one needs to take care that indeed the items do represent a reasonable approximation to the ideal where all have the same shape. Fortunately, item fit statistics have been developed that are directly sensitive to the issue of "equal slopes". There are global tests (Andersen,

1973) as well as tests that are diagnostic for each item (Wright & Masters, 1982; Wright & Stone, 1979).

Thus, the consequences for the measurement model are that the estimation of parameters of best-fit needs to be done under the set of constraints implicit in equation (3), or an equivalent. It doesn't matter that you might get better fit by estimating separate $\alpha_i s$, those models would not be *admissible* to as part of the family of models that affords a construct map interpretation. This sort of position is not at all inconsistent with good statistical modeling practice. Consider the following quotation from Annette Dobson's monograph on statistical modeling:

> "A criterion for a good model is one which 'explains' a large proportion of this variability... In practice this has to be balanced against other criteria such as simplicity. Occam's Razor suggests that a parsimonious model which describes the data adequately may be preferable to a complicated one which leaves little of the variability 'unexplained'." (Dobson, 1983, p. 8)

In fact, the first strategy described by Dobson is really only a partial strategy—one *always* must consider the need for parsimony, as well as the need to base a statistical model on the interpretational context, which will determine that some models are admissible, and others are not.

Of course, in carrying out strategy (d) above, it is very important that the constraint of equal α_i be reasonably well reflected in the data. That is why we need to test that particular fit aspect: using techniques that investigate item fit such as the weighted and unweighted fit indices (Wright & Masters, 1982), and fitting all-but-one α_i constrained models. These techniques pinpoint problem items, and can be helpful in diagnosing the causes of these problems, thus making it easier to come up with replacement items that do not have the same problems. One might argue that this results in a relatively low quality final item set, as it tends to lead one to seek alternatives to items with the highest slope parameters, which are commonly interpreted as the "best". But in fact this is simply a poor strategy—a better strategy is to seek alternatives to items with shallow slopes—this alternative strategy can keep the items with the steepest slope in the item set. Some would argue that *any* usage of the fit statistics to select the item set is a threat to validity. This is in direct contradiction with the standard procedures for developing instruments (AERA/APA/NCME, 1999), which insist that item analysis be used for deleting problem items. In the situation being described here, where the interpretability of the instrument is seen as being dependent on the integrity of the construct

map, it makes every sense to treat items that threaten that integrity as problematic—items with extremely different slopes will threaten the validity of the interpretation, and hence are problem items.

In review, the old question: "Should you 'make the model fit the data' or 'the data fit the model'?", is recast as a question of what are the *a priori* interpretational constraints that must be imposed on the model to have it make sense in the way that you want it to. This has been an issue that has been debated in the mainstream statistics literature. For example, consider what Gershenfeld had to say:

> "...the two central tasks are always choosing the functional form of the model and using the data to determine the adjustable parameters of the model." (Gershenfield, 1998, p. 113)

What we have been concentrating on in this paper is the former—the functional form of the model. Gershenfeld continues:

> "Decreasing one kind of error is likely to increase the other kind. This is called a *bias/variance tradeoff*—if you want less bias in the estimate of a model parameter, it usually costs you some variance. A more flexible model that can better represent the data may also more easily led astray by noise in the data." (Gershenfield, 1998, p. 113)

Thus the question becomes, should one make the model with constrained scores fit the data, or should one make the model with unconstrained scores fit the data? In other words: Should you make the items fit the construct or the construct fit the items.

Following the discussion above, if one wants to preserve the meaningfulness of the distance in a construct map, then you have no option but to seek item sets that fit the constrained model. In practice, within the item response modeling area, this becomes a strategy of finding ways to use the Rasch family of models to deal with the many complexities of measurement situations. In theory, non-parametric models may also be used, but there are limits to what you can achieve with those models.

Further directions that this line of research can take include the following. First, the extension of this argument to polytomous items is fairly straightforward, but, as is the case with respect to other aspects, the extension is not without its complexities. Generally, polytomous models that fall within the Rasch family, such as the partial credit model and the rating scale model (Wright & Masters, 1982) do indeed conform with the construct map logic displayed here, while models that involve a slope (α) parameter in

some form, such as the graded response model (Samejima, 1969) and the generalized partial credit model (Muraki, 1992) do not. This will be examined in a companion paper devoted to just this topic (Wilson, 2001). There are a number of ancillary issues that arise in the polytomous case, such as what to do about items with differing numbers of categories. Again, these issues are beyond the scope of the current paper. A second set of issues that deserves a similar analysis is the use of a linear model on the item parameters (Fischer, 1973)—also called "facets" models (Linacre, 1989)—the question is whether a similar invariance is needed across these facets. Third, there are some cases where a strong model exists that *determines* the scores. For example, Draney, Pirolli, and Wilson (1995) showed that a power law of learning corresponds to a logistic model of errors that looks like:

$$Pr(X_{jkt}=1 \mid \theta_i,\delta_j,\tau_k,\gamma)=\frac{\exp(\theta_i+\delta_j+\tau_k-\gamma\log(t))}{1+\exp(\theta_i+\delta_j+\tau_k-\gamma\log(t))} \tag{5}$$

where t is the number of attempts. Thus, the score for the γ (rate of learning) facet is $\log(t)$. Clearly, this model is within the Rasch family, yet it involves effects that are not easily related to the arguments above.

References

American Educational Research Association, American Psychological Association, & National Council for Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123-140.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Cella, D., & Chang, C-H. (2000). A discussion of item response theory and its application in health status assessment. *Medical Care, 38* (supplement II), II-66-II-72.

Dobson, A. J. (1983). *An introduction to statistical modeling*. London: Chapman and Hall.

Draney, K. L., Pirolli, P., & Wilson, M. (1993). A measurement model for a complex cognitive skill. In P. Nichols, S. Chipman, & R. Brennan, (Eds.) *Cognitively diagnostic assessment* (pp. 103-126). Hillsdale, NJ: Erlbaum.

Dwyer, J. H. (1983). *Statistical models for the social and behavioral sciences*. New York: Oxford University Press.

Fischer, G. H. (1973). The linear logistic model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G. H., & Molenaar, L. W. (1995). *Rasch models: Recent developments and applications*. New York: Springer-Verlag.

Fisher, A. (1994). Development of a functional assessment that adjusts ability measures for task simplicity and rater leniency. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice (Volume II)* (pp. 145-175). Norwood, NJ: Ablex.

Guttman, L. A (1944). A basis for scaling quantitative data. *American Sociological Review, 9*, 139-150.

Gershenfeld, N. (1998). *The nature of mathematical modeling*. Cambridge, UK: Cambridge University Press.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care, 38* (supplement II), II-60-II-65.

Hayes, R. D., Morales, L. S., & Reise, S. P. (2000). Item response theory and health outcomes measurement in the 21st century. *Medical Care, 38* (supplement II), II-28-II-42.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., deFur, E., & Angeles, J. (1998). *Should NCES adopt a standard? The response probability convention used in reporting data from IRT assessment scales*. Washington, D.C.: American Institutes for Research.

Linacre, J. M. (1989). *Many faceted Rasch measurement*. Unpublished doctoral dissertation, University of Chicago.

Ludlow, L., & Haley, S. (1992). Polytomous Rasch models for behavioral assessment: The Tufts assessment of motor performance. In M. Wilson (Ed.), *Objective Measurement: Theory into Practice* (pp. 121-137). Norwood, NJ: Ablex.

Masters, G. N., Adams, R. J., & Wilson, M. (1994). Charting student progress. In T. Husen, & T. N. Postlethwaite (Eds.), *International Encyclopedia of Education (2nd Edition)* (pp. 5783-91). Oxford: Pergamon Press.

Masters, G. N., & Wilson, M. (1997). *Developmental assessment.* Research Report. University of California, Berkeley.

McHorney, C. A., & Cohen, A .S. (2000). Equating health status measures with item response theory: Illustrations with functional status items. *Medical Care, 38* (supplement II), II-43-II-59.

Mokken, R. J. (1997). Nonparametric models for dichotomous responses. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 351-368). New York: Springer-Verlag.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16* (2), 159-176.

Raczek, A. E., Ware, J. E., Bjorner, J. B., Gandek, B., Haley, S. M., Aaronson, N. K., Apolone, G., Bech, P., Brazier, J. E., Bullinger, M., & Sullivan, M. (1998). Comparison of Rasch and summated rating scales constructed from the SF-36 Physical Functioning items in seven countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology, 51*, 1203-1211.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded responses. *Psychometrika Monograph*, *17*, 1-100.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In Thissen, D. & Wainer, H. (Eds.), *Test Scoring.* Mahwah, NJ: LEA.

Van der Linden, W., & Hambleton, R. K. (1996). *Handbook of modern item response theory.* New York, Springer.

Ware, J. E., & Gandek, B. (1998). Overview of the SF-36 Health Survey and the International Quality of Life Assessement (IQOLA) Project. *Journal of Clinical Epidemiology, 51*, 903-912.

Ware, J. E., Bjorner, J. B., & Kosinski, K. (2000). Practical implications of item response theory and computerized adaptive testing: A brief summary of ongoing studies of widely used headache impact scales. *Medical Care, 38* (supplement II), II-73-II-82.

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin*, *105*, 276-289.

Wilson, M. (2001). *On choosing a model for measuring.* Paper presented at the 3rd International Conference on Objective Measurement, Chicago.

Wilson, M. (2004). *Constructing Measures: An Item Response Modeling Approach.* Mahwah, NJ: Erlbaum.

Wilson, M. (in press). Subscales and summary scales: Issues in health-related outcomes. In J. Lipscomb, C. Gotay, & C. Snyder (Eds.), *Outcomes Assessment in Cancer.* Cambridge, UK: Cambridge University Press.

Wilson, M., & Draney, K. (1997). Partial credit in a developmental context: The case for adopting a mixture model approach. In M. Wilson, G. Engelhard, & K. Draney,

(Eds.), *Objective Measurement: Theory into Practice (Volume IV)* (pp. 333-350). Norwood, NJ: Ablex.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Education Measurement, 14*, 97-116.

Wright, B. D., & Stone, M. (1979). *Best Test Design*. Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Yen, W. M. (1985). Increasing item complexity: A possible cause of scale shrinkage for unidimensional Item Response Theory. *Psychometrika*, *50*, 399-410.