Issues not to be overlooked in the dominance versus ideal point controversy

Anna Brown
SHL Group, UK

Alberto Maydeu-Olivares
Faculty of Psychology
University of Barcelona, Spain

Commentary article on Drasgow et al, “75 Years After Likert: Thurstone Was Right!”

Industrial and Organizational Psychology: Perspectives on Science and Practice.

Address correspondence to: Anna Brown, Research, SHL UK, The Pavilion, 1 Atwell Place, Thames Ditton, Surrey, KT7 0NE, United Kingdom. E-mail: anna.brown@shlgroup.com

Issues not to be overlooked in the dominance versus ideal point controversy

Drasgow, Chernyshenko and Stark (2010) argue that "responses to questions requiring introspection involve a comparison process" and we agree with this statement. The key question is, however, what it is respondents compare themselves to? Drasgow and colleagues suggest that respondents *always* compare their self-perception to the statement's location (an **ideal point**) and therefore ideal point models should always be used. We feel the answer depends on the construct being measured and on the type of items used to measure it. Sometimes the response mechanism appears to be an ideal point process. But many other times we believe that given a statement and a binary (endorse/not endorse) rating scale, an individual will endorse the item if its utility is larger than a certain threshold, and will not endorse it otherwise. In this case respondents compare themselves to a **threshold**, which calls for the use of a dominance model.

In order to investigate whether a threshold or an ideal point mechanism best describes the response process, intermediate items should be used more frequently. This is because, as Drasgow et al. point out, if only extreme items are used, dominance and ideal point models will yield a similar fit. It is only considering intermediate items like "My life has had about equal amounts of ups and downs" we can tell apart an ideal point and a threshold process. However, classical statistics (i.e. low item-total correlations) is not the only reason why there are very few intermediate items in applications. There are other important issues with ideal point items and models that are overlooked in Drasgow and colleagues' account.

Good intermediate items are difficult to write

Good item writing practice should be maintained, regardless of whether an item is extreme or intermediate, and regardless of whether it reflects a dominance or an ideal point process. Items should be fair and should not confuse respondents. However, following this basic practice makes it difficult to write good intermediate items. To illustrate, consider descriptions of typical behaviors of an average scorer on a personality scale, as often given in questionnaire manuals, or in feedback reports. These descriptions are very fuzzy, with suggestions that some behaviors might apply in certain conditions, or in certain times, or that behaviors might apply to a lesser extent. This ambiguity very much applies to items with locations in the middle of the trait continuum.

The first problem occurs when some conditional clause is used to describe an average scorer, which leads to double-barreled items. Consider the item “Although I have a daily organizer, I have a hard time keeping it up to date”. Now consider a respondent trying to make sense of this item and decide how to respond. The typical feedback is that such items are incredibly frustrating. To some, the first part does not apply and it makes the whole item meaningless, thus considerably increasing the likelihood of responding randomly. To others, the first part applies but the second part does not apply, and they feel equally confused. We believe that regardless of the underlying response model, most of these items are to be thrown away. Not because of low item-total correlations, since we advocate the use of IRT modeling instead, but because we are concerned about the face validity of such items, and potentially about their construct validity too.

The second problem occurs when items explicitly require a judgment about how one compares with a reference group. An example of such an item is “My room neatness is about average”. This item requires making a judgment about the state of other people’s rooms. Such judgments are confounded by the respondent’s frame of reference, and tend to show item bias. This situation can occur also with extreme items such as “I am more outgoing than most people”. Such items tend to cause problems when people with very different frames of reference are compared, for instance in cross-cultural research. Consequently, outgoing Spaniards might score lower on Sociability than, let’s say, Finns.

The third problem occurs when an attempt to write an intermediate item leads to introducing particulars and contexts not clearly related to the construct of interest. Consider the item “I enjoy chatting quietly with a friend at a café” (as opposed to simply “I enjoy chatting”). This item is intended to measure extraversion, however, it is likely that the specificity of the context will introduce multidimensionality, and the item will have a low discrimination on the extraversion scale.

Developers of personality scales know that it is easy to write descriptions of high and low scorers on a personality trait. Because dominance items describe behaviors typical of the high or low scorers on the trait, they are also easy to write. They are easy to answer, too. From a respondent’s perspective, responding to the item “I am organized” is straightforward. This item is unambiguous and is easily related to one’s behavior. The fact that it is easier to write good dominance items than good ideal point intermediate items is an overlooked issue in Drasgow et al.'s account.

Estimation may not be so accurate for ideal point models

An IRT model is useless unless it can be shown that its item characteristic curves (ICC) can be estimated with enough precision in reasonably small samples and that latent trait estimation is precise enough. Having been used more widely to date, more software is capable of fitting dominance models, and requirements for successful item and latent trait estimation are better known than for ideal point models. Software considerations aside, it may well be that it is inherently more difficult to recover true item parameters for ideal point models than for dominance models. Maydeu-Olivares, Hernández and McDonald (2006) proposed the ideal point counterpart of the normal ogive model, the normal Probability Density Function (PDF) model. Yet, in simulation studies it was not possible to obtain nearly as accurate estimation of item parameters as with the normal ogive model (e.g. Forero & Maydeu-Olivares, 2009). Clearly, more research involving simulation studies is needed with ideal point models to show that item parameters, as well as respondent's traits, are estimated as well as they are with dominance models.

Ideal point models are not invariant to reverse scoring

Within an IRT framework, it is not necessary to reverse score items, regardless of whether a dominance model or an ideal point model is used. However, some pause is needed when choosing the direction of the construct using ideal point models. In these models, reverse scoring of the items leads to a different set of parameters and to a different goodness of fit. See Maydeu-Olivares, Hernández and McDonald (2006, p. 467) for an example involving modeling dissatisfaction with life as opposed to satisfaction with life. Dominance models, on the other hand, are invariant to reverse coding and allow equivalent modeling of either end of psychological constructs. That is, when dominance models are used and some or all items are reverse scored, parameters are transformed in a simple way and model fit remains unchanged. This is not the case with ideal point models.

IRT modeling of forced choice does not require ideal point items nor models

Some researchers mistakenly believe that forced-choice measurement using IRT requires ideal point items and models. The confusion probably stems from the fact that the same group of researchers happened to extensively use ideal point models and items (Stark, Chernyshenko, Drasgow & Williams, 2006; Chernyshenko, Stark, Drasgow & Roberts, 2007), and have introduced an IRT approach to creating forced-choice tests that involves an ideal point model (e.g. Stark, Chernyshenko & Drasgow, 2005).

Modeling forced-choice responses requires a model for the response mechanism by which respondents choose between two items. The model for this response mechanism is independent of the model used to link an item to a personality trait. The latter can be a dominance model or an ideal point model.

Stark, Chernyshenko and Drasgow's (2005) use the MUPPM model for forced choice items, in which the probability of preferring one item to another is approximated by the joint probability of endorsing one item and rejecting the other. To link the probability of endorsing the item to a personality trait, they use an ideal point model. However, there is nothing in the actual MUPPM model that stops it from being populated with dominance items and, consequently, using a dominance model.

The MUPPM is not the only model using an IRT approach to modeling forced-choice responses. Our own model (the Thurstonian IRT model, see Brown & Maydeu-Olivares, in press) uses Thurstone’s (1927) Law of Comparative Judgment to model how respondents choose between two items. The model posits that respondents choose the item with the largest utility at the time of comparison. How the utilities of individual items depend on the psychological constructs being measured is immaterial for the comparative law. An assertion by Drasgow and colleagues that using a dominance model to describe forced-choice judgments can lead to serious model misspecification (see footnote) is misleading. We happen to use dominance items and a dominance model, but ideal point items and model could be used just as well. The problem, as Drasgow and colleagues rightly point out, is that when scored traditionally these instruments produce ipsative data. However, it is not the use of dominance items that produces ipsative data; it is the classical way of scoring forced-choice instruments. The Thurstonian IRT model we proposed completely overcomes the problems of ipsative data in questionnaires with dominance items and ICCs, and latent traits are very accurately recovered (see Brown & Maydeu-Olivares, in press).

Furthermore, Tsai and Böckenholt (2001) have shown that ideal point and dominance models are undistinguishable from comparative data. That is, from preference choices made between two items it is impossible to determine if the items were linked to personality traits using an ideal point model or a dominance model. Hence, it does not matter if ideal point or dominance items are used in forced-choice formats. Clearly, much additional work is needed to evaluate implications of different item response models on comparative judgments.

Dominance items measuring multiple traits can be mistaken for ideal point items

In applications, even the best constructed unidimensional measures contain sizeable amounts of multidimensionality. Drasgow and colleagues report that their interest in ideal point models arose when fitting unidimensional dominance models to personality scales and finding that dominance models could not reproduce personality data well. Searching for better fitting models, they applied a non-parametric unidimensional model (Levine, 1984), which yielded shapes that suggested the use of ideal point models.

Unfortunately, their findings should not be taken to imply that an ideal point model is the 'true generating' model for the data. Rather, if the true model is a multidimensional dominance model, fitting Levine's model would reveal bumps and swirls similar to those for a unidimensional ideal point model (Levine, 1994; Maydeu-Olivares, 2005). While Drasgow and colleagues argue that some debates about personality dimensions “might be clouded by the application of misspecified models” (implying that multidimensionality might be wrongly suspected where a unidimensional ideal point process is present), it is equally probable that the opposite can occur (that a unidimensional ideal point process is wrongly suspected where a multidimensional dominance model is present) –see Peress and Spirling (2009). The presence of multidimensionality, however, might not be picked up if the construct is tested in isolation from other constructs, and no other items share variance specific to this other dimension. For instance, it might be impossible to tell from the data if the item “I enjoy chatting quietly with a friend at a café” is an ideal point item on the one-dimensional extraversion scale, or a dominance item measuring introversion and warmth. This admittedly messy situation suggests that one cannot differentiate between the two models on the basis of model fit alone, and we agree with Drasgow and colleagues that a good theory should guide judgments made about appropriateness of any model.

Conclusions

It is one thing to find that **some** items are better described (importantly, not only mathematically but also conceptually) by the ideal point response process. In this case it is indeed a welcome development to model such items appropriately. However, it is an entirely different thing to contend that applied researchers and practitioners should write **all** their items to fit an ideal point response model. We feel that there are clear advantages in using dominance items (and models) that should not be overlooked.

Our own approach regarding the choice of dominance vs. ideal point models is simple. We inspect the item stems and classify items as ideal point or dominance items. We then fit a dominance model if all items are dominance items and an ideal point model if all items are ideal point items. We never limit ourselves to unidimensional models. Rather, we always fit multidimensional models as well. In the case of items of mixed type, we fit first an ideal point model. This is because one can argue that many dominance items are simply a special case of ideal point items where the ideal point is located outside the region of high density of respondents. However, given the issues discussed above, we also fit a dominance model of the same class for comparison. Clearly, most items in use today are dominance items. Thus, we end up fitting more dominance models than ideal point models. Even when ideal point items are present, sometimes we end up choosing a dominance model for one or more of the reasons listed above.

Should we use ideal point models for all items? Our position is a clear “No”, and we have offered a number of arguments to support it. Should we discard existing dominance items and replace them by ideal point items? If dominance items work well and provide good measurement quality and validity, as they do in personality measures, our answer is

“Certainly not”. Should we develop more intermediate (including ideal point) items? “Of course”. Only by examining more intermediate items, in more situations, can we obtain clear answers about their properties, advantages and disadvantages, and resolve the ideal vs. dominance models controversy.

References

Brown, A. & Maydeu-Olivares, A. (in press). Item response modeling of forced-choice questionnaires. *Educational and Psychological Measurement*.

Chernyshenko, O., Stark, S., Drasgow, F. & Roberts, B. (2007). Constructing personality scales under the assumption of an ideal point response process: toward increasing the flexibility of personality measures. *Psychological Assessment*, *19*, 88-106.

Drasgow, F., Chernyshenko, O. S. & Stark, S. (2010). 75 Years After Likert: Thurstone Was Right! *Industrial and Organizational Psychology: Perspectives on Science and Practice, 3*.

Forero, C.G. & Maydeu-Olivares, A. (2009). Estimation of IRT graded models for rating data: Limited vs. full information methods. *Psychological Methods*, *14*, 275-299.

Levine, M.V. (1994). *Every data set that is well fit by a two-dimensional IRT model can be equally well fit by a lower dimensional model*. Paper presented at the 59th Annual Meeting of the Psychometric Society. Champaign, IL.

Maydeu-Olivares, A. (2005). Further empirical results on parametric vs. non-parametric IRT modeling of Likert-type personality data. *Multivariate Behavioral Research*, *40*, 275-293.

Maydeu-Olivares, A., Hernández, A. & McDonald, R.P. (2006). A multidimensional ideal point IRT model for binary data. *Multivariate Behavioral Research*, *44*, 445-472.

Peress, M. & Spirling, A. (2009). Scaling the critics: Uncovering the latent dimensions of movie criticism with an item response approach. *Journal of the American Statistical Association*, *105*, 71-83.

Stark, S., Chernyshenko, O. & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The

Multi-Unidimensional Pairwise-Preference Model. *Applied Psychological Measurement*, *29,* 184-203.

Stark, S., Chernyshenko, O., Drasgow, F. & Williams, B. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology*, *91*, 25-39.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Tsai, R.C. & Böckenholt, U. (2001). Maximum likelihood estimation of factor and ideal point models for paired comparison data. *Journal of Mathematical Psychology*, *45*, 795-811.