DIGITALCOMMONS @WAYNESTATE

Journal of Transportation Management

Volume 29 | Issue 1 Article 5

7-1-2018

About item response theory models and how they work

Nell Sedransk
National Institute of Statistical Sciences, nsedransk@niss.org

Recommended Citation

Sedransk, Nell. (2018). About item response theory models and how they work. Journal of Transportation Management, 29(1), 35-44. doi: 10.22237/jotm/1530446640

ABOUT ITEM RESPONSE THEORY MODELS AND HOW THEY WORK

Nell Sedransk
National Institute of Statistical Sciences*

ABSTRACT

This article is about FMCSA data and its analysis. The article responds to the two-part question: How does an Item Response Theory (IRT) model work differently . . . or better than any other model? The response to the first part is a careful, completely non-technical exposition of the fundamentals for IRT models. It differentiates IRT models from other models by providing the rationale underlying IRT modeling and by using graphs to illustrate two key properties for data items. The response to the second part of the question about superiority of an IRT model is, "it depends." For FMCSA data, serious challenges arise from complexity of the data and from heterogeneity of the carrier industry. Questions are posed that will need to be addressed to determine the success of the actual model developed and of the scoring system.

INTRODUCTION**

This article is about FMCSA data and its analysis. The essential question posed to this author was, in the context of FCMSA data analysis, was: - How does an Item Response Theory (IRT) model work differently to make it better than current FMCSA practice or better than any other model? The quick answer is that IRT models are a class of data-based models that are different from other kinds of models because IRT models establish their relevance and validity differently from other kinds of models or scoring systems. From a practical point of view, IRT models focus on items and assign a weight to each one in accord with the acuteness of each item's ability to distinguish between lower- and higher-scoring (safer and less safe) individuals (carriers).

Whether an IRT model performs better or worse than another model depends on whether the assumptions required for an IRT model are met sufficiently well and also depends on key technical decisions that define the specific IRT model being developed.

An IRT model is no different from any other data-based model in three important ways:

- A data-based model can detect a pattern in the data and give a mathematical or numerical definition for this pattern that can be used to estimate or to predict.
- What can be modeled is determined – and limited - by the information present in the data (unless external or theoretical components are imposed on the otherwise data-based model).
- A data-based model cannot determine the veracity of any datum, whether aberrant or consistent with the pattern.

A longer answer requires first understanding the conceptual basis for IRT models. Then the method for constructing the model and computing a score is illustrated. Finally, attention can turn to the particular challenges for FMCSA data and to the elements that determine how well the model can perform: 1) IRT model requirements: the premises built into the structure of an IRT model, 2) Data used to fit the model: data selected, also both properties and the form of data input, 3) Specifications for the particular model; structure, model precision and accuracy, minimum information required for reliable scoring, and ultimately, 4) Model-based scoring or decision-making: the implementation and reporting of the model and individual scores.

Before going further note that in 4) above, *how* a model or a score is reported or is used depends on administrative decision-making and is not intrinsic to the model or scoring system itself. The kind of model for FMCSA data that is discussed in the NAS report is complex and belongs to the class of

confirmatory MIRTs – Multi-dimensional Item Response Theory models (van der Linden, W., 2018). "Multi-dimensional" means that several distinct aspects of safety will be addressed. In this case six aspects are drawn from BASIC information (excluding the category "Crash Indicator"). "Confirmatory" means that those six aspects have been pre-determined and that the items that address them have already been categorized accordingly.

Thus, this MIRT can be thought of in two stages: modeling separately for each aspect using the relevant items, then assembling the results for the individual aspects into a single score.[1] The guiding concept is the same at each stage.

HOW AN IRT MODEL WORKS

Fundamentals of (Any) Data-based Model

The most general concept for a data model is a specified computation that combines data for a collection of observations/factors/items/measures into a summary statistic. For a data-based model (these include IRT models), data are also used in setting the specifications for that computation.

Models come in many forms. They can be simple (a mean or a total) or complicated; they can be theory-based or empirical; they can be linear, non-linear or they can have no closed form to write down as an equation. Model computations can be pre-specified, be data-based or they can combine a pre-specified computation with a data-based computation.

Regardless of the particular form, all data-based models take in a collection of observations and generate a summary statistic (whether uni-dimensional or multi-dimensional or complex function). The value of any model is limited first by the scope of the factors included in the data and second by the quality (truth, precision, accuracy and relevance) of the data. Whether in addition the model is "fit for purpose" depends on its relevance and the intended purpose.

Different kinds of models and scores lend themselves naturally to different ways of establishing *relevance* (validity) of the model. Prediction accuracy is one measure of relevance when there is an external measurable quantity for comparison (i.e., the true value or a gold standard). If the true value is measured with error, model adequacy can be formulated in terms of the error component.

In the absence of a gold standard, other kinds of data-based models may utilize other auxiliary measures, recruit independent data or consult an expert resource. In any of these cases, without a gold standard the calculation of relevance is subject to variation depending on the particular selection of independent data or expert opinion.

IRT models differ conceptually (Hambleton, R.K., Swaminathan, and Rogers, H. J., 1991). Essentially an IRT model postulates the existence of such a standard (fundamental factor or trait) that is fixed but that is only observable indirectly. Consequently an individual's true score can only be inferred or predicted based on indirect information. An IRT model optimizes this inference given the available data without recourse to exogenous information.

Concept Underlying IRT Models

Constructing an IRT model of an unobservable fundamental factor depends on having indirect information that can be used to infer/predict that factor's value based on the indirect information about an individual. In this case, the fundamental factor is referred to as "Safety Culture;" fundamental factors at the first stage are the indices for "Unsafe Driving," "Vehicle Maintenance," "Driver Fitness" etc. The indirect information is the data (reported items) that make up the FMCSA data base for each of these first-stage groups. The purpose of the model is to infer/predict each individual's true factor score, first for each of the first stage factors and then overall.

There are three essential components for constructing an IRT model: 1) the postulated numerically scaled fundamental factor, 2) the *difficulty* of each item, 3) the *discrimination* of each item.

The advantage of an IRT model is that *both* attributes – *difficulty* and *discrimination* – are utilized to infer an individual's factor score the scale from "safest" to "least safe."

Item *difficulty* is not enough – suppose a candidate item is: Does the operator's birthday date contain a "5"? In one sense this qualifies as "difficult," meaning that fewer than 10% of operators' birthday dates will be either 5, 15 or 25. But as this conveys no information about an operator's safe/unsafe driving, its inclusion in a model or score can only add noise. To be a useful item, its difficulty (i.e., likelihood of a positive response) must align with the scale of the fundamental factor.

IRT models anchor each item's relevance to the fixed but unobservable fundamental factor with a numerical scale. Hence the importance of *discrimination,* i.e., the capability of each item separately to correctly place an individual (carrier) on the numeric scale based on the individual's response. Since item *discrimination* is the basis for the item weights, the failure of an irrelevant item to *discriminate* along the Safety Culture scale will result in its being given no weight in the score calculation. Other items that fail to discriminate are those where the response is uniform across the safety scale, as for example when the overwhelming number of carriers report that they transport loads of all types.

Conceptually, the relationship among these three elements is depicted by plotting the probability of a "present" response to an item as a function of the true factor score. The graph in figure 1 shows the Item Characteristic Curves for two items, each with an absent/present response where "present" is associated with an unsafe practice. The x-axis gives the true factor score. For each item, the point on the curve (y-coordinate) gives the probability that an individual with factor score given by the x-coordinate will respond "present" to the item.

Item *difficulty* (language borrowed from the education origins of IRT) is defined to be the x-coordinate corresponding to the 50% y-coordinate point of the curve. In this illustration, the curve for item #2 lies toward higher values on the "safer to less safe" scale; therefore item #2 is "more difficult" (i.e., associated with greater safety risk).

Item *discrimination* can also be visualized from the curve graphed in Figure 1. Note that for item #1 the curve is steeper, indicating that there is a narrow range of true scores for which there is a considerable mixture of responses "present" and "absent." Hence, item #2 discriminates better than item #1 where a response of "present" would be consistent with a broader range of true scores (also true for a response of "absent"). Technically the definition of discrimination is the steepness (slope of the curve) at the y=0.50 point of the curve where the difficulty is also defined (by the x-coordinate).

IRT Models and Scores

The method for constructing an IRT model is most easily seen by solving a simpler problem first. Consider how a model and scoring system would be developed IF the gold standard or true scores could be known for a large enough number of individuals to build a good model. Then the solution for this case can guide the development of a model when the gold standard is unknowable.

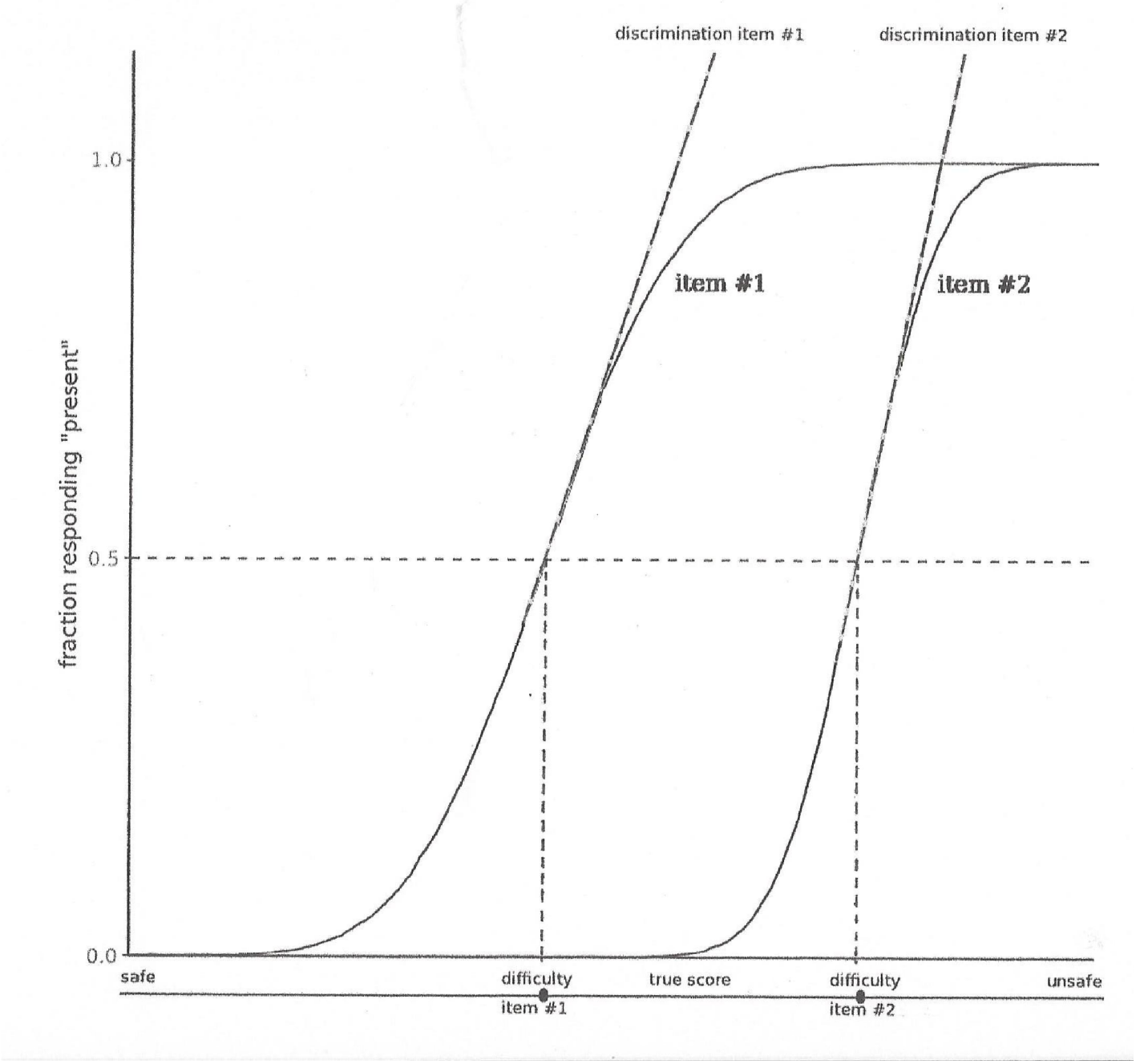
This case is simpler but still needs to create a model to solve two problems – the properties of the individual items and the scoring of the individuals (carriers).

Problem Specification: For just a single aspect/dimension (e.g., "Driving Safety," one aspect in stage one for an MIRT), suppose the gold standard scores are known for the majority of the population of carriers. For the remainder, a model is needed from which to estimate true (gold standard) scores for the remainder and for future carriers.

Assume that there are five items for this aspect and each requires an absent/present response.

Step One: Take each item one at a time and graph its curve (as shown in Figure 1). The item's difficulty and discrimination can be determined from

FIGURE 1:
CONCEPTUAL RELATIONSHIPS AMONG FACTOR SCORE, ITEM DIFFICULTY AND ITEM DISCRIMINATION



this curve. The curve is created as described above: at each point on the true score scale, record the percentage of "present" responses for all individuals with that true score.[2] Difficulty can be measured in a usual way, i.e., find the score associated with responses equally divided between "absent" and "present," i.e., the true score (x-coordinate) for the 50% "present" responses (y-coordinate = 0.50). Discrimination is the slope of the curve at this point.

Step Two: Assemble all the items for this single factor. To see how items compare, the curves for all the items can be plotted together. Items with curves lying toward the left have lower difficulty; curves for

items of greater difficulty (increasingly unsafe practices) are located to the right.

This multi-item graph is also the basis for defining the likelihood for each possible set of responses to the complete set of items (5 items in this illustration). For the true score marked on the graph, the vertical line intersects each of the 5 item curves at the probability of a "present" response (marked on the vertical right-hand axis). So the probability of a "present" response to item #1 is p_1, to item #2 is p_2, etc., and the probability of a "absent" response to item #1 is $(1- p_1)$, to item #2 is $(1- p_2)$, etc. With a crucial assumption that responses to the separate items are independent, the likelihood of every possible combination of responses for an individual with the true score depicted can be calculated. For instance, a response (present, absent, present, absent, absent) would have the likelihood:

$$\text{Probability of } \{1,0,1,0,0\} = p_1 \times (1- p_2) \times p_3 \times (1- p_4) \times (1- p_5) .$$

Step Three: Assign an (inferred) factor score based on an individual's responses when the individual's true score is unknown. First, the probability of that individuals set of responses can be calculated at each point along the true factor scale. Maximum likelihood assignment means choosing the number on the scale with the highest probability.

One important result arises from basing the score on the probability of the specific set of responses. If one item must be deleted from consideration for an individual because it cannot be recorded or data are lost, the probability for the rest of the response set can still be calculated across the true factor scale and the score assigned based on those probabilities. This is referred to as *invariance*[3], with the consequence that the inclusion/exclusion of any particular item does not bias the scoring – always assuming that missing a response is not in and of itself informative.

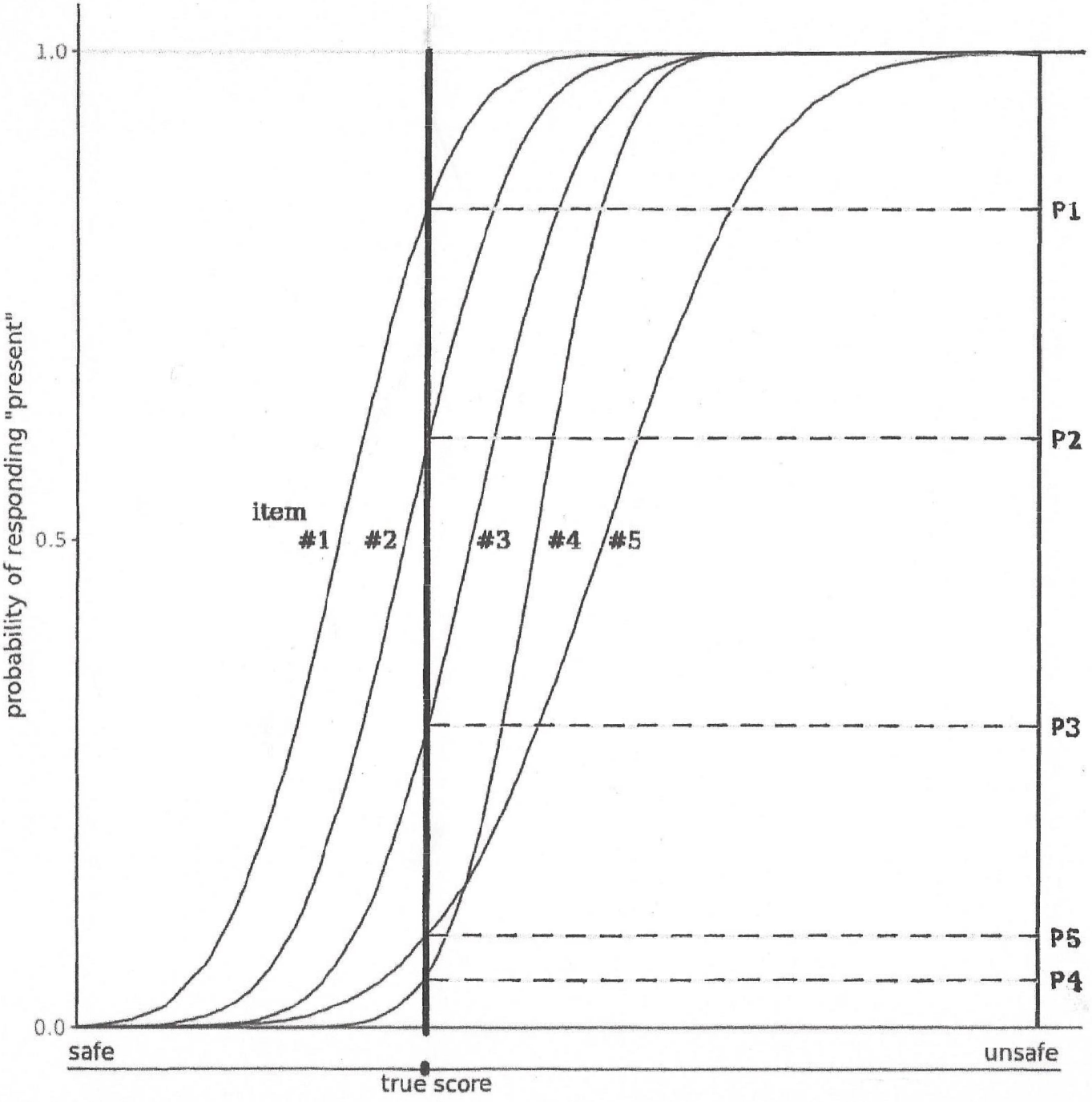
Since the true factor score that maximizes the joint probability of the five item responses also maximizes the sum of probabilities for those responses, IRT literature often refers to "summing the probabilities for the individual items." This also leads to an alternative in scoring paradigm by using the item discrimination to weight the item probabilities, then proceeding to find the true factor score that maximizes this weighted sum.

Reality – Unknown True Factor Scale: An IRT model is self-contained in the sense that it is constructed using its own data base, i.e., responses from the individuals who are to be assigned scores. Therefore without a known True Factor Scale, defining both a scale and scores becomes a joint optimization problem. Computational approaches can involve sophisticated algorithms and be quite efficient in reaching the optimal solution.

However, it is possible to see from a more elementary optimization approach that the (optimal) solution can be attained, albeit laboriously. One such approach alternates between defining the items' curves and scoring individuals. Consider starting from individuals' approximated "true scores" (historical ranks could be used, even randomly assigned ranks could be used although this would be very inefficient). Assume that these scores will be fairly accurate for many individuals and erroneous for others. Based on these scores, create the items' curves – then with these curves, re-score all the individuals. Then iterate as many times as needed, each time recalculating to obtain the items' new curves – and once again re-scoring all the individuals. When the factor scale and the item curves stop changing, scoring cannot change and the process terminates. And both the factor scale and the individuals' scores are determined, completing the process.

Modifications for Different Responses: If responses for some items go beyond absent/present, to includes more categories or even to a continuous measure such as a rate, the concept and paradigm for constructing an IRT model do not change. But the mechanics do and these have been worked out theoretically and computationally. The roles of difficulty and discrimination do not change. For instance, to assess discrimination, look at the

FIGURE #2:
RELATIONSHIP OF MULTIPLE ITEMS TO FACTOR SCORE



distribution of responses at each point along the factor scale. Then compare those distributions to determine their overlap. A highly discriminating item will show relatively little overlap for nearby factor scores and almost no overlap for more distant factor scores. Difficulty can still be linked to the median score.

A Single Comprehensive Score: There are several computational strategies for calculating the single score at stage 2, but the principle illustrated above applies for combining the information in the several (six) stage 1 factor scores. Some software proceeds to model the hierarchy of fundamental

factors at both stages simultaneously; other software proceeds sequentially.

The technical differences and the computational advantages/disadvantages to these modes of solution are beyond the scope of this article. However the other challenges for modeling FMCSA data will have much greater influence on the model's success.

HOW WELL CAN AN IRT MODEL WORK FOR FMCSA?

Challenges for IRT Modeling of FMCSA Data

An IRT model and the scoring system it provides have the potential to work well for FMCSA data, with the three key advantages. First, the model can allow weighting individual items according to their abilities to discriminate all along the scale from safer to less safe. Second, scoring incurs no bias when items are inapplicable or missing for some carriers; and no imputation is made for missing responses. Third, the model is stable since it does not depend on selection of an expert or on the choice of secondary data source or reference to some other resource that could change over time. Fourth, there is no mathematical magic in constructing an IRT model, although there may be computational cleverness especially for very large data files.

To be successful the actual MIRT must satisfy the premises underlying its mathematical construction. The crucial challenges for constructing an MIRT for FMCSA data, however, lie in how the model handles heterogeneity – heterogeneity of the population and heterogeneity in the item responses.

Premises for IRT Models

An IRT model is therefore a mathematical solution that gives the best simultaneous set of item measures (discrimination and difficulty) together with scores for individuals. The mathematics require meeting several conditions.

Premise #1: The presumption underlying an IRT model is that only indirect information is available about an important factor. So the first premise is:

- Taken altogether, the available indirect information (items) gives complete information about the important factor.

The bottom line is that, regardless of its name or its intended meaning, the factor will *only* reflect the indirect information in the actual items used to define it. If new items are added without expanding the coverage of the factor, then the factor will not change and scores can be calculated with or without inclusion of these items. If however, new items are added to expand the scope of the factor, these will modify the definition of the factor.

Premise #2: Like any data-based model, an IRT model depends on the data quality. So the second premise is:

- The data (responses to items) are true, accurate and precise.

The data will be modeled whether they are correct or not; hence any systematic bias will become part of the model. If the bias is strong enough, then when those responses are subsequently corrected, the item curve might even change enough to require model recalibration. Of course, if there is a large amount of measurement error, the model might still be correct but the discrimination would be poor.

Premise #3: The structure of an IRT model is built using two attributes of each item that provides the indirect information: difficulty and discrimination. So the third premise is:

- Relative difficulty of one item is independent of whether a carrier's practices are safer or less safe; and also, difficulty is independent of any other circumstances that would vary among carriers.

In essence this requires that a response of "present" to one item must always represent "less safe" than a response of "absent" by the same or any other respondent. For a scaled response, "4" must always designate greater safety than "5." If counts are used jointly with scaled scores, "4" must be equivalent to 2 x "2."

Premise #4: IRT models reflect the relevance of each item to the underlying factor in contrast with scoring algorithms that most often weight items equally or weight items by difficulty. So the fourth premise is:

- Weights for calculating the summary statistic (final score) should depend on the degree to which each item discriminates between "safer" (lower) and "less safe" (higher) scoring carriers.

Premise #5: Scores calculated from IRT models are valid within the range of factor scores that are represented in the data base used to build the model. So the fifth premise is:

- The data base used to construct the IRT model includes carriers across the full range of "Safety Culture" and across the full range of each of the six component aspects.

While intuitively obvious, it is clear from graphing the all the curves together that at each extreme, all items have probabilities of near zero or all items have probabilities near one. (Figure 2 illustrates this at each extreme of "safe" and "unsafe.") Therefore, distinguishing among scores on either side of the middle range becomes impossible.

Heterogeneity

When IRT models were originally created, they were predicated on the assumptions that the fundamental factor was similarly germane for all the individuals it would be applied to. It was also crucial to inclusion of an item that each possible response have a single meaning.

Thus for IRT modeling of FMCSA data, heterogeneity poses a major challenge; and whether it can be addressed satisfactorily will be a determinant of the success of the model and the scoring system.

The first step is to understand which sources of heterogeneity require attention in constructing an IRT model and to assess the magnitude of the impact. The second step is to develop an approach to address the heterogeneity wisely. Some of the many(!) options include restricting the referent group for each carrier, for example by constructing a separate IRT model for each (large) relatively homogeneous subgroup. Other options focus on the items themselves, e.g., expanding a particular item into a set of items that separately address different subgroups or reweighting responses to an item perhaps using an exposure measure based on carrier attributes or services.

Regardless of the approach or approaches taken, validation of the factors and of the scoring system for each important subgroup is essential to ensure fairness of the scores and to give confidence in the results.

Heterogeneity of population of carriers: The motor carrier industry is extraordinarily heterogeneous and carriers provide multiple kinds of service over greatly different geographic regions and routes. An IRT – or any other data-based – model presumes homogeneity in the absence of information characterizing differences among individuals. Therefore differences in services provided (e.g., long-distance hazmat versus short-distance farm-to-market) could result in different item curves and hence in different scoring equations. On the other hand, for some aspects (e.g., Controlled Substances) the underlying factor may be essentially the same for carriers that are otherwise dissimilar.

Two immediately apparent sources of heterogeneity are the relative "exposure" of each carrier to violation based on auxiliary factors such as geographic distribution of mileage and differences among carriers of the relevance or the comparative importance of particular violations. A third source of heterogeneity is the amount of information available for each carrier and hence the precision with which each can be scored.

Heterogeneity of response information: Serious difficulties are posed when responses are anticipated to be essentially limited to "1," "2," or "3" for one carrier's service type or service region while another carrier over the same time frame or

mileage can realistically incur a "4" or a "5" (e.g., wintertime traveling in the South versus the northern tier of states).

Adjustments are possible for salient differences (e.g., miles traveled in states with low versus high ratios for "speeding : exurban miles traveled"), and again, there are a variety of logically defensible approaches.[4] These adjustments might be made at any level, i.e., in response definitions or transformations or at the first level of the MIRT model where item weights define each aspect (fundamental factor) or at second level of the MIRT model where a single overall measure is created based on the factors together. The purpose is to achieve equivalence of response meaning (as a safety item) across all responders. How successfully the model handles the response heterogeneity will be a determinant of the model's effectiveness.

Additional Questions to Ask

- How are responses being recorded for each item whether binary, polytomous or continuous; are these scaled? In what form are the reported responses for each item entered into the model?
- How is the model being constructed so that it applies to categories of carriers and also to the complete population of carriers?
- How much information (responses to how many items and response distributions based on how many responders) must be available before a carrier can be assigned a score?
- How is the precision of each carrier's score being quantified and quoted?
- How much impact can any single item contribute to a carrier's score? Is there a limit?
- How is the model being vetted or validated for overall performance? How is the model being vetted or validated for performance with respect to important subgroups of carriers?
- How will model performance be monitored for anomalies once the model is put into use?
- How will scores be published? What referent group (total population or specified subgroup) will be used in publishing scores?

SUMMARY

The good news is that an IRT model has the potential to provide a stable and fair scoring system. Whether it can achieve this goal will depend on the availability of accurate relevant data on all the important aspects of "Safety Culture." Success will also depend on how well the truly difficult challenges of the heterogeneity of carrier industry can be encompassed by the final model and scoring system. The details will be telling – until these are known and the model is fully vetted the IRT model remains a potential waiting to be realized.

ENDNOTES

1: For an MIRT (multidimensional)model, separate scores are often reported for each dimension. There are ways of combining those separate scores into a single score, but that typically occurs outside the actual model-fitting process.

2: If information is sparse or lacking at some points along the true factor scale, then the standard practice would be either to fit a smooth function to the available responses or to interpolate smoothly so that the final curve is monotone increasing.

3: Technically the term *invariance* is typically used to imply that the item parameters, the difficuilty and the discrimination, stay the same regardless of which respondent is considered or which population is used to develop the model or which population is is

applied to. Likewise the response parameters, and hence the score(s), for each respondent stay the same regardless of which items are administered. Application to the case of FMCSA data is considered in later sections of this article.

4: Of the wide range of options, a few examples are rescoring or rescaling responses based on auxiliary information. For instance, mileages could be separated by state or reweighted based on an exposure measure such as an index for a state's rate of issuing violations. Alternatively, responses could be relativized (actual compared expected) based on a "norm" or expectation for comparable carriers taking into account the relevant carrier attributes or transport and route patterns. At the level above responses to individual items, aspects could be weighted separately for different types of carriers or reweighted in accord with carrier attributes. It would also be possible to reweight aspects in accord with total information available, equivalent to reweighting in accord with the precision of quantification of the aspect. It would also be possible to take heterogeneity into account through score calculation, the determination of the referent group of carriers and /or by the relative risk or measure of exposure. This does not begin to exhaust the potential logical approaches, but rather to underscore the options for effectively handling heterogeneity.

REFERENCES

Hambleton, R.K., Swaminathan, H., and Rogers, H.J. (1991). *Fundamentals of Item Response Theory*, Newbury Park, California: Sage Publications, Limited.

National Academies of Sciences, Engineering and Medicine (2017). *Improving Motor Carrier Safety Measurement*, Washington, DC: National Academies Press. doi: https://doi.org/10.17226/24818.

van der Linden, Wim. (2018). *Handbook of Item Response Theory*, Boca Raton, Florida: CRC Press.

DISCLAIMERS AND NOTES

*- Disclaimer and Acknowledgments - The content of this article, the representation of the concepts, and views expressed herein are those of the author and should not be construed to represent those of the National Institute of Statistical Sciences.

** -This article is written about IRT methodology per se. The author is not privy to the actual modeling being done on contract to FMCSA, nor is it known how closely this modeling follows the recommendations in the National Academies of Sciences, Engineering and Medicine report (2017).

BIOGRAPHY

Nell Sedransk received her PhD in Statistics from Iowa State University. She is an Elected Fellow of the American Statistical Association, the International Statistical Institute, and the American Association for the Advancement of Science. She has coauthored three books and published research on statistical theory and application in refereed statistical, scientific and engineering journals. Following her career as professor of mathematics and statistics, she moved to National Institute of Standards and Technology as Chief of Statistical Engineering, later joining the National Institute of Statistical Sciences, ultimately becoming its third Director. E-Mail: nsedransk@niss.org