

## Making Multiple Regression Narratives Accessible: The Affordances of Wright Maps

Alexander Mario Blum  
*Stanford University Graduate School of Education*  
*Enrich Your Academics*

James M. Mason  
*University of California, Berkeley*


Aaryan Shah  
*Stanford University*


Sam Brondfield  
*University of California, San Francisco*


Wright Maps have been an important tool in promoting meaning making about measurement results for measurement experts and substantive researchers alike. However, their potential to do so for latent regression results is underexplored. In this paper, we augmented Wright Maps with hypothetical group mean locations corresponding to realistic scenarios of interest. We analyzed data from an instrument measuring cognitive load experienced by medical fellows and residents while performing inpatient consults. We focused on extraneous load (EL: i.e., distraction) and variables potentially associated with distraction. Through collaborative examination of the Wright Map, we found not only corresponding regions to construct levels but also a region with important practical consequences, namely that the third threshold represented a critical level of cognitive load, which could impact patient care. We augmented the Wright Map with the locations of two typical scenarios differing only in novelty of the consult, representing the lowest and highest levels of novelty, respectively. These group locations were plotted on the Wright Map approximately 1.5 logits apart, allowing for a kind of *visual relative effect size*, as this difference can be perceived relative to other features of the Wright Map. In this case, both scenarios were within the same band of the Wright Map, leading to the practical interpretation; although EL was significantly increased, the risk of cognitive overload was not. However, because of the problematic nature of the third threshold, a 1.5 logit difference does not have the same practical consequences along the entire scale; other realistic scenarios with increased initial EL are possible, where increased novelty could lead to cognitive overload. This area of visualization techniques, along with a combinatorial view of a multiple-regression analysis, could be helpful in other substantive and practical contexts, and with more complex regression models.

**Keywords:** Latent regression, Wright map, item-response theory, visual relative effect size, combinatorial interpretation

---

Alexander Mario Blum  <https://orcid.org/0000-0002-5887-7417>

James M. Mason  <https://orcid.org/0000-0002-3549-638X>

Sam Brondfield  <https://orcid.org/0000-0001-7656-7490>

Requests for reprints should be sent to Alexander Mario Blum, 1900 Powell St., Suite 700 Office #712, Emeryville, CA 94608, USA; email: Alexander.M.Blum@gmail.com

Wright Maps, aka item-person maps (Embretson, 1996), are “central tool[s]” (Wilson, 2017, p. 80) to understand the dynamics of different types of cognition being modeled using Rasch family models (Adams et al., 1997; de Boeck & Wilson, 2004; Masters, 1982; Wright & Mok, 2004). They have been applied to latent variables (Borsboom, 2005) as diverse as sense of belonging (Stachl & Baranger, 2020), inferential thinking (Blum et al., 2020), and social skills (Jolin & Wilson, 2022). Item-person maps for different variables often look different. The unique appearance of, and patterns suggested by, each item-person map serve as a springboard for meaningful interpretation of the variable being modeled.

An essential ingredient to producing an item-person map is the use of a Rasch-family model. Such an analysis, whether using a dichotomous model such as Rasch’s simple logistic model (Rasch model; Rasch, 1980) or a polytomous model such as the partial credit model (PCM; Masters, 1982), produces a table of numeric coefficients for items (often called “difficulties” in achievement contexts) and numeric coefficients for persons (often called “abilities” in achievement contexts). Critically, if the model fits, these coefficients (whether for persons or items) can be interpreted as being on the same scale. Note that this scale is an *interval* scale (Stevens, 1946), so the origin (or “zero”) of the scale is arbitrary. These coefficients are only interpretable relative to each other: the *differences* between coefficients are what matter, rather than their particular values. In an achievement context, items that have numerically higher difficulty (relative to other items) are considered “harder” items, and persons with a numerically higher proficiency (relative to other persons) are considered to demonstrate having more of the construct being modeled. The inverse applies to numerically lower item difficulties (“easier” items) and numerically lower proficiencies (respondents having less of the construct). Numeric differences between person parameters and item parameters are a function of the relative probability of a “correct” response to the item (in

an achievement context; other interpretations of “1” vs. “0” apply in other contexts). We will discuss the nature of this function in detail in the Methods section.

However, looking at a series of tables and comparing numerical values across respondents and items can be cumbersome, especially if there are a lot of respondents and items. The quantity of interpretations to be made may be overwhelming, making patterns hard to discern. For example, imagine a list of cities along a particular highway with the mile marker numbers for each city. One could technically compute a large number of distances from the differences between all the mile markers, but it would be a lot easier to look at a visualization of those between cities on a map. A map could lower the cognitive load of interpreting these relative distances and allow a perception of the locations of the cities. Note that for a visualization to be considered a map, it is necessary that the underlying scale be at least interval.

Benjamin Wright was a big champion of the item-person map as a visualization technique. Because of his contributions both in advancing item map techniques, and in promoting their use as a central tool in interpreting measurement results, item maps are often called “Wright Maps” (Wilson, 2017, p. 80) in his honor. A Wright Map begins with the vertical axis as a “ruler” representing a useful range of the interval scale (in logits) generated by the Rasch analysis. The person coefficients are plotted to the left of the ruler, often as a histogram. In the dichotomous case, the item coefficients are plotted to the right of the ruler; in the polytomous case, Thurstonian thresholds are plotted instead. Either way, various horizontal arrangements of the items are possible. This visualization, through the metaphor of a map, affords the interpretation of these coefficients as *locations*.

If latent variables represent different types of cognition, generating a Wright Map can be seen as a sort of cognitive cartography. By doing so, one can visually perceive the

locations of items and respondents, as well as relationships between them. Furthermore, items can be arranged in various ways, and color-coded for their types, helping to illuminate patterns of items relative to each other and to respondents (Zhao et al., 2023). For example, polytomous items can be arranged in order of increasing the threshold location (e.g., Zhao et al., 2023), or any other threshold of interest. This allows inferences to be made not only by methodologists but also by substantive researchers, leading to rich interdisciplinary conversations, potentially generating new research ideas.

Regression analysis is often used to address substantive questions in many disciplines, as well as to provide validity evidence for measurement instruments, for example, evidence related to relationships with other variables (American Educational Research Association et al., 2014). If the dependent variable is latent, then a latent regression (de Boeck & Wilson, 2004) is more appropriate because, while linear regression accounts only for sampling error, latent regression also accounts for the measurement error inherent in the use of a latent variable. As with linear regression, both simple and multiple latent regression are possible.

In a simple regression (i.e., on a single variable), the regression coefficient is much easier to interpret than the coefficients in a multiple regression. For example, in a simple (latent) regression context, one could say that for every one-unit increase in the regressor, respondents are estimated to have, on average, a certain increase (or decrease) in the latent variable, based on the estimated regression coefficient. But often researchers want to control for other independent variables when estimating the effect of their variable of interest, or they may be interested in the effects of multiple independent variables. Either way, a multiple (latent) regression analysis is needed. When this is done, the effects of each of the regressors must be interpreted as the effect of that variable alone, when “holding the other variables constant.”

But what does this mean? Technically, this means that the regression is comparing the average of the independent variable when the regressor in question differs, but the values of the other regressors do not. But this is a highly decontextualized view of the variables, considering each variable in isolation, without considering the cumulative effect of all the regressors on the dependent variable. This decontextualized view may be sufficient for purely academic research (e.g., publishing a paper on the effect of a variable), but making use of such research is another matter. In a practical context, values of dependent variable often represent people; practitioners, therefore, will be interested not only in the effect of one variable in isolation but also in the effect of that variable for a particular person, considering the context implied by the values of the other variables in a combinatorial fashion. Consider an example where the effect of a dichotomous regressor  $X_1$  is estimated to be a two-unit increase in the dependent variable  $Y$ . Depending on the values of the other regressors, this might mean that  $Y$  changes from 2 to 4 or from 8 to 10. However, the dependent variable  $Y$  may have a non-linear effect on the real world (for example, 9 might be a critical level for  $Y$ ). Examples of this can be diverse as stress (a slightly increased stress may matter more for a patient who is already stressed) and end-of-year exams (a slightly higher math score may allow a student who is almost proficient to pass while making little difference for a student who is much further behind); recognizing this type of situation requires the perspective of substantive researchers in the discipline and also, sometimes, of practitioners.

This is the situation we address in this paper: just as Wright Maps help facilitate interdisciplinary discussions around a measurement context, they also can help one understand results from latent regression—but how? There are two features that are central to this method: (a) that the regression coefficients are decontextualized because they represent the effect of each regressor in isolation, and (b) that the regression coefficients are in logits (i.e.,

they are differences in locations on the logit scale).

An illustrative case is provided by (Brondfield et al., 2021), which measures three types of cognitive load in medical fellows providing inpatient consults, and then examines the effects of several contextual variables on these types of cognitive load (see methods for more in-depth details of the study). For purposes of this paper, we focus on one type of cognitive load, *extraneous load* (i.e., distraction), and four predictor variables: the fellows’ self-rated expertise relative to the consult, their self-rated evaluation of the novelty and difficulty of the consult, and the number of prior consults the fellow has done during that shift. This is a particularly illustrative example because the authors of that study were interested not only in the isolated effects of each variable but in their *combined* effect on the fellows’ extraneous load, since too much distraction is undesirable.

### Research Question

How can we use a Wright Map to represent a holistic lens to interpret multiple regression analysis, rather than a silo lens, to make practical meaning? As applied to this case, how can a Wright Map be used to contextualize the results of a latent regression of extraneous load on expertise, novelty, difficulty, and prior consults in a way that is accessible and pragmatic to substantive researchers and practitioners?

## Methods

### Overview

As mentioned above, this is a secondary analysis of data collected as part of a study that was published in (Brondfield et al., 2021) investigating various types of cognitive load. In this paper, we focus specifically on one type of cognitive load, extraneous load (EL), and address both measurement issues and issues of substantive interpretation and practical application that become apparent during analysis and collaborative discussions centered on Wright Maps.

### Materials

We used the four EL items from the Consult Cognitive Load instrument (Brondfield et al., 2021). In that study, the instrument had good fit with no item fit statistics exceeding 1.33, and there was a consistent increase across mean location of respondents within each category, adding to an argument for the ordinal nature of the construct and its internal structure validity. Reliability estimates for this construct were 0.78. These items, which we will call Item 1 through Item 4, respectively, ask, using a four-point Likert scale, ranging from (1) strongly disagree to (4) strongly agree. Each item has the same sentence stem, “During this consult...”. Item 1, interruptions distracted me; Item 2, thoughts or emotions that were unrelated to the consult distracted me (these could be positive or negative, work-related or not.); Item 3, extraneous, irrelevant, or redundant information (oral or written) distracted me; Item 4, “technology or equipment distracted me (these could be work-related or not. Exclude devices directly used to perform a procedure on the patient). For an abbreviated use of item names, Item 1 will be named “interruptions;” Item 2, “thoughts and feelings;” Item 3, “information;” Item 4, “technology.”

Response levels were recoded to 0, 1, 2, 3 from 1, 2, 3, 4 to avoid empty categories.

In terms of the relationship between predictor variables and the construct under investigation, Novelty of consult topic to the learner predicted higher EL, and number of prior consults during the shift predicted higher EL. The authors anticipated these results and the cognitive load literature (Sewell et al., 2019).

### Participants

Participants were fellows in several medical specialties at a large system of teaching hospitals in California (see Brondfield et al., 2021). As found in Brondfield et al. (2021, p. 1735), “from March to September 2019, excluding July to avoid the transitional month, we emailed the survey to all 253 rotating

internal medicine fellows and psychiatry residents and fellows identified by their program directors at five University of California clinical campuses (Davis [n = 39], Irvine [n = 35], Los Angeles [n = 59], San Diego [n = 65], and San Francisco [n = 55]) and sent weekly reminders. As with the pilot, participants were asked to complete the survey within 24 hours of a new consult.”

Modeling

Software

Item-response models were fit using ACER Conquest Version 3 (Adams et al., 2012).

Model 1: The Partial Credit Model

The partial credit model (PCM; Masters, 1982, 2016; Wright & Masters, 1982) was used to measure extraneous cognitive load. The formula is as follows for Equation 1:

Prob(X<sub>vi</sub> = m | θ<sub>v</sub>, δ<sub>ij</sub>) = 
$$\frac{\exp \sum_{j=0}^m (\theta_v - \delta_{ij})}{\sum_{k=0}^x \exp \sum_{j=0}^k (\theta_v - \delta_{ij})}$$
 (1)

This is the Rasch-family model, so if it fits, it can be said to place the items and the respondents’ extraneous load on a common logit scale. However, PCM models polytomous items (items with more than two response levels). As can be seen from the left side of the equation, PCM models the probability that a person with EL of θ<sub>v</sub> will respond to item *i* at category *m* rather than any other category. An item with categories 0, 1, . . . *m<sub>i</sub>* (*m<sub>i</sub>* + 1 categories) will be characterized by *m<sub>i</sub>* item parameters δ<sub>i1</sub>, . . . δ<sub>im<sub>i</sub></sub>. These parameters are often called *step* difficulties because, as can be seen from exponential sums on the right side of the equation, they utilize *adjacent-category* logits. (This means, for example, that δ<sub>i3</sub> is based on the odds of attaining Level 3 on item *i*, conditional on *already having attained* Level 2 on the item). It is often more appropriate to compute and use Thurstonian thresholds for the Wright Map, as these thresholds have

*cumulative-category* semantics (e.g., based on the odds of attaining Level 3 or above on item *i* vs. Level 2 or below on that item). EL has four levels; see Brondfield et al. (2021) for the EL Construct Map and Wright Map for EL.

Model 2: Latent Regression PCM

Latent Regression (de Boeck & Wilson, 2004) can be seen as a special type of regression model which incorporates an item-response measurement model for the dependent variable or as a type of explanatory item-response model that uses regression to explain the effects of various predictors on the latent variable being measured. The formula for a Latent Regression PCM can be obtained from the PCM formula, by replacing θ<sub>ν</sub> with a regression model as follows for Equation 2:

$$\theta_{\nu} = \beta_0 + \sum_{j=1}^J \beta_j X_{pj} + \varepsilon_{\nu}$$
 (2)

In our case, *J* = 5; the five regressors are described in Table 1. In this equation, *X<sub>pj</sub>* is person *p*’s value regressor *j* and β<sub>*j*</sub> is the fixed regression weight for that regressor. The intercept is β<sub>0</sub>, and we assume that ε<sub>ν</sub> is a normally-distributed error term, representing the residual EL after the effects of all the regressors on person *ν* are accounted for.

Since latent regressions incorporate both a measurement model and a regression model, they are able to account for the sources of error modeled by each (i.e., measurement error and sampling error, respectively). When the dependent variable is latent, this is more appropriate than a regular multiple regression using *score estimates* from a measurement instrument since doing so treats these estimates as though they were observed variables.

Another advantage of using a latent regression is that the estimated coefficients for each regressor in a multiple regression analysis (the regression weights β<sub>0</sub>, . . . β<sub>6</sub>), are in the same units as the measurement model (i.e., in the logits from the PCM). Accordingly, we can use the regression outputs, including the intercept, to plot them on the same scale as the

locations of the respondents and items. This adds a new layer of interpretation, allowing results from latent regression to be interpreted relative to the distribution of persons and especially relative to item thresholds.

How to Plot Multiple Regression Analysis on a Logit Scale: An Estimated Hypothetical Group Location

Results from regression analysis, latent regression included, are typically presented as a table of regression coefficients (usually including the intercept, though sometimes this is omitted) along with standard errors, p-values, confidence intervals, or some other indicator of statistical significance. This is a very information-dense and efficient way of presenting the results, providing a quick view of which regressors were statistically significant and some idea of the effect size of each regressor in isolation. However, as mentioned above, this is a decontextualized view, generally more suited to research purposes and less suited to bridging that research into practice. Recontextualizing latent regression results involves computing the cumulative effect of all the regressors (both those being viewed as “predictor” variables and those being viewed as “controls”) on the location of a person or group of people. While regression can be viewed either from an individual standpoint (using the full model, including the error term and its distribution) or a group standpoint

(including the conditional mean), *predictions* from regression, because they involve only the linear predictor part of the model, are first and foremost about the means of subgroups (often hypothetical subgroups) of the population. It is this latter view that we follow: we compute means for two or more subgroups, then contrast these means in such a way that each contrast can show the effect of a variable of interest while keeping the others at specified values.

Selecting these values for all the variables in the model is another case where collaboration with substantive researchers and practitioners is important. These groups, defined by the values chosen, represent scenarios that are both realistic and different in ways that are of substantive interest.

Collaborative Example

Since a key part of this method is collaboration between measurement, substantive researchers, and practitioners, we conducted additional analyses on the (Brondfield et al., 2021) data set, which has been a fertile ground for such rich collaborations. We ran the PCM for EL, followed by a multiple latent regression using a reduced set of predictor variables (compared to Brondfield et al., 2021): (a) Novelty, (b) Expertise, (c) Difficulty, (d) Year in the Program, (e) Prior Consults that day, and (f) Hours of Sleep, as predictor variables. These variables are described in Table 1.

Table 1  
Regression Variables

Variable	Item	Scale
Novelty	Prior to this consult, how familiar were you with consults similar to this one?	1-very familiar through 4-very unfamiliar
Expertise	Before you began this consult, what was your level of expertise with the primary [medical/psychiatric] issue in this consult?	1-novice through 5-expert
Difficulty	How easy or difficult was this consult relative to others you’ve done [during your current fellowship]?	1-very easy through 5-very difficult
Prior consults	During the current workday or shift, how many other new consult requests (not including curbsides) did you receive prior to this one?)	Count
Hours of sleep	How many hours of sleep did you get the night before you received this consult request (i.e., the night before this shift started)?	Hours

We chose values of the predictor variables to contrast two realistic training scenarios:

**Scenario 1:** Keeping these variables constant with the following values multiplied times their corresponding coefficient: Sleep (6 hours), Difficulty (3: about average), Expertise (3: competent), Prior Consults (their first consult of the day), and Set Novelty to (1: very familiar).

**Scenario 2:** Keeping these same variables constant with the exception of setting novelty to (4: very unfamiliar).

We then used these variables in the (linear predictor portion of) the latent regression, along with the estimated regression coefficients, to generate predicted group means for these scenarios as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$\hat{Y}_1 = \beta_0 + 1\beta_{novelty} + 3\beta_{expertise} + 3\beta_{difficulty} + 0\beta_{consults} + 6\beta_{sleep}$$

$$\hat{Y}_2 = \beta_0 + 4\beta_{novelty} + 3\beta_{expertise} + 3\beta_{difficulty} + 0\beta_{consults} + 6\beta_{sleep}$$

This was to tease out the impact of novelty being extremely familiar or novel with other variables held constant.

**Results**

**Model 1**

The item and person locations shown in Figure 1 were estimated using Model 1 (PCM). As noted above, the items are about interruptions (Item 1); thoughts or emotions (Item 2); extraneous, irrelevant, or redundant information (Item 3); and technology or equipment (Item 4). As seen on the right side of the figure, the second thresholds (“strongly agree” & “agree” vs. “disagree” & “strongly disagree”) of most items (except Item 1)

were around 0 logits, and the third thresholds (“strongly agree” vs. “agree” and below) were around 3 logits. As both in Figure 1 and in Table 2, the second and third thresholds for Item 1 (about interruptions) were over 1 logit lower than the other three items, at -1.16 and 1.37 logits, respectively. The distribution of persons (the histogram of “X”s) was roughly symmetrical around 0 logits, with about two-thirds between -2 and 1.5 logits, but with a heavy upper tail with a few persons as high as 5 logits (who were therefore not well measured, being distant from all the item thresholds).

In collaborative discussions around the Wright map, it became clear that the third threshold, the level at which respondents were equally likely to choose “strongly agree” versus any of the lower categories, was of particular importance. This was deemed to be a very undesirable level of EL, where the level of distraction could lead to concerns around patient safety. With Item 1 having the lowest third threshold (1.37 logits), it could be viewed as a canary in a coalmine: reaching this level of EL on even one item is problematic in and of itself and indicative of more serious problems to come.

**Table 2**

*Thurstonian Thresholds From Model 1*

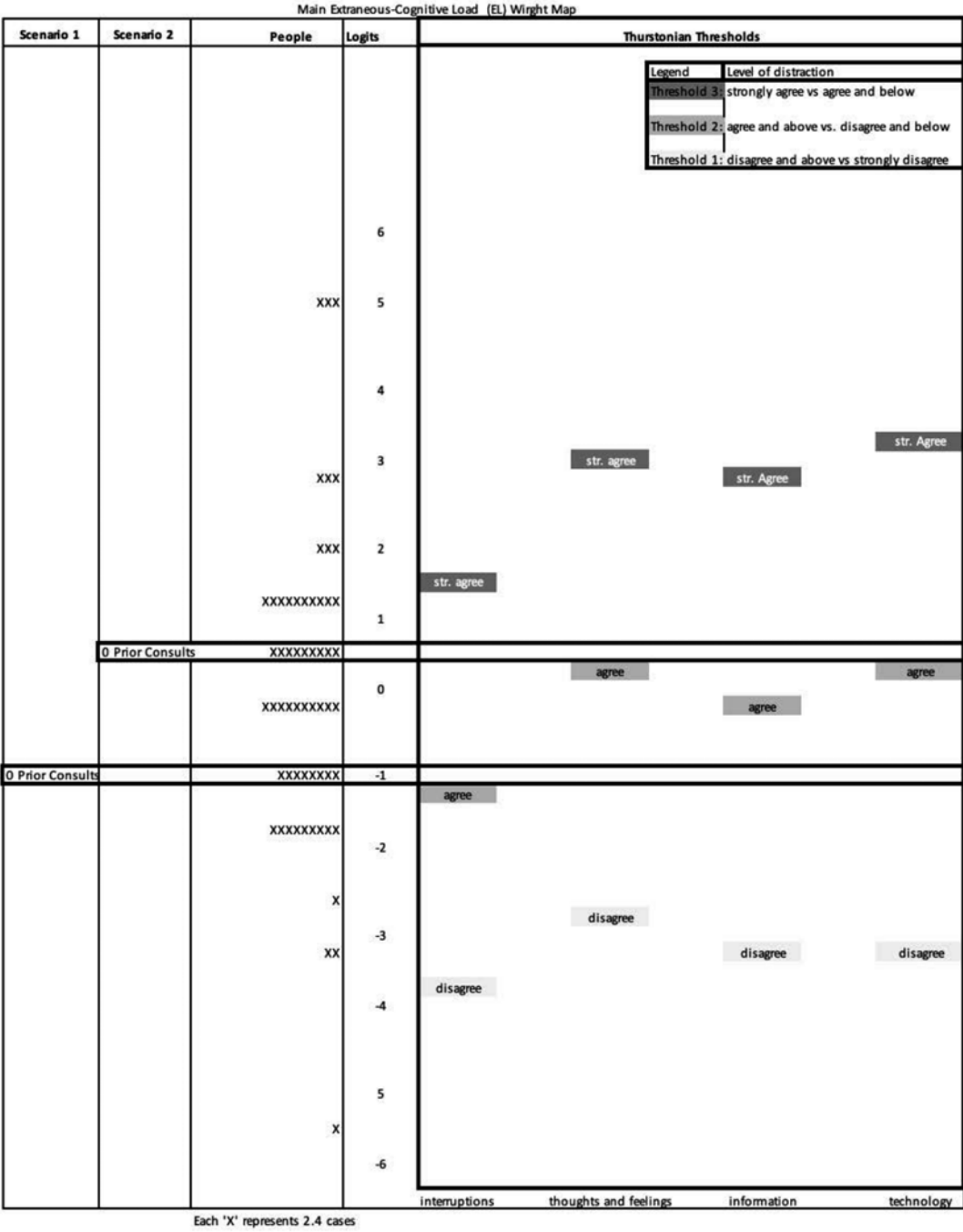
	Thresholds		
	1st	2nd	3rd
Item 1	-3.80	-1.16	1.37
Item 2	-2.87	0.32	3.04
Item 3	-3.31	-0.07	2.76
Item 4	-3.37	0.37	3.32

**Model 2**

For Model 2, Latent Regression PCM was fit with item parameters anchored to their values from Model 1. This ensures that the scales generated by the two models are the same, and thus, Figure 1 can still be used for interpretation.

**Figure 1**

*Wright Map for EL*



Note. \*The original version of the Wright Map for EL can be found in the supplemental documents in Brondfield et al. (2021).

**Table 3**  
*Latent Regression Coefficients From Model 2*

Variable	Coefficient	SE
Intercept	0.237	(1.239)
Novelty	0.506*	(0.248)
Expertise	0.290	(0.197)
Difficulty	−0.417	(0.232)
Prior consults	0.386*	(0.077)
Hours of sleep	−0.234	(0.132)

The regression parameters estimated by the model are shown in Table 2. Only Novelty and Prior Consults were statistically significant at the  $p < .05$  level. Each one-unit increase in novelty (on a four-point scale) was associated with an estimated increase in mean EL of 0.506 logits, and each additional Prior Consult was associated with an estimated increase in mean EL of 0.386 logits. We explore the implications of this former effect (about half a logit increase per unit of novelty) using a Wright Map as a key visualization and focus of collaboration. Each scenario represents the same predictors and regression coefficients. Only Prior Consults are labeled to focus on the implications of the first consult of the day within two given scenarios. So, to be clear, the location of each scenario is a different combination of the same regression coefficients set to different values with varying degrees of novelty. Scenario 1 represents computed group means for the hypothetical groups representing Scenarios 1 and 2 as follows:

$$\begin{aligned}\hat{Y}_1 &= 0.237 + 1(0.506) + 3(0.290) + 3 \\ &\quad (-0.417) + 0(0.386) + 6(-0.234) \\ &= 0.237 + 0.506 + 0.870 - 0.441 + 0 - \\ &\quad 1.404 \\ &= -1.042 \\ \hat{Y}_2 &= 0.237 + 4(0.506) + 3(0.290) + 3 \\ &\quad (-0.417) + 0(0.386) + 6(-0.234) \\ &= 0.237 + 2.024 + 0.870 - 0.441 + 0 - \\ &\quad 1.404 \\ &= 0.476\end{aligned}$$

The locations for the two scenarios (−1.042 logits and 0.476 logits, respectively) are indicated in the two left columns of Figure 1 (left of the person histogram). Additionally, horizontal bars across the Wright Map help show these locations relative to the person distribution and the item thresholds.

Contrasting Scenarios 1 and 2, we see that a three-unit difference in novelty (i.e., the greatest possible difference) is associated with approximately a 1.5 logit difference in EL, as expected from the regression coefficients. As seen in Table 2, the difference between successive thresholds (Thresholds 1 to 2 or Thresholds 2 to 3) is around 3 logits. Since we have an interval scale, such differences can be compared; we find the effect of novelty on EL is about half of the inter-level difference. The size of this difference on the Wright Map can be seen as a kind of *visual relative effect size*.

However, the implications of such an effect may not be the same across the entire scale; this is a matter of practical interpretation. As noted above, the third threshold is of particular importance since it indicates a potentially dangerous level of EL. In this case, both Scenario 1 and Scenario 2 are well within the second level (the band defined by the second thresholds, where fellows may be distracted, but not dangerously so). Fellows in Scenario 2 would be near the high end of this band, whereas fellows in Scenario 1 would be near the low end (i.e., regardless of whether novelty is 1 or 4, fellows are still likely to be endorsing items in the same threshold range). In other words, novelty matters here, but not critically so. For fellows in Scenario 1, increasing the novelty of the consult could probably be done safely and may even be beneficial to their training. However, there are other realistic scenarios where this would not be the case. Imagine another low-novelty scenario, but where other variables are different (e.g., Less Sleep or more Prior Consults) so that the estimated location is around 1 logit (at the very top of the second level); in this case, a three-unit increase in novelty would result in a location

around 2.5 logits, which is well above the third threshold for interruptions and just below the other third thresholds. This is well into the third level, a very undesirable EL indeed. Again, it was because of collaboration with substantive researchers and practitioners that we were able to see this very different interpretation of the practical consequences of increased novelty (a potential training benefit in one case vs. dangerous cognitive overload in the other).

### Discussion

The juxtaposition of latent regression and Wright Maps provides some interesting affordances. The locations of hypothetical groups (or scenarios) can be interpreted both relative to the person distribution and also relative to the construct (via the item thresholds).

Note that if measurement is done based on a theoretically driven construct with meaningful levels, and if the theory is borne out by the data, then resultant Wright Maps *can be* interpreted relative to that construct. Of course, if there was *not* a construct, person locations (i.e., scores) are not interpretable (Blum et al., 2020), even if the model fits. The logit scale generated (if the model fits) is interval, and therefore distances can be compared, as they can on a map... but a *map of what?* If the measurement was based on a construct, which was supported by the Wright Map, then it is indeed a map of the variable defined by the construct.

### Visual Relative Effect Size

Typically, in interpreting a linear regression, the focus is first on which variables achieved statistical significance (i.e., are we confident that the effect we saw was real rather than a result of sampling error), effectively treating statistical significance as a climax of the research effort. But statistical significance is just the beginning of interpretation as this is the most decontextualized presentation of the findings. At the other end of the spectrum, we have the raw effect size. This is typically some form of difference in means, and as such, it

is in the same units of the outcome measure. These units might be pounds, raw test scores, or logits. If the units are interpretable, such as miles or pounds, then the raw effect size can be useful. A standardized effect size is obtained by dividing the raw effect size by the spread of the sample (i.e., standard deviation), but in this process, it becomes decontextualized, losing its units (standardized effect sizes are always in units of standard deviations). People understand what 100 lbs. (45 kg) means and what a 500-mile drive can mean, but when this is not the case, standardized effect sizes are, maybe, necessary for interpretability. Are logits directly interpretable (after all, they represent a latent construct)? That depends on whether you have waypoints, which is one of the four building blocks of the BEAR assessment system (i.e., construct mapping; Wilson, 2004, 2023).

The visual relative effect size is a representation of a raw effect size on a Wright Map, enabling interpretation relative to waypoints if available. That is, we now have (a) information about its impact in different ranges along the scale, (b) if the impact is relatively more meaningful (such as crossing waypoints), and (c) whether the impact is being extrapolated, either outside the range of the waypoints or the range of the sample. After all, the regression line doesn't go forever on both ends; it has a starting point and an end point defined by the boundaries of the external variable.

This visual representation of the range of the external variable's relative effect and its relationship to waypoints is what brings a multiple regression analysis to life by providing more context for interpreting and discussing the construct among interdisciplinary team members.

### Towards a Type of Cognitive Atlas

Measurement is always done in some context. Thus, the variable so measured is not context-free. To understand that context requires collaboration with substantive researchers and especially with practitioners in that context.

The Wright Map is a very valuable visualization, but the map is not the territory. For example, there may be clear banding or other patterns of thresholds on the Wright Map, marking out interesting zones of the variable relative to the construct levels (i.e., waypoints). Returning to the map metaphor from the introduction, these waypoints can indicate the location of a city or a point of interest, but they don't include important local knowledge, such as topology, traffic, the safety of neighborhoods, and the significance of local landmarks. In other words, to really understand the *landscape of the variable* in a particular context requires "local knowledge" from those familiar with that context.

#### Person Side: The Value of Seeing Groups Relative to the Person Distribution

How realistic are these hypothetical groups in defining the contrasts used to contextualize the results of the latent regression? Ideally, the groups would represent realistic combinations of regressor values, but that doesn't mean that the locations of these groups can have sufficient support to be well estimated by the latent regression.

By plotting the location of these groups near the person histogram on the Wright Map, it is possible for each group to see how many cases are located near it. If there are many cases in the vicinity, then we can be fairly confident in our conclusions about group location. On the other hand, we may have less confidence the cases are distant from the estimated location: if the cases are on both sides, then the estimation is an interpolation; if most of the cases are on one side, then the estimation is an extrapolation (i.e., we are predicting out of sample), and caution is warranted in interpreting the results.

An additional type of visual relative effect size can be seen here by comparing the difference between scenarios to the overall spread of the person histogram.

#### Item Side: The Value of Seeing Groups Relatives to the Landscape of the Variable

By placing estimated group locations on the Wright Map, they can be visually compared to construct levels (via waypoints). Moreover, these locations can be interpreted relative to potential real-world consequences, if collaboration with substantive researchers and practitioners has illuminated some of the landscape of the variable.

In this case, the highest thresholds were seen as an important critical level of the variable as an indication of overload or potential burnout. Additionally, it was apparent that fellows were more sensitive to interruptions, allowing that item (which had a lower top threshold) to act as a warning sign of imminent trouble. This problematic nature of the highest threshold also afforded the interpretation that a large difference in novelty was not dangerous in the contrast we selected but could have been for other potential contrasts.

#### Conclusion

To bring this all together, the mission of this paper is to provide a tool to champion collaborative discussion among all stakeholders. By contextualizing multiple regression analyses in this holistic way, more informed inferences can be made, backed by empirical evidence. Now, inferential relationships between external variables and the construct can be situated relative to waypoints, giving rise to pragmatic interpretations.

#### References

- Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21(1), 1–23. <https://doi.org/10.1177/0146621697211001>
- Adams, R. J., Wu, M. L., & Wilson, M. (2012). *ACER ConQuest: Generalised item response modelling software* (Version 3) [Computer software]. Australian Council for Educational Research.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in

Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

- Blum, A. M., Mason, J. M., Kim, J., & Pearson, P. D. (2020). Modeling question-answer relations: The development of the integrative inferential reasoning comic assessment. *Reading and Writing: An Interdisciplinary Journal*, 33(8), 1971–2000. <https://doi.org/10.1007/s11145-020-10026-4>
- Borsboom, D. (2005). Latent variables. In *Measuring the mind: Conceptual issues in contemporary psychometrics* (pp. 49–84). Cambridge University Press.
- Brondfield, S., Blum, A. M., Lee, K., Linn, M. C., & O'Sullivan, P. S. (2021). The cognitive load of inpatient consults: Development of the consult cognitive load instrument and initial validity evidence. *Academic Medicine: Journal of the Association of American Medical Colleges*, 96(12), 1732–1741. <https://doi.org/10.1097/ACM.00000000000004178>
- de Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer. <https://link.springer.com/book/10.1007/978-1-4757-3990-9>
- Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment*, 8(4), 341–349.
- Jolin, J., & Wilson, M. (2022). Developing a theory of two latent soft skills progress variables using the BEAR assessment system: Validity evidence for the internal structure of the social evaluative in the workplace instrument. *Journal of Psychoeducational Assessment*, 40(3), 381–399. <https://doi.org/10.1177/07342829211057641>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Masters, G. N. (2016). Partial credit model. In

W. J. van der Linden (Ed.), *Handbook of item response theory* (Vol. 1, pp. 109–126). Chapman and Hall/CRC.

- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (Expanded ed.).
- Sewell, J. L., Maggio, L. A., Ten Cate, O., van Gog, T., Young, J. Q., & O'Sullivan, P. S. (2019). Cognitive load theory for training health professionals in the workplace: A BEME review of studies among diverse professions: BEME Guide No. 53. *Medical Teacher*, 41(3), 256–270.
- Stachl, C. N., & Baranger, A. M. (2020). Sense of belonging within the graduate community of a research-focused STEM department: Quantitative assessment using a visual narrative and item response theory. *PloS One*, 15(5), e0233431. <https://doi.org/10.1371/journal.pone.0233431>
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. <https://doi.org/10.1126/science.103.2684.677>
- Wilson, M. (2004). *Constructing measures: An item response modeling approach* (1st ed.). Routledge.
- Wilson, M. (2017). Things I learned from Ben. In M. Wilson & W. P. Fisher, Jr. (Eds.), *Psychological and social measurement: The career and contributions of Benjamin D. Wright* (pp. 75–81). [https://doi.org/10.1007/978-3-319-67304-2\\_8](https://doi.org/10.1007/978-3-319-67304-2_8)
- Wilson, M. (2023). *Constructing measures* (2nd ed.). Routledge. <https://doi.org/10.4324/9781003286929>
- Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.
- Wright, B. D., & Mok, M. M. C. (2004). An overview of the family of Rasch measurement models. *Introduction to Rasch Measurement*, 1(1), 1–24.
- Zhao, N. W., Mason, J. M., Blum, A. M., Kim, E. K., Young, V. N., Rosen, C. A., & Schneider, S. L. (2023). Using item-response theory to

improve interpretation of the trans woman voice questionnaire. *The Laryngoscope*, 133(5), 1197–1204. <https://doi.org/10.1002/lary.30360>