Journal of Mathematical Psychology 55 (2011) 229–239

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Signal detection theory with item effects

Lawrence T. DeCarlo

Department of Human Development, Teachers College, Columbia University, 525 West 120th Street, Box 118, New York, NY 10027, United States

ARTICLE INFO

Article history:
Received 19 June 2010
Received in revised form
13 January 2011
Available online 5 March 2011

Keywords:
Item effects
Random coefficient
Mixture model
Medical imaging
Word recognition

ABSTRACT

Applications of signal detection theory (SDT) often involve presentations of different items on each trial, such as slides in a medical imaging study or words in a memory study. If factors particular to the items themselves, apart from being a signal or noise, affect observers' responses, then 'item effects' are present. One way to model these effects is to use a latent continuous variable as an item 'factor', such as item 'difficulty'. Details of SDT models with item effects are clarified via derivations of their implied conditional means, variances, and covariances. Intra-item correlations are defined and suggested as measures of the magnitude of item effects. The SDT-item models are simple random coefficient models and can be fit with standard software. More general models, such as item models with mixing and/or with random observer effects, are also considered.

1. Signal detection theory with item effects

Applications of signal detection theory (SDT; Green & Swets, 1966) often involve the presentation of different items over trials, such as slides in a medical imaging study or words in a memory study. It has long been recognized that factors particular to the items themselves, other than being a signal or noise, might affect observers' responses (cf. Clark, 1973), and so 'item effects' might be present. Examined here are simple extensions of SDT that allow for item effects (also see Freeman, Heathcote, Chalmers, & Hockley, 2010, Morey, Pratte, & Rouder, 2008, Pratte, Rouder, & Morey, 2010, Rouder & Lu, 2005 and Rouder et al., 2007).

One approach is to use a latent continuous variable (i.e., an item factor) to represent an item effect. Exactly how the latent variable is introduced into the SDT model, however, depends on the conceptualization of the situation. For example, if the item effect is thought of as item 'difficulty', then a latent 'difficulty' variable that affects the discrimination parameter can be introduced; the latent variable affects the discrimination parameter because more difficult items are harder to discriminate. Thus, the latent variable *moderates* the relation between the response and presentation of a signal or noise. Basic SDT models that follow from these and other considerations are derived. Important details of the models are clarified via the derivation of conditional means, variances, and covariances they imply. Measures of the magnitude of item effects, intra-item correlations, are also proposed.

An example is a medical training study where observers attempted to detect fractures in X-ray slides of ankles. The traditional SDT model gives estimates of the observers' ability to discriminate fractures from non-fractures and their use of response criteria. In addition, an item-effect SDT model allows the slides to differ with respect to difficulty. A model where only fractures differ in difficulty is compared to a model where both fractures and non-fractures differ in difficulty. The models are simple random coefficient models and can be fit with maximum likelihood estimation (MLE) using standard software.

The traditional unequal variance SDT model as applied to multiple observers is first briefly reviewed. Next, models that generalize the equal variance SDT model to allow for item effects are introduced; the models include latent variables that represent item 'factors'. It is shown that the presence of item effects leads to non-zero correlations across observers as well as larger variance within observers. The SDT-item models are illustrated with applications to data from a medical imaging study and word recognition studies.

2. Signal detection theory with multiple observers

2.1. Unequal variance SDT

A basic idea in SDT is that observers arrive at a response by using response criteria along with their perceptions. The use of response criteria is one component of SDT, which consists of a *decision rule*, such as

$$Y_j = k \quad \text{if } c_{j,k-1} < \Psi_j \leq c_{jk}, \tag{1}$$

where Y_j is a rating response for the jth observer with discrete values k that range from 1 to K_j, where K_j is the number of response categories (typically the same across observers), Ψ_j is a latent

E-mail addresses: decarlo@tc.edu, decarlo@exchange.tc.columbia.edu.
URL: http://www.columbia.edu/~ld208.

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.01.002

continuous random variable that represents the jth observer's perception, and c_{jk} is the kth criteria for the jth observer, which are strictly ordered, $c_{j1} < c_{j2} < \cdots < c_{j,k-1}$, with $c_{j0} = -\infty$ and $c_{jK} = \infty$. Note that Y_j and Ψ_j are random variables for the jth observer, where the variation is over items (i.e., trials; a subscript i could also be used, but it is implicit here, which simplifies the notation).

A second component of SDT is the *perceptual*, or more generally, the *structural model*, which is concerned with the relation between observers' perceptions and observed (or unobserved) events that give rise to them (DeCarlo, 2010). For example, the structural model for the traditional unequal variance SDT model is,

$$\Psi_j = d_j x + \sigma_j^x \varepsilon_j, \tag{2}$$

where $x = 0$ or 1 for noise or signal, respectively (i.e., the mean of the noise distribution is used as the zero point; note that a subscript j is not needed on x because the items are the same across all observers, but see below), d_j is the distance of the mean of the signal distribution from the mean of noise distribution for the jth observer (scaled with respect to the square root of the error variance), σ_j is a scale parameter (that allows the signal variance to differ from the noise variance), and ε_j is random variation in the jth observer's perception; ε_j is assumed to have a mean of zero, $\mathrm{E}(\varepsilon_j) = 0$, and variance $\mathrm{V}(\varepsilon_j)$, where E is the expectation operator and V is the variance operator; it is also assumed that ε_j is uncorrelated with x. For purposes of identification, $\mathrm{V}(\varepsilon_j)$ for the normal model is set to unity for each observer.

The decision and perceptual models of Eqs. (1) and (2) together give the unequal variance SDT model. In particular, it follows that

$$\begin{aligned} p(Y_j \le k \mid X = x) &= p(\Psi_j \le c_{jk} \mid x) = p(d_j x + \sigma_j^x \varepsilon_j \le c_{jk}) \\ &= p[\varepsilon_j \le (c_{jk} - d_j x)/\sigma_j^x]. \end{aligned}$$

If $\varepsilon_j \sim \mathrm{N}(0, 1)$ then

$$p[\varepsilon_j \le (c_{jk} - d_j x)/\sigma_j^x] = \Phi[(c_{jk} - d_j x)/\sigma_j^x]$$

and so

$$p(Y_j \le k \mid x) = \Phi[(c_{jk} - d_j x)/\sigma_j^x],$$

which is the unequal variance SDT model (e.g., DeCarlo, 2003).

Basic aspects of SDT can be illustrated in terms of conditional means and variances of the latent variable Ψ. For example, it follows from Eq. (2) that the conditional expectations and variances (over items) for observer j are

$$\mathrm{E}(\Psi_j \mid x = 0) = \mathrm{E}(\varepsilon_j) = 0 \qquad \mathrm{V}(\Psi_j \mid x = 0) = \mathrm{V}(\varepsilon_j)$$

$$\mathrm{E}(\Psi_j \mid x = 1) = d_j \qquad \mathrm{V}(\Psi_j \mid x = 1) = \sigma_j^2 \mathrm{V}(\varepsilon_j),$$

where d_j (and the criteria c_{jk}) is scaled with respect to $\mathrm{V}(\varepsilon_j)$ and, as noted above, $\mathrm{V}(\varepsilon_j)$ is set to unity for the normal model. The above shows that, for a given observer j, the signal variance can be less than, equal to, or greater than the noise variance if $\sigma_j < 1$, $\sigma_j = 1$, or $\sigma_j > 1$, respectively. It also follows from Eq. (2) that, conditional on x, Ψ is a linear transformation of ε, and so normal ε implies normal $\Psi \mid x$. A fit of the unequal variance SDT model to individual observer's data gives estimates of d_j and σ_j, along with estimates of the criteria locations c_{jk}.

For the situation examined here, there is a potential effect of items that is common across observers, and so conditional covariances must also be considered. For example, it follows from Eq. (2) that the conditional covariance for observers' j and j' is, for noise,

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 0) = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0,$$

where $\mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0$ follows from the assumption of independence across observers. Similarly, for signal,

$$\begin{aligned} \mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 1) &= \mathrm{Cov}(d_j + \sigma_j \varepsilon_j, d_{j'} + \sigma_{j'} \varepsilon_{j'}) \\ &= \sigma_j \sigma_{j'} \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0, \end{aligned}$$

given that d_j, $d_{j'}$, σ_j, and $\sigma_{j'}$ are constants (over items) for a given observer. The above shows that, conditional on the presence or absence of a signal, perceptions (and responses) are not correlated across observers in the traditional unequal variance SDT model; note that the above represents *error* covariance, that is, covariance across observers that is not accounted for by the presentation of signal or noise. The possible presence of item effects means that the assumption of zero covariance might be violated.

3. SDT models with item effects

Two simple extensions of the equal variance version of SDT are presented, one that includes an item effect for signal alone, and one that includes item effects for both signal and noise. It is shown, using the algebra of variances and covariances, that item effects lead to non-zero correlations across observers and to larger variance within observers.

3.1. Item effects for signals

An *item effect* means that items have one or more characteristics that have a *common effect* across observers, over and above being a signal or noise. The result is a residual correlation across observers, which is what the SDT extensions considered here attempt to account for. For example, in the study analyzed below, observers attempted to detect fractures in X-ray slides of ankles. It seems reasonable to assume that it might be more difficult to detect a fracture in some slides as compared to other slides because of characteristics of the slides besides the simple presence of a fracture (e.g., the type of fracture, the location of the fracture, noise in the slides such as shades, streaks, etc.). One approach is to model the overall effect of these characteristics via a latent continuous variable, such as a 'difficulty' variable. This allows for *heterogeneity* among the slides, that is, fractures in some slides might be more difficult to detect than others. Note that discrimination will be lower for more difficult slides and higher for easier slides, and so the latent variable affects discrimination. Thus, the latent variable *moderates* the relation between a response (i.e., perception) and the presence or absence of a fracture, rather than directly affecting the response (for an example of the latter approach, see Qu & Hadgu, 1998). Because of the common effect of the latent variable, perceptions are correlated across observers, and so the independence assumption made in the traditional SDT model is no longer appropriate.

Consider the situation where an item effect is present for signals (e.g., slides with a fracture). If the effect is basically one of 'difficulty', then a latent continuous variable that represents item difficulty, say φ (which varies over items), can be introduced. Because of the interpretation in terms of difficulty, the latent variable affects discrimination, and so, for an equal variance normal SDT model, the extended structural model is

$$\Psi_j = (d_j - \varphi)x + \varepsilon_j, \tag{3}$$

with $\varphi \sim \mathrm{N}[0, \mathrm{V}(\varphi)]$, $\mathrm{Cov}(\varphi, \varepsilon_j) = 0$, $\mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0$, and $\mathrm{Cov}(\varepsilon_j, x) = 0$. It follows from Eq. (3) that, for larger values of φ (i.e., more difficult items), discrimination is lower.

From Eq. (3), the conditional variances for observer j are

$$\mathrm{V}(\Psi_j \mid x = 0) = \mathrm{V}(\varepsilon_j) \qquad \mathrm{V}(\Psi_j \mid x = 1) = \mathrm{V}(\varphi) + \mathrm{V}(\varepsilon_j).$$

With the assumption $\mathrm{V}(\varepsilon_j) = 1$, the conditional variances for the (normal) model are 1 for noise and $\mathrm{V}(\varphi) + 1$ for signal. Thus, if there is an item effect for signals, the signal variance will be larger than the noise variance (for each observer). The above shows that the presence of an item effect in an equal variance SDT model provides a possible theoretical reason as to *why* the signal variance is larger than the noise variance within observers—because of an item effect.

The presence of an item effect also means that perceptions (and responses) are correlated across observers. In particular, the conditional covariances of Ψ_j for observers' j and j' are,

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 0) = \mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0,$$

for noise and

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 1) = \mathrm{Cov}(d_j - \varphi + \varepsilon_j, d_{j'} - \varphi + \varepsilon_{j'}) = \mathrm{V}(\varphi),$$

for signal. It follows that, if $\mathrm{V}(\varphi)$ is greater than zero, perceptions are correlated across observers.

A conditional *intra-item correlation* (IIC), in this case for signal, can be defined as

$$\mathrm{Corr}(\Psi_j, \Psi_{j'} | x = 1) = \frac{\mathrm{Cov}(\Psi_j, \Psi_{j'} | x = 1)}{\sqrt{\mathrm{V}(\Psi_j | x = 1)\mathrm{V}(\Psi_{j'} | x = 1)}} = \frac{\mathrm{V}(\varphi)}{\mathrm{V}(\varphi) + \mathrm{V}(\varepsilon_j)} \quad (4)$$

where the second line follows from the conditional variances and covariances derived above. The intra-item correlation is the correlation of perceptions for two observers judging the same item (for signal in this case); it is suggested here as a measure of the magnitude of item effects. Eq. (4) shows that the intra-item correlation also has an interpretation as the proportion of signal variance for observer j, $\mathrm{V}(\varphi) + \mathrm{V}(\varepsilon_j)$, that is due to the item effect, $\mathrm{V}(\varphi)$.

With the decision rule of Eq. (1), the resulting SDT model is

$$p(Y_j \le k \mid x, \varphi) = \Phi[c_{jk} - (d_j - \varphi)x], \quad (5)$$

which is a *random coefficient* model, in that the coefficient of x is not a constant (d_j) for each observer, but rather includes a random component, the item effect φ, that varies over items (trials). The model is a random coefficient model, as discussed in DeCarlo (2010, Eq. (9)), but differs in that, for item models, a *common* latent variable (φ) is assumed to have an effect across observers, whereas a *different* latent variable would appear for each observer in Eq. (9) of DeCarlo as applied to multiple observers (i.e., the random coefficient would be γ_j for observer j).

Eq. (5) is a probit model with a random slope (when written using an inverse link, see DeCarlo, 1998) and can be fit, using maximum likelihood estimation (MLE), with software for latent class modeling, such as Latent Gold (Vermunt & Magidson, 2007). The model is illustrated with data from X-ray and word recognition studies.

3.2. Item effects for signal and noise

Eq. (5) introduces an item effect for signals, such as slides that show a fracture. It is also possible, however, that there are item effects for noise. For example, for the X-ray study considered below, it might be more difficult to determine if an ankle is normal (no fracture) in some slides as compared to others, because of various characteristics of the slides, such as subtle variations in light, shade, angle, and other "noise".

A model that allows for item effects for both signal and noise extends the structural model of Eq. (3) as follows,

$$\Psi_j = (d_j - \varphi_s)x + \varphi_n(1 - x) + \varepsilon_j, \quad (6)$$

where $\varphi_s \sim \mathrm{N}[0, \mathrm{V}(\varphi_s)]$, $\varphi_n \sim \mathrm{N}[0, \mathrm{V}(\varphi_n)]$, $\mathrm{Cov}(\varphi_s, \varepsilon_j) = 0$, $\mathrm{Cov}(\varphi_n, \varepsilon_j) = 0$, $\mathrm{Cov}(\varphi_s, \varphi_n) = 0$, $\mathrm{Cov}(\varepsilon_j, \varepsilon_{j'}) = 0$, and $\mathrm{Cov}(\varepsilon_j, x) = 0$.[1] It follows from Eq. (6) that the perceptual distribution is at $d_j - \varphi_s$ when $x = 1$ (signal) and is at φ_n when $x = 0$ (noise); note that higher values of φ_n indicate more difficult normal items, in that shifting Ψ_j to the right (for noise) results in lower confidence that an item is normal (noise).

It follows from Eq. (6) that,

$$\begin{aligned} &\mathrm{V}(\Psi_j \mid x = 0) = \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j) \qquad \mathrm{V}(\Psi_j \mid x = 1) = \mathrm{V}(\varphi_s) + \mathrm{V}(\varepsilon_j) \\ &\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 0) = \mathrm{V}(\varphi_n) \qquad \mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 1) = \mathrm{V}(\varphi_s). \end{aligned} \quad (7)$$

In this case, the covariance can be non-zero for either signal or noise, and so one can obtain intra-item correlations (Eq. (4)) both for signal (IIC_s) and for noise (IIC_n). Eq. (7) also shows that, within observers, the signal variance can be greater than, equal to, or less than that of noise, depending on the relative magnitudes of $\mathrm{V}(\varphi_s)$ and $\mathrm{V}(\varphi_n)$. It is also important to note that the model places a constraint on the conditional variances and covariances, in that a larger covariance between observers implies that the variance within observers is also larger.

Using the decision rule of Eq. (1) together with Eq. (6) gives,

$$\begin{aligned} p(Y_j \le k \mid x, \varphi_s, \varphi_n) &= \Phi[c_{jk} - (d_j - \varphi_s)x - \varphi_n(1 - x)] \\ &= \Phi[c_{jk} - \varphi_n - (d_j - \varphi_s - \varphi_n)x], \end{aligned} \quad (8)$$

where the latter form shows that the model (written with an inverse link) has a random intercept and slope. A fit of Eq. (8) provides estimates of d_j, c_{jk}, $\mathrm{V}(\varphi_s)$, and $\mathrm{V}(\varphi_n)$. Eqs. (5) and (8) will together be referred to as *SDT-item models*. Note that the model with item effects for both signal and noise is closely related to that used by Pratte et al. (2010; also see Morey et al. 2008), except that Pratte et al.'s model is parameterized slightly differently and also includes random observer effects, whereas observer-specific fixed effects are used here (i.e., the usual d_j for each observer), and so no distributional assumptions are made about d_j across observers, as in traditional SDT. Some comments on models with random observer effects are made below.

3.3. An empirical extension

A simple yet useful model-based way to obtain information about the variances and covariances (with constraints) is to generalize Eq. (8) as follows,

$$p(Y_j \le k \mid x, \varphi_s, \varphi_n) = \Phi[c_{jk} - (d_j - a_j\varphi_s)x - b_j\varphi_n(1 - x)] \quad (9)$$

with $\varphi_s \sim \mathrm{N}(0, 1)$ and $\varphi_n \sim \mathrm{N}(0, 1)$. Note that the coefficients a_j and b_j can be positive or negative, and so φ_s and φ_n might no longer have simple interpretations as common item effects (i.e., the effects can be in different directions for different observers). It follows from the structural model associated with Eq. (9) that, for observers' j and j',

$$\begin{aligned} &\mathrm{V}(\Psi_j \mid x = 0) = b_j^2 + 1 \qquad \mathrm{V}(\Psi_j \mid x = 1) = a_j^2 + 1, \\ &\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 0) = b_j b_{j'} \qquad \mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x = 1) = a_j a_{j'}. \end{aligned} \quad (10)$$

Eq. (10) shows that the coefficients a_j and b_j are scale parameters that allow the variances and covariances to differ across the J observers; the coefficients also allow for positive or negative covariances, depending on the signs of a_j and b_j; these aspects are what make the model useful for exploratory purposes.

Eq. (9) is offered here as a useful empirical model, in that if all or nearly all of the estimates of a_j or b_j have the same sign (and similar magnitudes),[2] then it follows from Eq. (10) that the covariances are

[1] One could also use a single random variable φ in place of φ_s and φ_n to allow for a single common item effect. For the data considered here, both BIC and AIC were smaller for the two factor model as compared to the single factor model in all cases (i.e., the two factor model had relatively better fit), and so the focus is on the two factor model.

[2] One should be aware that, over repeated runs, the signs of the coefficients can reverse from positive to negative because of 'label-switching' (see DeCarlo, 2008). All that matters in the exploratory analysis suggested here is whether or not the J estimates of a_j (or b_j), generally have the same sign across observers (i.e., either positive or negative).

positive, and so a model with item effects might be useful, given that the model predicts a positive covariance across observers. On the other hand, if there is a mix of positive and negative estimates (with some large values), then a simple item-effect model will likely not suffice because items no longer have a simple common effect across observers. Thus, Eq. (9) provides useful information about the structure of the data.

The next section illustrates the use of the above models in a situation where it was expected that there might be non-trivial item effects: the detection of fractures in X-ray slides of ankles. Item effects were expected because it seems likely that fractures might be more difficult to detect in some slides as compared to others, because of the type of fracture and characteristics of the particular image. In addition, slides without fractures (normals) might also differ with respect to how difficult it is to detect a normal ankle. Thus, models with item effects for both signal and noise are considered. The models are compared to the traditional equal and unequal variance SDT models.

4. Detecting fractures in ankle X-rays

The study consisted of 48 observers who examined 234 X-rays of ankles that were presented on a computer screen (with three different views). Of the 234 slides, 88 showed ankles with a fracture. For each slide, an observer made a decision as to whether a fracture was present or absent on a four point scale, 1 = definitely normal, 2 = probably normal, 3 = probably abnormal, and 4 = definitely abnormal. Further details can be found in Boutis, Pecaric, Seeto, and Pusic (2010).

4.1. Results

Although each slide was presented in a different (random) order for each observer, the data should be sorted, for the analysis, so that the observers' responses are matched with respect to the presented item (i.e., for the data in multivariate form, with observers as columns and items as rows). It is useful to start with Eq. (9) to explore the data. For the X-ray data, the estimates of b_j (normal slides) ranged from 0.19 to 1.07 with a mean of 0.63 and standard deviation of 0.22, whereas a_j (fracture slides) ranged from 0.38 to 1.15 with a mean of 0.81 and standard deviation of 0.20. Thus, all of the coefficients were positive (and significant), which suggests the presence of item effects for both signal (fracture) and noise (normal). Note that the mean magnitudes of $a_j(0.81)$ and $b_j(0.63)$ also suggest that the correlation is larger (on average) for fractures than for normals.

4.2. Relative fit

The upper part of Table 1 shows relative fit statistics, BIC and AIC, for the equal variance SDT model, the unequal variance SDT model, and the two item-effect models (Eqs. (5) and (8)). The table shows that the relative fit of the item-effect models is better than either the equal or unequal variance SDT models, in that both BIC and AIC are considerably smaller, and in particular, BIC and AIC are smallest for the SDT model with item effects for both signal and noise (Eq. (8)). The results indicate that the item models, which account for non-zero covariances across observers, offer an improvement over simply fitting the equal or unequal variance SDT model to the data of each observer.

4.3. Parameter estimates

Unequal variance SDT model. The left panel of Fig. 1 presents a plot of estimates of d_j for a fit of the unequal variance SDT model to the data of each of the 48 observers. The estimates range from 0.49 to 2.63 with a mean of 1.45 (standard deviation of 0.46) and are approximately normally distributed. The lower panel shows estimates of the signal standard deviation, σ_j, which range from 0.62 to 1.62, with a mean of 1.08, and so the ratio of fracture to normal standard deviations is 1.08 (on average), which indicates (slightly) larger variance for the signal distribution (fractures) than noise (normals).

SDT-item model. The right panel of Fig. 1 presents a plot of estimates of d_j for a fit of Eq. (8). The estimates range from 0.38 to 2.76 with a mean of 1.61, a standard deviation of 0.64, and are approximately normally distributed. Estimates of $V(\varphi_s)$ for fractures (signal) and $V(\varphi_n)$ for normals (noise) are 0.60 (standard error of 0.04) and 0.38 (0.04), respectively. It follows from Eq. (7) that the (predicted) ratio of fracture to normal standard deviations is $\sqrt{(0.60+1)}/\sqrt{(0.38+1)} = 1.08$, which is the same as the average estimate of σ_j found for fits of the unequal variance SDT model above. Thus, with respect to the variance, the unequal variance and item SDT models both lead to the same conclusion—for each observer, the perceptual distribution for fractures has larger variance than that for normals; of course, the SDT-item model also accounts for the correlation between observers, whereas the unequal variance SDT model does not.

Estimates of the intra-item correlations (Eq. (4)), from the variance estimates, are $0.38/(0.38+1) = 0.275$ for normal slides and $0.60/(0.60+1) = 0.375$ for fractures. Thus, fairly large correlations are found, which indicate the presence of item effects, as expected. In addition, the intra-item correlation for slides with fractures is larger than that for slides without fractures, and so the item effects appear to be larger for fractures. In this case, the item effects are consistent with both non-zero correlations across observers and larger variance (for fractures) within observers.

Note that one can also obtain 'estimates' of φ_s and φ_n (they are more accurately described as 'predictions'), just as one can get factor scores in latent variable modeling (e.g., Skrondal & Rabe-Hesketh, 2004); these are provided, for example, by Latent Gold (i.e., empirical Bayes estimates) and other software packages. For the X-ray study, one goal was to examine the utility of the item scores (i.e., the item difficulties) in future training studies with the same items (by, for example, using the factor score estimates to select difficult or easy slides); this can provide additional evidence for or against the validity of the item model and the utility of the scores, which is an important next step.

Table 1
Information criteria for traditional SDT and item-effect SDT models.

Model	Ankle X-rays (48 observers)		
	#par	BIC	AIC
Equal variance SDT[a]	191	23030.6	22367.2
Unequal variance SDT	239	23150.8	22321.5
SDT-item (Eq. (5))	192	21491.9	20828.4
SDT-item (Eq. (8))	193	20167.2	19500.4
Model	**Word recognition (21 observers)**		
	#par	BIC	AIC
Equal variance SDT	105	6800.2	6507.5
Unequal variance SDT	126	6883.0	6481.8
SDT-item (Eq. (5))	106	6777.8	6482.4
SDT-item (Eq. (8))	107	6690.0	6391.7
Model	**Word recognition (97 observers)**		
	#par	BIC	AIC
Equal variance SDT	579	139148.8	136732.2
Unequal variance SDT	676	138447.9	135831.7
SDT-item (Eq. (5))	580	137855.2	135434.4
SDT-item (Eq. (8))	581	135951.3	133526.3
SDT-item (Eq. (8), correlated φ)	582	135937.0	133507.9

[a] Note: One observer used only 3 of the 4 response categories.

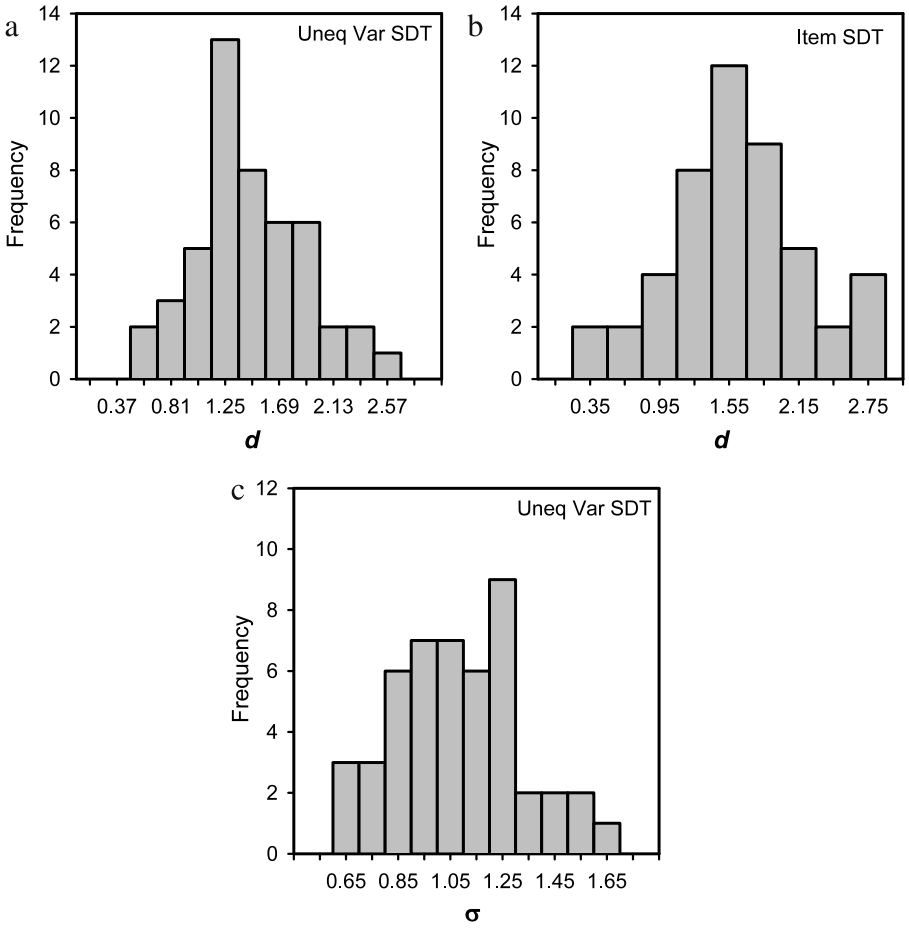

Fig. 1. The left panel shows, for the unequal variance SDT model, the distribution of estimates of d_j obtained for 48 observers in a medical imaging study. The lower panel shows the distribution of estimates of σ_j for the same observers. The right panel shows, for the SDT-item model, the distribution of estimates of d_j.

5. Word recognition

This section applies the SDT-item models to recognition memory studies with words. In the typical word recognition study, a list of words is presented to observers for a study period, followed by a test where observers rate their confidence as to whether a presented word is old (on the study list) or new (not on the study list). Some words might have characteristics that make them more difficult (or easier) to remember, and so there might be item effects for words. This can again be modeled by introducing a latent 'difficulty' variable that affects discrimination, as done above. New words might also differ in difficulty, which leads to an SDT-item model with effects for both signal and noise. Given that words used in word recognition studies are typically selected so as to have little variation on other dimensions (e.g., word length, word frequency, etc.), it was expected that item effects might be smaller than those found for the medical imaging study, and so the word studies offer an interesting comparison.

In the first word recognition study, the list of study words was the same across all observers, which is directly analogous to the medical imaging study (i.e., particular slides cannot be fractures for some observers and normal for others). In the second study, whether a word was old or new differed across observers, that is, a particular word was old for some observers and new for others; consequences of this are discussed below.

5.1. Word recognition: same study words across observers

The experiment was conducted as part of a laboratory course (DeCarlo, 1997). Twenty one undergraduate students were run individually. The participants were first shown 60 words in a study condition where all 60 words were presented together on a computer screen (10 rows of 6 words each) for 2 min. A test followed immediately with the 60 study words presented one at a time in a random order along with 60 new words; participants rated their confidence that a word was old or new on a five point scale. The Appendix gives a Latent Gold program to fit the SDT-item model of Eq. (8); the data are available at the author's website.

5.2. Results

It is again useful to begin with Eq. (9) to obtain some basic information. For new words, the estimates of b_j ranged from 0.04 to 0.78 with a mean of 0.46 and standard deviation of 0.22. Thus, all the estimates are positive, which suggests positive correlations, and so item effects might be present for new words. For old words, the estimates of a_j ranged from -0.72 to 0.73 (with significant positive and negative values) with a mean of 0.08 and standard deviation of 0.35. Thus, the presence of simple item effects for old words is not apparent (given that the mean effect is near zero); in fact, the large positive and negative estimates of a_j for old words suggests, from Eq. (10), that the mean covariance (of which there are positive and negative values) might be near zero, whereas the variance, $a_j^2 + 1$, can be large, which is predictive of results found below.

5.3. Relative fit

The middle panel of Table 1 shows information criteria for fits of the equal variance SDT model, the unequal variance SDT model,

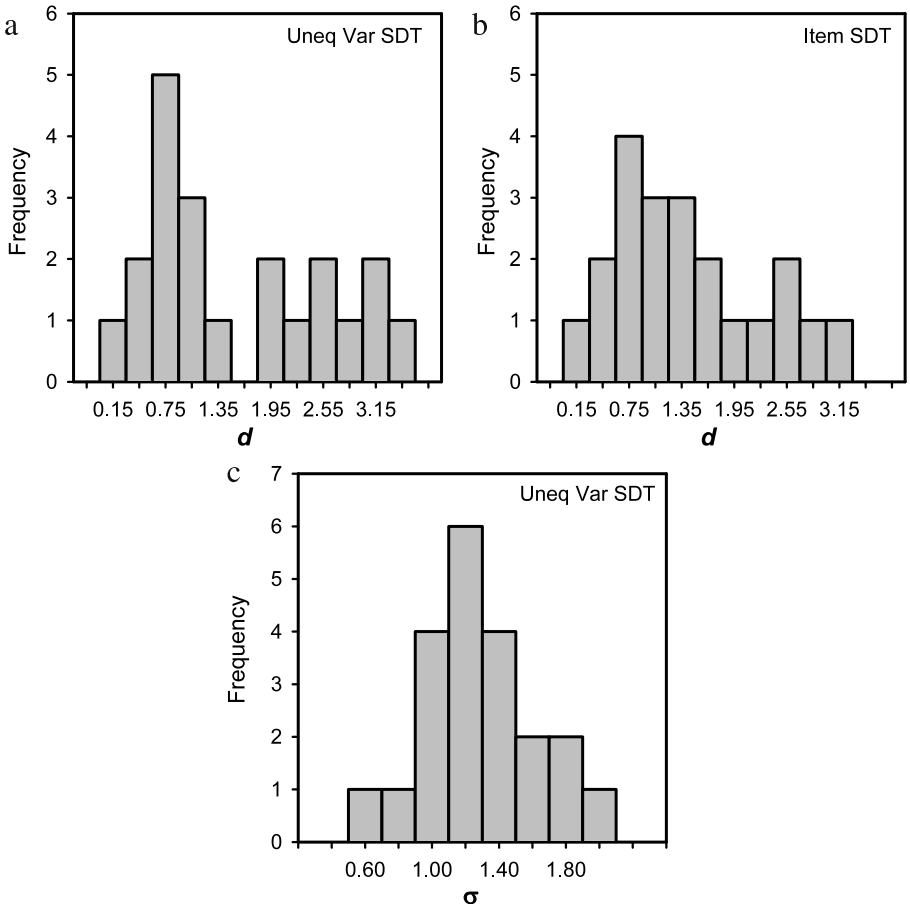

Fig. 2. The left panel shows, for the unequal variance SDT model, the distribution of estimates of d obtained for 21 observers in a word recognition study. The lower panel shows the distribution of estimates of σ_j for the same observers. The right panel shows, for the SDT-item model, the distribution of estimates of d_j.

and the item-effect models. Both BIC and AIC indicate that the item-effect SDT model of Eq. (8) provides the best relative fit, which again suggests the need to account for non-zero covariances across observers.

5.4. Parameter estimates

Unequal variance SDT. The left panel of Fig. 2 shows a plot of estimates of d_j obtained for a fit of the unequal variance SDT model to the data of each observer. The estimates range from 0.26 to 3.37 with a mean of 1.59 and standard deviation of 1.03. The number of observers is fairly small (21), however the plot suggests two clusters, observers with lower values of d (around 0.75) and those with higher values (1.9 and higher), perhaps reflecting motivated and non-motivated observers; note that the fixed effects approach does not make any distributional assumptions about d_j across observers. The lower panel of Fig. 2 shows a plot of estimates of σ_j, which range from 0.52 to 1.96 with a mean of 1.28, which reflects the usual finding of larger variance (on average) for signal (old words) than noise (new words).

SDT-item model. The right panel of Fig. 2 shows a plot of estimates of d_j obtained for a fit of the item model of Eq. (8). The estimates range from 0.29 to 3.20 with a mean of 1.42 and standard deviation of 0.83. The figure suggests two distributions of estimates, for lower and higher d, as also found for fits of the unequal variance SDT model (although there appears to be some slight smoothing).

Estimates of the variances (and standard errors) of φ_s and φ_n are 0.02 (0.02) and 0.19 (0.05) for old and new words, respectively. Note that this implies that the ratio of signal to noise standard deviations is $\sqrt{(0.02+1)}/\sqrt{(0.19+1)} = 0.93$ (from Eq. (7)), and so the results differ from those found for the unequal variance SDT model in that the item-effect model suggests that the variance for signal is slightly smaller than that for noise (ratio of standard deviations of 0.93), whereas the unequal variance SDT model gave larger variance for signal (average ratio of 1.28). This suggests the influence of other factors, as discussed below.

The estimate of the intra-item correlations are $0.19/(0.19+1) = 0.16$ for new words and $0.02/(0.02+1) = 0.02$ for old words. Thus, for new words, there appears to be an item effect, but the magnitude of the effect, 0.16, is considerably smaller than that found in the medical imaging study above. For old words, the near zero intra-item correlation suggests the absence of a simple item effect (as also suggested by the exploratory analysis presented above).

The consistency of the results for word recognition is next examined using data from another (published) word recognition study. It is also shown that the particular experimental design that is used determines whether or not item-effect correlations can be estimated.

5.5. Word recognition: different study words across observers

The example is a word recognition study by Pratte et al. (2010), details of which can be found in their article. 240 study words were presented to 97 observers. The test consisted of the 240 studied words and 240 new words. An interesting aspect of the study is that, unlike the medical imaging or word recognition examples presented above, the list of old words differed across observers, in that the 240 study words were randomly selected for each observer

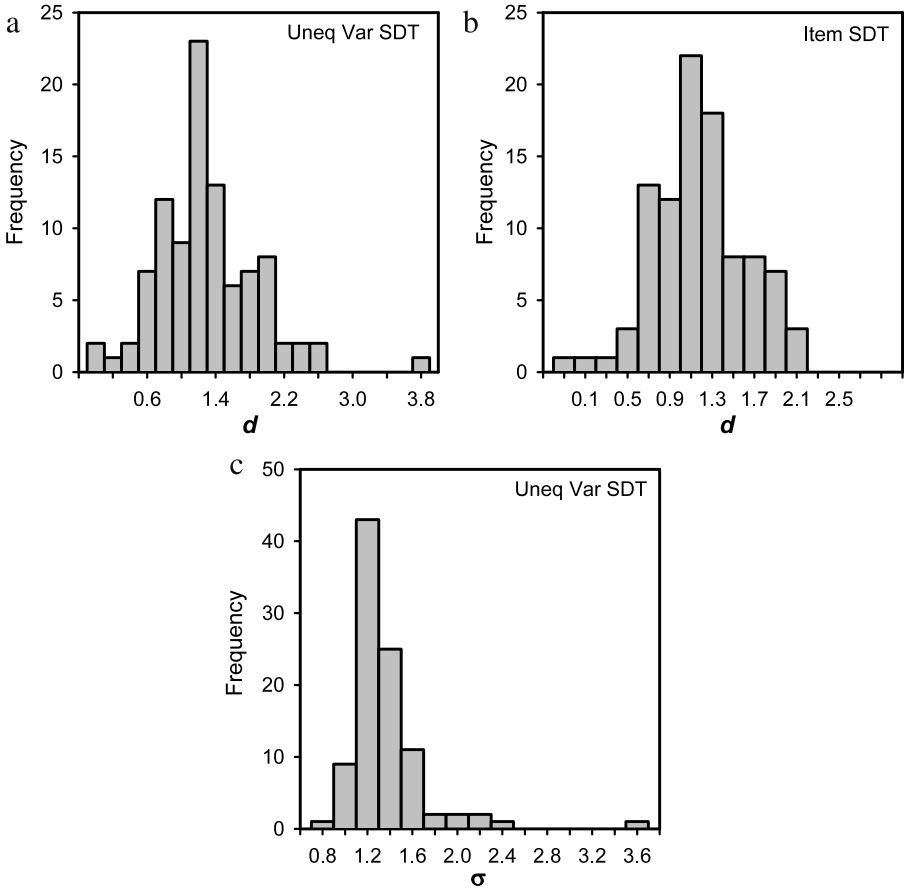

Fig. 3. The left panel shows, for the unequal variance SDT model, the distribution of estimates of d obtained for 97 observers in a word recognition study. The lower panel shows the distribution of estimates of σ_j for the same observers. The right panel shows, for the SDT-item model, the distribution of estimates of d_j.

from the pool of 480 words. Thus, whether a particular word was old or new word differed across the observers (i.e., each word was old for some observers and new for others), in contrast to the above examples.[3] In this case, there is information about a possible correlation between item effects.

The design requires that a subscript j be added to x because whether a particular item was old or new can now differ across the J observers. The conditional covariances are

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x_j = 0, x'_j = 0) = \mathrm{V}(\varphi_n)$$
$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x_j = 1, x_{j'} = 0) = \mathrm{Cov}(\varphi_s, \varphi_n)$$
$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x_j = 1, x_{j'} = 1) = \mathrm{V}(\varphi_s)$$
$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x_j = 0, x_{j'} = 1) = \mathrm{Cov}(\varphi_n, \varphi_s).$$

Note that when the items are both new for a given pair of observers, $x_j = 0$ and $x_{j'} = 0$, or old, $x_j = 1$ and $x_{j'} = 1$, then the covariances are the same as in Eq. (7), that is, $\mathrm{V}(\varphi_n)$ and $\mathrm{V}(\varphi_s)$. However, when the item types (new or old word) differ across pairs of observers (which does not happen in the earlier examples), then the conditional covariances of Ψ depend on $\mathrm{Cov}(\varphi_s, \varphi_n)$, as shown above, and so there is information about item-effect covariances.

5.6. Results

It is again useful to start with Eq. (9) to explore the data. For new words, estimates of b_j ranged from 0.06 to 0.71 with a mean of 0.38 and standard deviation of 0.14. The positive values suggest the presence of a common item effect for new words. For old words, estimates of a_j range from −0.04 to 0.65 with a mean of 0.34 and a standard deviation of 0.16; there were only two negative values, neither of which was significant. The results indicate that there are positive correlations across observers and suggest the presence of item effects for both old and new words.

5.7. Relative fit

The bottom section of Table 1 shows information criteria for fits of the equal variance SDT model, the unequal variance SDT model, and the two item-effect SDT models. The table shows that the SDT-item models again provide better relative fits than either the equal or unequal variance SDT models. BIC and AIC are both smallest for the SDT-item model of Eq. (8), which has item effects for both signal and noise.

5.8. Parameter estimates

Unequal variance SDT. The left panel of Fig. 3 presents a plot of estimates of d_j for the 97 observers for the unequal variance SDT model. The range is from 0.07 to 3.79 with a mean of 1.30 and standard deviation of 0.59; the distribution is approximately normal, except for one extreme value of 3.79. The lower panel of Fig. 3 shows a plot of estimates of σ_j for the 97 observers; the estimates range from 0.81 to 3.61 with a mean of 1.36, and so the signal variance is larger than the noise variance (note that Pratte et al., 2010, obtained an estimate of σ of 1.36 for a fit of their model, see their Fig. 2). Note that the observer with the largest estimate of d_j (3.79) also had the largest estimate of σ_j (3.61).

[3] Note that the design is more complex; for example, it differs from the design where the set of old and new items is the same across all observers in that different subsets of observers (rather than all observers) judge each item as old or new.

SDT-item model. The right panel of Fig. 3 shows a plot of estimates of d_j for the item model of Eq. (8). The estimates of d_j range from −0.01 to 2.12 with a mean of 1.18 and standard deviation of 0.43; the distribution is approximately normal. In this case, the observer who appeared above to be an 'outlier' for the unequal variance SDT model now has an estimate of d_j of 2.12 (and so the estimates are smoothed by the item model). The estimate of the variance of φ_n (new words) is 0.16 (0.01) whereas that for φ_s (old words) is 0.10 (0.01); it follows that the ratio of old to new word standard deviations is predicted to be $\sqrt{(0.10+1)}/\sqrt{(0.16+1)} = 0.97$. Thus, as for the word study examined above, the SDT-item model suggests that old words have equal or smaller variance than new words, whereas the unequal variance SDT model suggests that old words have larger variance.

The estimate of the intra-item correlations (using Eq. (4)) are $0.16/(0.16+1) = 0.14$ for new words and $0.10/(0.10+1) = 0.09$ for old words. Thus, the correlation for new words is larger than that for old words, as was also found in the first word study. The intra-item correlations are again small, 0.14 or less. Finally, the estimate of the covariance of φ_s and φ_n is −0.03 (0.01) which, together with the variance estimates of φ_s and φ_n, gives an estimate of −0.03 for the correlation of new and old item effects.

Note that the finding of larger variance for φ_n than φ_s found here is consistent with results reported by Pratte et al. (2010), in that their estimates of the item-effect standard deviations, $\sigma_{\beta,n}$ and $\sigma_{\beta,s}$ in their notation, were larger for new words than for old words (see their Fig. 3(C)) and similar in value to those found here. Freeman et al. (2010) also recently presented results where the item-effect variance was larger for new words than for old words (see their Fig. 5, bottom row). Thus, using an SDT model with random observer and item effects (and estimation via Markov chain Monte Carlo methodology), as in Pratte et al.'s and Freeman et al.'s approach, gives the same results as the item models presented here, in that the intra-item correlations are larger for new words than for old words.

6. Discussion

The SDT-item models presented here are simple extensions of the conventional equal variance SDT model that allow for item effects. The models offer an improvement in relative fit over the unequal variance SDT model (fit to the data of each observer) and account for correlations between observers over and above that due to signal or noise (i.e., residual correlations). A fit of the models allows one to estimate intra-item correlations, which were fairly large in the medical imaging study, in the range of 0.28–0.38, and were smaller in the word recognition studies, 0.16 or less. This was expected to some extent because words in recognition memory studies are usually selected so as to not vary too much in terms of various characteristics, such as word length, word frequency, and so on. Thus, the intra-item correlations are informative about the presence and magnitude of item effects. A next step in future research is to attempt to gain some experimental control over the item effects, by for example manipulating a variable that might increase the effects and examining the resulting intra-item correlations.

An interesting finding for the word recognition studies is that the intra-item correlation appears to be larger for new words than old words. This result has not been noted before, to my knowledge, although it appears in other recent studies: the item variances (and so the covariances) were larger for new items than old items in both the studies of Freeman et al. (2010) and Pratte et al. (2010), even though (slightly) different models and estimation methods were used. Thus, this result appears across both mixed and random versions of the models and across different estimation methods; whether or not this is generally the case requires further investigation. Of course, one has to use the models to discover their advantages and limitations; the ability to easily fit the basic SDT-item models with standard software and MLE should motivate researchers to use them in their research, in addition to the traditional unequal variance SDT model. In any case, the larger intra-item correlation for new words is noted here as a result of theoretical interest for researchers concerned with word recognition and item effects.

It is shown that the conditional variances and covariances associated with a particular model clarify important details. For example, a fairly strong constraint that follows from the item models is that larger covariance (between observers) should be accompanied by larger variance (within observers, see Eq. (7)). Put simply, item effects not only lead to correlations between observers, but also add to the variance within observers. The results for the medical imaging study were consistent with this relation, in that fractures showed both larger intra-item correlations and larger variance. However, the results for the word recognition studies were not, in that new words showed larger intra-item correlations, and so they should also have larger variance, yet the unequal variance model suggested smaller or equal variance for new words. This suggests that, although the item models are useful, they are not completely satisfactory, at least for word recognition. Potential extensions are noted next.

7. Extended item models

The results for word recognition suggest that an SDT-item model that frees the relation (shown in Eq. (7)) between the conditional variances and covariances might be needed in some cases. Different ways in which this can be done are noted here, as are other extensions, such as models with random observer effects.

7.1. Larger signal variance

To start, note that the model used by Pratte et al. (2010) frees the relation between the conditional variances and covariances by including a variance parameter, that is, by extending the unequal variance SDT model rather than the equal variance model. An unequal variance extension of the structural model of Eq. (6) is,

$$\Psi_j = (d_j - \varphi_s)x + \varphi_n(1-x) + \sigma^x \varepsilon_j.$$

It follows that the conditional variances and covariances are

$$\mathrm{V}(\Psi_j \mid x=0) = \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j)$$
$$\mathrm{V}(\Psi_j \mid x=1) = \mathrm{V}(\varphi_s) + \sigma^2\,\mathrm{V}(\varepsilon_j)$$
$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=0) = \mathrm{V}(\varphi_n) \qquad \mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=1) = \mathrm{V}(\varphi_s).$$

The term $\sigma^2\mathrm{V}(\varepsilon_j)$ in the above shows that the signal variance can be inflated or deflated by σ^2. This means that, even if $\mathrm{V}(\varphi_s) < \mathrm{V}(\varphi_n)$, as in the word recognition studies above, the signal variance can still be larger than noise if $\sigma^2 > 1$. Note that, using Pratte et al.'s notation, $\mathrm{V}(\varphi_s) = \sigma^2_{\beta,s}$ and $\mathrm{V}(\varphi_n) = \sigma^2_{\beta,n}$, estimates of which are about $0.40^2 = 0.16$ for old words and $0.45^2 = 0.20$ for new words (from their Fig. 3(C)). Using the conditional variances and covariances shown above, estimates of the intra-item correlations are approximately $0.20/(1+0.20) = 0.17$ for new words and, using Pratte et al.'s estimate of σ of 1.36, $0.16/(0.16+1.36^2) = 0.08$ for old words. Thus, for the unequal variance version of the item model, the intra-item correlation is again larger for new words than old words (and the estimates of 0.17 and 0.08 are similar to those found above, 0.14 and 0.09), and so this result appears to be consistent.

A related way to extend the model is to include an additional random coefficient γ as follows,

$$\Psi_j = (d_j - \varphi_s + \gamma)x + \varphi_n(1-x) + \varepsilon_j,$$

where $\varphi_s \sim \mathrm{N}[0, \mathrm{V}(\varphi_s)]$, $\varphi_n \sim \mathrm{N}[0, \mathrm{V}(\varphi_n)]$, $\gamma \sim \mathrm{N}[0, \mathrm{V}(\gamma)]$, $\mathrm{Cov}(\gamma, \varphi_s) = 0$, $\mathrm{Cov}(\gamma, \varphi_n) = 0$, and ε_j is independent of the other terms. A convenient aspect of the resulting model is that it is simple to implement in standard software (i.e., it is a random coefficient probit model); for details of the relation of this approach to the unequal variance extension, see DeCarlo (2010). It follows from the above structural model that

$$\mathrm{V}(\Psi_j \mid x=0) = \mathrm{V}(\varphi_n) + 1$$
$$\mathrm{V}(\Psi_j \mid x=1) = \mathrm{V}(\varphi_s) + \mathrm{V}(\gamma) + 1,$$
$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=0) = \mathrm{V}(\varphi_n) \qquad \mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=1) = \mathrm{V}(\varphi_s).$$

The above shows that the term $\mathrm{V}(\gamma)$ in the conditional variance for signal frees up the relation between the covariance and variance, as in Pratte et al.'s (2010) model, and so the random coefficient extension offers a closely related model (with additive instead of multiplicative effects).

A theoretical model. The above offer potentially useful empirical models that can capture additional aspects of the data (i.e., smaller covariance yet larger variance). Empirical models, however, do not provide a reason as to *why* the variance is larger, they simply allow for it. A mixture SDT model (DeCarlo, 2002), on the other hand, offers a theoretical motivation for the larger variance. The basic idea is that observers do not attend to some of the study words, and so old words consist of a mixture of attended and non-attended words, which is what leads to the apparent larger variance for old words (see DeCarlo, 2010). Some steps towards combining mixture SDT models with item models are noted here.

When formulating a mixture SDT-item model, it is important to pay attention to the underlying conceptualization. In particular, if some of the old items are not attended to, then it seems that the item effect for those items should be the same as for new items, that is, the item effect for unattended old items should be φ_n (and not φ_s). The structural model of Eq. (6) can be extended, recognizing this restriction, as follows

$$\Psi_j = [\delta_j(d_j - \varphi_s) + (1-\delta_j)\varphi_n]x + \varphi_n(1-x) + \varepsilon_j, \tag{11}$$

where δ_j is a latent dichotomous (zero/one) variable that indicates, for observer j, attention ($\delta_j = 1$) or a lack of attention ($\delta_j = 0$) on a given trial (see DeCarlo, 2010). Note that the model for new items is the same as in Eq. (6), however for old items, the distribution is located at $(d_j - \varphi_s)$ for attended old items (as in Eq. (6)), but is at φ_n for non-attended old items. The resulting model is

$$\begin{aligned} &p(Y_j \le k \mid x, \varphi_s, \varphi_n) \\ &\quad = \Phi\{c_{jk} - [\delta_j(d_j - \varphi_s) + (1-\delta_j)\varphi_n]x - \varphi_n(1-x)\} \\ &\quad = \Phi[c_{jk} - \varphi_n - \delta_j(d_j - \varphi_s - \varphi_n)x], \end{aligned}$$

which is a simple extension of the item model of Eq. (8).

With the assumptions that $\varphi_s \sim \mathrm{N}[0, \mathrm{V}(\varphi_s)]$, $\varphi_n \sim \mathrm{N}[0, \mathrm{V}(\varphi_n)]$, $\delta_j \sim$ Bernoulli(λ_j), and δ_j and ε_j are independent of each other and the other terms, it follows from Eq. (11) that the conditional covariance for noise is the same as in Eq. (7),

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=0) = \mathrm{V}(\varphi_n),$$

whereas (for uncorrelated item effects)

$$\mathrm{Cov}(\Psi_j, \Psi_{j'} \mid x=1) = \lambda_j\lambda_{j'}\mathrm{V}(\varphi_s) + (1-\lambda_j)(1-\lambda_{j'})\mathrm{V}(\varphi_n).$$

The variance for noise is also the same as in Eq. (7),

$$\mathrm{V}(\Psi_j \mid x=0) = \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j).$$

The variance for signal, however, is a bit trickier. Conditional on *both* x and δ (i.e., $x = 1, \delta = 1$), the signal variance is simply the same as in Eq. (7), that is, $\mathrm{V}(\varphi_s) + \mathrm{V}(\varepsilon_j)$, but this is not the case conditional on x alone. Note that simply taking the conditional variance of Ψ_j as given by Eq. (11) gives a complex expression for $\mathrm{V}(\Psi_j \mid x)$. A simpler approach is to use the following decomposition of the variance: for two variables Y and X with a joint distribution,

$$\mathrm{V}(Y) = \mathrm{E}_X[\mathrm{V}(Y \mid X)] + \mathrm{V}_X[\mathrm{E}(Y \mid X)].$$

In the current context, one can use the above by conditioning on δ as follows (ignoring the conditioning on x for a moment),

$$\mathrm{V}(\Psi) = \mathrm{E}_\delta[\mathrm{V}(\Psi \mid \delta)] + \mathrm{V}_\delta[\mathrm{E}(\Psi \mid \delta)].$$

Using the above, it can be shown that the conditional variance for the mixture SDT-item model is, for signal,

$$\mathrm{V}(\Psi_j \mid x=1) = d_j^2\lambda_j(1-\lambda_j) + \lambda_j\mathrm{V}(\varphi_s) + (1-\lambda_j)\mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j).$$

The above shows that the mixture SDT-item model leads to additional terms in the signal variance (cf. DeCarlo, 2010) and so it frees up (with restrictions) the relation between the variance and covariance shown in Eq. (7), as in the empirical extensions. The model in essence attributes the apparent larger signal variance to mixing that occurs because of lapses in attention to the signal.

An example. To obtain some information about how the results for a fit of a mixture SDT-item model compare to those for the simple SDT-item model, the model was fit to the data of the first word recognition study discussed above. More complex models of this sort can be fit using software for a Bayesian approach, in which case the model is as given above, with the difference that priors are specified for the model parameters; specifically, normal priors were used for d_j and c_{jk} (i.e., the first criterion was normal with the remaining criteria obtained as zero-truncated normal increments, to maintain the ordering), normal priors were used for φ_1 and φ_2, and a uniform prior was used for λ_j. Markov chain Monte Carlo (MCMC) estimation was used with the software OpenBugs (Thomas, O'Hara, Ligges, & Sturtz, 2006). 10,000 burn-ins were used, followed by 20,000 iterations for posterior inference; this gave Monte Carlo errors that were less than 5% of the sample standard deviations, which suggests an adequate number of iterations (see Spiegelhalter, Thomas, Best, & Lunn, 2003).

For the mixture SDT-item model, the posterior means for d_j ranged from 0.50 to 4.7 with a mean of 2.48 and standard deviation of 1.10; note that d_j tends to be larger for the mixture model than for the simple item model because, from the mixture perspective, the item model provides an estimate of $\lambda_j d_j$ and not simply d_j (see DeCarlo, 2010). The posterior means for λ_j ranged from 0.11 to 0.96 with a mean of 0.64 and standard deviation of 0.23, which suggests that, on average, about 64% of the items were attended to.

Of particular interest are the item variances. For the SDT-item model, the posterior means (and standard deviations) were 0.20 (0.05) and 0.02 (0.02) for $\mathrm{V}(\varphi_n)$ and $\mathrm{V}(\varphi_s)$, respectively, which are virtually identical to the MLE estimates found above (0.19 and 0.02). For the mixture SDT-item model, the posterior means for $\mathrm{V}(\varphi_n)$ and $\mathrm{V}(\varphi_s)$ were 0.24 (0.05) and 0.04 (0.04). Thus, the mixture-item model, compared to the simple item model, gives a small increase in the variance for new words (from 0.20 to 0.24) and also possibly for old words (0.02–0.04). Most important, the results lead to the same conclusion, in that the IIC is again larger for new items than for old items, and so this result appears to be consistent across the different models.

7.2. Random observer effects

The design considered here is *mixed*, in that the item effects, φ_s and φ_n, are random whereas the observer effects, c_{jk} and d_j, are fixed (i.e., they are observer-specific fixed effects, as in traditional SDT). Here it is noted that one can also introduce random

observer effects; this would be of interest, for example, if the goal was determine whether or not observers are exchangeable. This touches upon issues that go beyond the scope of the current article, however some basic results for versions of SDT-item models with random observer effects are given.

The structural model of Eq. (6) can be extended to allow for random observer effects as follows,

$$\Psi_j = \alpha_j + (d_j - \varphi_s)x + \varphi_n(1-x) + \varepsilon_j, \tag{12}$$

where d_j is now random, specifically $d_j \sim \mathrm{N}(\mu_d, \sigma_d^2)$ with $\mu_d = \mathrm{E}_j\mathrm{E}(\Psi_j \mid x = 1) = \mathrm{E}_j(d_j)$, and $\alpha_j \sim \mathrm{N}(0, \sigma_\alpha^2)$, where E_j indicates expectation over the J observers (whereas E is expectation over the items). The parameter α_j is a random intercept; this basically allows the response criteria to vary across observers (i.e., shifted up or down; see DeCarlo, 2010); a restriction on the criteria across observers is also necessary in order for the model to be identified (for example, Pratte et al. set the middle criterion to zero). Note that it is now assumed that, for each item, there is a perceptual distribution *across observers*, with population means $\mu_d = \mathrm{E}_j\mathrm{E}(\Psi_j \mid x = 1)$ and $\mu_\alpha = \mathrm{E}_j\mathrm{E}(\Psi_j \mid x = 0) = 0$, whereas this assumption is not made in the mixed model (i.e., perceptions are random within observers in the mixed model, with no distributional assumptions about perceptions across observers). The above model is closely related to that of Pratte et al. (2010),[4] also see Rouder and Lu (2005) and Rouder et al. (2007).

Note that the IIC of Eq. (4) can be used for models with random observer effects, in that it is defined for given observers (i.e., it is conditional on the observers), and so it does not include observer variance in the denominator. In particular, the conditional covariances for the structural model of Eq. (12) are the same as those given for Eq. (7), that is, $\mathrm{V}(\varphi_s)$ and $\mathrm{V}(\varphi_n)$, because $\mathrm{Cov}(d_j, d_{j'})$ and $\mathrm{Cov}(\alpha_j, \alpha_{j'})$ are zero, given the assumption that d_j and α_j are independently and identically distributed (and are constants for observers j and j'). The conditional variances within observers are also the same as in Eq. (7), because, for a given observer, $\mathrm{V}(d_j \mid x = 1) = 0$ and similarly for α_j (i.e., d_j and α_j are random across observers but constant within observers). It follows that the intra-item correlation of Eq. (4) is the same for models with fixed or random observer effects; the IIC measures how much of the signal or noise variance within an observer is due to item effects.

One can also develop versions of the IIC that use a 'total' or 'unconditional' variance in the denominator, that is, a variance over both items i and observers j, $\mathrm{V}_j = \mathrm{E}_j\mathrm{E}[\Psi_j - \mathrm{E}_j\mathrm{E}(\Psi_j)]^2$, where the subscript j on V_j is used to indicate that the variance is across observers (whereas V is variance within observers). For a model with random observer effects, as given above, it follows from Eq. (12) that the 'total' variances are

$$\mathrm{V}_j(\Psi_j \mid x_j = 0) = \sigma_\alpha^2 + \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j)$$
$$\mathrm{V}_j(\Psi_j \mid x_j = 1) = \sigma_\alpha^2 + \sigma_d^2 + \mathrm{V}(\varphi_s) + \mathrm{V}(\varepsilon_j).$$

The above shows that the total variance includes variance due to observers, $\sigma_\alpha^2 = \mathrm{V}_j(\alpha_j)$ and $\sigma_d^2 = \mathrm{V}_j(d_j)$; this approach applies to a population of observers.

Intra-item correlations based on the total variance can be defined for noise and signal, respectively. In particular, it follows from Eq. (12) that,

$$\mathrm{IIC}_{2n} = \frac{\mathrm{V}(\varphi_n)}{\sigma_\alpha^2 + \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j)}$$
$$\mathrm{IIC}_{2s} = \frac{\mathrm{V}(\varphi_s)}{\sigma_\alpha^2 + \sigma_d^2 + \mathrm{V}(\varphi_s) + \mathrm{V}(\varepsilon_j)}. \tag{13}$$

Eq. (13) presents versions of the intra-item correlation that give the covariance between observers relative to the total variance across and within observers (for a model with uncorrelated random observer and item effects). The main consequence is that the IIC_2 of Eq. (13) will tend to be smaller than the IIC of Eq. (4), because of the additional terms in the denominator (i.e., observer variance); note that IIC_2 is also now dependent on observer characteristics (i.e., the variation in d_j and α_j across observers). The IIC_2 for signal also has an extra term in the denominator, which follows from the conceptualization in terms of shifting criteria (and so the signal and noise distributions both shift together); the extra term could account for a smaller signal IIC, however it did not for the word studies examined above (i.e., the signal IIC was still smaller without it). The important aspect to note is that the finding above of a larger IIC for new words than old words still holds for IIC_2. Note that one can also define *intra-observer* correlations (i.e., using the covariance across items) as well as other types of intra-item correlations, which require additional theoretical and empirical study.

In Pratte et al.'s (2010) parameterization of the model, the signal and noise distribution locations are random (rather than the response criteria, and so the distributions shift separately rather than together) and it can be shown that the resulting intra-item correlations are

$$\mathrm{IIC}'_{2n} = \frac{\mathrm{V}(\varphi_n)}{\sigma_{\alpha_n}^2 + \mathrm{V}(\varphi_n) + \mathrm{V}(\varepsilon_j)}$$
$$\mathrm{IIC}'_{2s} = \frac{\mathrm{V}(\varphi_s)}{\sigma_{\alpha_s}^2 + \mathrm{V}(\varphi_s) + \sigma^2\mathrm{V}(\varepsilon_j)} \tag{14}$$

where the term σ^2 is included because the unequal variance model is generalized, as noted in the previous section. For Pratte et al.'s data, estimates of $\mathrm{V}(\varphi_n)$ and $\mathrm{V}(\varphi_s)$ are 0.20 and 0.16, as noted above, whereas estimates of $\sigma_{\alpha_n}^2$ and $\sigma_{\alpha_s}^2$ are, from their Fig. 3(C), $0.43^2 = 0.18$ for both signal and noise. Using these estimates, together with Pratte et al.'s estimate of σ of 1.36 (see their Fig. 2) give values of IIC'_{2n} and IIC'_{2s} of 0.14 and 0.07, respectively (which are slightly smaller than the values of 0.17 and 0.08 found above for IIC, as expected), and so once again the intra-item correlation is larger for new words than for old words.

The difference between IIC_2 and IIC is analogous to the difference between 'reliability' (agreement) versions of the intra-class correlation (ICC), where the 'total' variance includes rater (observer) variance, versus 'consistency' versions of the ICC, which do not include rater variance (e.g., see McGraw & Wong, 1996 and Shrout & Fleiss, 1979); note that consistency versions of the ICC(2) and ICC(3) given by Shrout and Fleiss (i.e., random and mixed versions of the ICC, respectively), give the same results, just as the IIC of Eq. (4) gives the same results for random and mixed models. Similarly, in classical test theory, two measures of measurement precision are defined (conditional and unconditional, that is, for a given subject or for a population of subjects; see Mellenbergh, 1996); also see the generalizability coefficient and index of dependability used in generalizability theory (see Brennan, 2001). Of course, the intra-item correlations examined here apply to *residual* correlation across observers, that is, correlation *not* due to the item being a signal or noise, and not *true score* variance, and so they are not measures of reliability, but rather measure the proportion of excess signal or noise variance that is due to an item effect. Nevertheless, earlier discussions about issues related to the choice between the different types of measures are relevant. The choice of measure depends on the purpose and design of the study; for studies of item effects and factors that affect them, for example, the IIC of Eq. (4) is likely to be adequate.

Further research on SDT-item models and extensions should include derivations of the conditional variances and covariances, which clarify important details of the models, as well as a presentation of the decision and structural models. Together, this

[4] Note that, for the current model, as well as for Pratte et al.'s (2010) model, the 'random' observer effects are not fully random, in that, apart from the random slope (d_j) and intercept (α_j), the spacing between the criteria are fixed effects.

will help to further our knowledge and understanding of item effects in applications of SDT.

Appendix. A Latent Gold program for the SDT-item model

```
infile 'C:\ Documents and Settings\ Desktop\ word_97.sav'
model
options
  algorithm nr
  tolerance = 1e–008 emtolerance = 0.01 emiterations = 250 nriterations = 50;
  startvalues
  seed = 0 sets = 10 tolerance = 1e–005 iterations = 50;
bayes
  categorical = 0 variances = 0 latent = 0 poisson = 0;
  quadrature nodes = 11;
  standarderrors = standard;
  output parameters profile bvr;
  //Note: next line saves the factor scores to an SPSS file//
  outfile 'word_97_out.sav' classification;
variables
  dependent y1 probit, y2 probit, y3 probit, y4 probit, y5 probit,
  y6 probit, y7 probit, y8 probit, y9 probit, y10 probit, y11 probit,
  y12 probit, y13 probit, y14 probit, y15 probit, y16 probit,
  y17 probit, y18 probit, y19 probit, y20 probit, y21 probit;
  independent x;
  latent phi_s continuous, phi_n continuous;
  //Note: LG models p(Y > k), which reverses the signs for symmetric dist.//
  equations
  phi_s; phi_n;
  y1 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y2 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y3 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y4 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y5 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y6 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y7 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y8 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y9 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y10 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y11 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y12 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y13 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y14 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y15 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y16 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y17 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y18 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y19 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y20 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
  y21 <- 1 + x + (−1)phi_s x + (−1)phi_n x + (1)phi_n;
end model
```

References

Boutis, K., Pecaric, M., Seeto, B., & Pusic, M. (2010). Using signal detection theory to model changes in serial learning of radiological image interpretation. *Advances in Health Sciences Education*, *15*, 647–658.

Brennan, R. L. (2001). *Generalizability theory*. NY: Springer.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*, 335–359.

DeCarlo, L.T. (1997). [Word recognition, 120 trials, 21 observers]. Unpublished raw data.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205.

DeCarlo, L. T. (2002). Signal detection theory with finite mixture distributions: theoretical developments with applications to recognition memory. *Psychological Review*, *109*, 710–721.

DeCarlo, L. T. (2003). Using the PLUM procedure of SPSS to fit unequal variance and generalized signal detection models. *Behavior Research Methods, Instruments, & Computers*, *35*, 49–56.

DeCarlo, L. T. (2008). Studies of a latent class signal detection model for constructed response scoring. ETS research report no. RR-08-63. Princeton, NJ: Educational Testing Service.

DeCarlo, L. T. (2010). On the statistical and theoretical basis of signal detection theory and extensions: unequal variance, random coefficient, and mixture models. *Journal of Mathematical Psychology*, 304–313.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language*, *62*, 1–18.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. NY: John Wiley & Sons.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293–299.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*, 30–46. (Correction, 1, 390).

Morey, R. D., Pratte, M. S., & Rouder, J. N. (2008). Problematic effects of aggregation in zROC analysis and a hierarchical modeling solution. *Journal of Mathematical Psychology*, *52*, 376–388.

Pratte, M. S., Rouder, J. N., & Morey, R. D. (2010). Separating mnemonic process from participant and item effects in assessment of ROC asymmetries. *Journal of Experimental Psychology Learning, Memory, and Cognition*, *36*, 224–232.

Qu, Y., & Hadgu, A. (1998). A model for evaluating sensitivity and specificity for correlated diagnostic tests in efficacy studies with an imperfect reference test. *Journal of the American Statistical Association*, *93*, 920–928.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Sun, D., Speckman, P. L., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*, 420–428.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models*. NY: Chapman & Hall/CRC.

Spiegelhalter, D., Thomas, A., Best, N. G., & Lunn, D. (2003). WinBUGS user manual. Retrieved from: http://www.mrc-bsu.cam.ac.uk/bugs/.

Thomas, A., O'Hara, B., Ligges, U., & Sturtz, S. (2006). Making BUGS open. *R News*, *6*, 12–17. Retrieved from: http://www.rni.helsinki.fi/~boh/publications/Rnews2006-1.pdf.

Vermunt, J. K., & Magidson, J. (2007). *LG-syntax™ user's guide: manual for latent gold 4.5 syntax module*. Belmont, MA: Statistical Innovations Inc.