Automatic Response Option Sampling for Situational Judgment Items

Anne Thissen-Roe, Ph.D.
Comira

Stephen Gunter, Ph.D.
Camber Corporation
Orlando, Florida
jgunter@camber.com

Automatic Item Generation (AIG) has been the subject of considerable interest in recent years, and for good reason. In conjunction with Computerized Adaptive Testing (CAT) or Linear On-The-Fly Test Assembly (LOFT) systems, automatically generated items, with appropriately developed and validated Item Response Theory (IRT) scoring models, allow for a proliferation of alternate test forms that produce comparable, equally valid scores on a common scale. Automatically generated items conserve and extend scarce item development resources, while supporting exposure control and consistent item quality.

In AIG, multiple choice items can be automatically generated based on a prototype item, by systematically altering or substituting elements of the stem, the response options, or both. For some types of multiple choice items, it is essential that one response option be objectively correct, but for others, it is only necessary that the response options be meaningfully comparable on one or more measurement dimensions of interest. For items in the latter category, the same stem can be used with many overlapping or disjoint sets of response options. Multidimensional Forced Choice (MFC) personality items (see, e.g., Chernyshenko, Stark, Prewett, Gray, Stilson & Tuttle, 2009; Brown, 2010; Heggestad, Morrison, Reeve & McCloy, 2006), for example, have completely response-independent stems; Situational Judgment Item (SJI) stems may also be response-independent within narrower semantic limits.

In this paper, we present a system of response option substitution, Automatic Response Option Sampling, as a form of Automatic Item Generation, and demonstrate its application to Situational Judgment Items.

Situational Judgment Items

Situational Judgment Items come in numerous formats and with various response instructions. In general, they take the form of a stem describing (or showing, through images, video, or a game-like simulation) a work situation, followed by four to six response options describing (or showing) possible actions a worker might take in the situation. A candidate may be asked to select responses based on their effectiveness or congruence with the candidate's own preferences and behaviors (i.e., "should do" or "would do"; McDaniel, Hartman, Whetzel, & Grubb, 2007); candidates may be asked to rate several response options independently, or to provide a partial ranking, such as by selecting the most effective single response option (Brown, 2010). In this study, we focus on the modeling and administration of SJIs with partially and fully-ranked response options.

Despite their strong face validity and empirical correlations with job performance (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001; McDaniel et al., 2007), SJIs have proven difficult to model and thus to incorporate into systems such as CAT, let alone AIG. In particular, SJIs generated through the critical incidents methodology are largely atheoretical, and any scoring key, much less item model, is imposed after the construction of the stem and response options (Bergman, Drasgow, Donovan, Henning, & Juraska, 2006); two SJIs on the same test may be highly dissimilar in measurement properties. Further, due to the nature of the judgment being tested, SJIs are not generally amenable to modeling in a manner that would allow response options to be generated, and valid scoring models obtained, without a coding or calibration based on human judgment -- whether that amounts to expert review, or empirical pre-testing and predictive validation.

Certain attempts to identify and validate the constructs measured by SJIs are promising. Under an Implicit Trait Policy (ITP; Motowidlo, Hooper & Jackson, 2006) model, response options are coded or deliberately constructed for the expression of personality traits in work contexts. Knowledge of the effectiveness of expressing personality traits such as agreeableness at work (i.e., "the customer is always right") is a component of job knowledge, and can mediate the effect of a candidate's personality traits on job performance. One or more ITPs may be applicable in a given situation. Similarly, Stemler and Sternberg (2006) created SJI response options of various effectiveness that each reflected one of seven general approaches to work problems: comply, consult, confer, avoid, delegate, legislate or retaliate. The items produced did not measure the approaches as orthogonal factors; Harvey (2016) reported an attempt to produce items through the same method such that the factor structure was empirically reproducible, in which he came up with a six-factor variation. Based on these results, we know that SJIs can be used to measure multiple specific constructs, not merely "general domain knowledge," "problem solving" or "good judgment."

One stumbling block in the application of classical and modern test theory has been that SJIs are often *intrinsically* multidimensional. That is, not only do the various situations differentiate candidates on more than one job-relevant latent trait, response options to a single situation may reflect non-proportional levels of the latent traits. Is deferential agreeableness, brutal honesty, bargaining, cheerleading, or strict rule-following the most effective strategy in a particular work situation? The candidate is being asked to compare responses that reflect different job-relevant competencies. The item is itself multidimensional, rather than merely being embedded in a multidimensional test.

There are three major approaches to dealing with the challenge of scoring and validating intrinsically multidimensional SJIs. Most often in practice, an atheoretical model is adopted. Response options are scored according to a consensus or expert model of contextual effectiveness, or according to an empirical criterion-referenced model (e.g., option scores are derived from a regression model predicting job performance). This approach foregoes niceties like form equating and compatibility with CAT systems.

A second approach is to modify the SJI development and/or vetting process so as to produce intrinsically unidimensional items, embedded in a multidimensional test. Lievens and Motowidlo (2016) advocate "lining up" the responses to a given situation such that expert-rated response effectiveness corresponds closely to the level of a job-relevant trait. The correspondence is enforced via item writing instructions, and also through a two-rater verification process that rejects sufficiently multidimensional items. Lievens and Motowidlo (2006) proposed a set of classical (e.g., accumulative) scoring procedures for such items. However, they are also clearly compatible with the use of standard multidimensional item response theory models such as the Nominal Response Model (NRM; Thissen, Cai & Bock, 2010). Zu & Kyllonen (2012) successfully applied the NRM to entire tests of SJIs designed to measure single constructs; the method generalizes even when the between-items factor model is complex.

The third and final approach embraces the intrinsic multidimensionality of SJIs and incorporates it in the model used for scoring and validation. Thissen-Roe (2013) fit the Multidimensional Nominal Response Model (MNRM; Bolt & Johnson, 2009) to a set of SJIs used to screen for customer service behaviors in hourly job candidates. The items collectively, and in some cases individually, measured three dimensions: overall effectiveness, an orientation toward more active or more passive responses, and a relative prioritization of coworkers' needs versus the company's rules. The items were not written to assess, and the dimensions did not align with, well-established traits such as conscientiousness and agreeableness; rather, the secondary axes appear to be equi-effective contrasts of traits. The constructs measured were consonant with those measured by a set of co-administered pairwise preference personality items, although the relationship was established through simultaneous calibration of a single mixed-type IRT model across both sets of items, rather than traditional convergent and discriminant validation.

Even intrinsically multidimensional SJIs can be modeled, using the MNRM. However, such a model lends no clues to the behavior of any modified form of a given SJI. By themselves, MNRM models of SJIs are not a pathway to AIG.

Automatic Response Option Sampling

For this reason, we have developed Automatic Response Option Sampling (AROS) as a means of Automatic Item Generation for SJIs. SJIs may be assembled from a stem and an automatically-selected subset of several available, calibrated response options. Even with a fixed stem, switching out the response options results in a considerable number of item instances. For example, an SJI might be written with nine response options, each reflecting an ITP for expression of a different personality trait, and the AROS algorithm might be set to select four response options for each candidate, creating a total of 630 combinations. A combination of response options could be chosen, without human intervention, during form construction or even during test administration, "on the fly."

Depending on the design of the test, the response options might be selected according to different principles. They might be selected entirely at random, of course, but more sophisticated algorithms are possible. It would not be difficult to build a LOFT system that used AROS to meet specifications for the number of items testing each ITP, or pairwise combination of ITPs. A CAT system could select the traits where the most additional information was needed, or where it was available, based on the match of the candidate's current posterior probability distribution with the information profile resulting from each combination of response options.

In order to score an SJI assembled by AROS, it is necessary to have a configurable IRT model which permits the behavior of the fully assembled item to be inferred from attributes of its component response options. While the MNRM was useful for beginning to understand the behavior of intrinsically multidimensional SJIs, it does not serve this purpose.

A more suitable general class of model has been developed for the purpose of studying and administering Multidimensional Forced Choice (MFC) personality items. Models such as the Multi-Unidimensional Pairwise Preference Model (MUPP; Stark, Chernyshenko and Drasgow, 2005) use abstract models of choice and decision-making to describe the behavior of a combination of response options with known individual response characteristics. Within this class of models, as reviewed by Brown (2016), various combinations of component models (for individual response options) and decision models for choice behavior may be situationally appropriate, depending on the format of available calibration data (e.g., dichotomous or polytomous) and the nature of the items and responses themselves.

Case Study

We present findings from a simple prototype of an AROS system. Data were available from an operational test in which SJIs were administered as 4-6 independently rated response options per situation. We modeled the operational data, and simulated candidate responses to versions of the items in which only three options were presented and the instructions were for partial or full ranking. AROS was used to assemble the triads independently within each simulation, and an MFC-style combinatoric-assembly model was used to impute responses.

Source Data

Operational data were drawn from regular skills testing of specialized employees at a federal agency. Test results were used for internal certification. Candidates were incumbent supervisors of service representatives. The test comprised SJIs testing leadership and interpersonal skills in the context of the specific position. Items required specific job knowledge, especially of procedures and terminology, to complete.

Items on the operational test were presented as blocks, in which a situation was followed by four to six response options. The candidate rated each response option for effectiveness on a seven-point Likert-type scale. Operationally, responses were scored according to their proximity to consensus ratings made by a panel of subject matter experts; a test score was accumulated across response options and items.

For the AROS prototype development, 46 blocks were used, each of which had been used in some, but not all, of a sequence of eight consecutive administrations. A calibration sample size of N=1,253 candidates was obtained, with each candidate supplying a necessarily-partial response pattern.

We had initially conceptualized AROS as being used with response options each constructed to measure a single, distinct ITP, following the MFC model in which each response option targets a different personality dimension. This is an ideal situation in that it allows for dimension reduction in item model assembly and candidate trait estimation. However, to achieve or approximate the ideal of simple structure in practice, the items must be intentionally written such that the response options measure distinct constructs. However, as is typical of SJTs, our operational items had been written from job analysis data to reflect the job-related situations that these supervisors would likely face on the job and possible responses that the supervisors might choose to solve those situations. Thus, our attempts to fit a simple structure model to the data were unsuccessful. Our best working model of the items allowed many of the response options to load on more than one of its four latent traits.

Within the four-dimensional measurement space, we modeled each rating using either the Generalized Partial Credit Model (GPCM; Muraki, 1992) or the Nominal Response Model (NRM; Bock, 1972; Thissen, Cai & Bock, 2010). GPCM was chosen over similar models for ordered polytomous responses because it can be written as a constrained form of the NRM; for item assembly purposes, GPCM-modeled items were treated as NRM items with certain predictable parameters.

It was necessary to use the NRM rather than GPCM to model some responses, especially those for which the "best" rating was in the middle of the scale, in order to achieve a better-fitting model. These response options display a behavior which is, in essence, the reverse of an unfolding-type item. In an unfolding item model, candidates whose trait levels are in the middle (near the location of the item) respond using the highest categories, while candidates whose trait levels are well above or well below the item both respond using the low categories. In this case, candidates whose trait levels were high responded using middle categories, while candidates whose trait levels were lower were more likely to respond using either the high or low categories. NRM is capable of re-ordering the responses such that the middle options are located at the top, and both extremes are at the bottom, of the latent trait. Its minimization process infers the high value of a middle option, and the intermediate positions of the various other options.

In order to anchor the calibration process such that an interpretable model would result, within each situation, GPCM was used for the response or responses with the highest expert consensus rating. For those items, the responses were assumed to proceed in ascending category order across levels of the

latent trait. Spot-checking by removing the constraint for individual responses suggested that the assumption was valid, or at least a good approximation of reality.

GPCM was also used for certain response options that were not consistently available to candidates, having been removed from or added to a situation that was used in multiple administrations. Those response options did not have sufficient sample size to robustly estimate NRM parameters.

We make no claim at this juncture to have used *the correct* model for the set of SJIs. For our purposes, it was sufficient to choose the best of those available. Subsequent analyses are structured to compensate for model inaccuracy. However, in an operational use of AROS, the accuracy of the component models would be paramount. Strategies to achieve an accurate set of component item models include planful item writing, designed to facilitate confirmatory item factor analysis; consistent (or at least symmetrically assorted) presentation of response options to all candidates seeing a situation; and the pursuit of a calibration sample size at least an order of magnitude greater than 1,253.

Item Assembly

For each of the 46 blocks modeled, our simple AROS system selected three response options to present to each candidate. If the system were operational, the items would be presented to candidates with "most effective - least effective" type response instructions, in order to obtain a full ranking, or "most effective" instructions alone in order to obtain a partial ranking.

In a fully ranked triad, six possible responses are available, each representing one *most-least* combination, or, equivalently, one *most-middle-least* ranking. An item characteristic surface (ICS) exists for each of the six responses; they need not be ordered in any dimension.

We extended the work of previous authors in the MFC domain in order to calculate the ICS for each option combination in a fully ranked triad, rather than a pairwise preference item. Stark et al. (2005) outlined procedures for construction and calibration of item stems on several dimensions, enabling the combinatoric synthesis of item characteristic surfaces from single-stimulus item characteristic curves. To date, Stark et al.'s combinatoric method has been applied solely to dyads, whereas a method for ranking data, related to confirmatory factor analysis and structural equation modeling, has been applied to triads and above (Chernyshenko et al., 2009; Brown, 2010; Brown, 2016). However, the combinatoric method is readily extensible to triads, tetrads, pentads and sextads, and transferable to SJIs. In this section, we focus on triads.

We derived our predicted ICS equation from Luce's choice axioms (Luce, 1959; Luce, 1977), under Brown's (2016) framework for forced choice type models. We followed Andrich (1989) and Stark et al. (2005) in defining the available options as a set of allowable combinations of would-be ratings under the component models; like the foregoing authors, we deemed inadmissible any combination resulting in a

tie between options. However, those authors used dichotomous component models to predict pairwise preference. For example, the endorsement of option *t* over option *s* is modeled as follows:

$$P(t > s \mid \theta_s, \theta_t) = \frac{P_s(0)P_t(1)}{P_s(0)P_t(1) + P_s(1)P_t(0)} \quad (1)$$

where $P_s(0)$ indicates non-endorsement of option *s*, $P_t(1)$ indicates endorsement of option *t*, and so on. As far as we know, assembly of polytomous component models under Luce's choice axiom has not been presented before in the forced choice literature (Brown, 2016). To construct a triad ranking from polytomous component models, we needed to write the available options more generally as strict inequalities:

$$P(t > u > v \mid \theta_v, \theta_t, \theta_u) = \frac{P(k_v < k_u < k_t)}{P(k_v < k_u < k_t) + P(k_v < k_t < k_u) + P(k_t < k_v < k_u) + \cdots} \quad (2)$$

(Option *v* has replaced option *s,* for clarity of subsequent expressions.)

With polytomous component models having seven response options each, the inequalities describing each allowable configuration expand to thirty-five specific combinations of would-be ratings apiece. That is, the inequality $k_v < k_u < k_t$ describes the case where $k_v = 7$, $k_u = 4$, and $k_t = 1$, as well as the case where $k_v = 4$, $k_u = 3$, and $k_t = 2$, and thirty-three other such cases. The probability $P(k_v < k_u < k_t)$ is equal to the sum of the probabilities of those thirty-five disjoint cases.

Although extant assembly models of multidimensional forced choice items are predicated upon simple structure, it is possible to predict the ICSs of an assembled item even if simple structure is not present. We demonstrate this possibility by assembling ICSs from component single-stimulus items modeled with the NRM in a multidimensional latent trait space.

The NRM in four dimensions can be written as:

$$P(k) = \frac{e^{((a_1\theta_1 + a_2\theta_2 + a_3\theta_3 + a_4\theta_4)s_k + c_k)}}{\sum_k e^{((a_1\theta_1 + a_2\theta_2 + a_3\theta_3 + a_4\theta_4)s_k + c_k)}} = \frac{e^{(\sum_{j=1}^{4} a_{tj}\theta_j s_k + c_k)}}{\sum_k e^{(\sum_{j=1}^{4} a_{tj}\theta_j s_k + c_k)}} \quad (3)$$

.

Allowable configurations then look like this:

$$P(k_v, k_t, k_u \mid k_v < k_u < k_t) \quad (4)$$

$$= \sum_{k_t=3}^{7} \sum_{k_u=2}^{k_t} \sum_{k_v=1}^{k_u} \left(\left(\frac{e^{(\sum_{j=1}^{4} a_{tj}\theta_j s_{tk_t} + c_{tk_t})}}{\sum_{k_t} e^{(\sum_{j=1}^{4} a_{tj}\theta_j s_{tk_t} + c_{tk_t})}} \right) * \left(\frac{e^{(\sum_{j=1}^{4} a_{uj}\theta_j s_{uk_u} + c_{uk_u})}}{\sum_{k_u} e^{(\sum_{j=1}^{4} a_{uj}\theta_j s_{uk_u} + c_{uk_u})}} \right) * \left(\frac{e^{(\sum_{j=1}^{4} a_{vj}\theta_j s_{vk_v} + c_{vk_v})}}{\sum_{k_v} e^{(\sum_{j=1}^{4} a_{vj}\theta_j s_{vk_v} + c_{vk_v})}} \right) \right)$$

.

Notably, the denominator of every case within every allowable configuration -- all two hundred ten cases -- is identically the product of the three denominators of the individual NRM components. When each assembled ICS is normalized, that denominator appears top and bottom and therefore cancels, simplifying calculation of the assembled ICS considerably:

$$P(k_v, k_t, k_u \mid k_v < k_u < k_t) = \left(\frac{\sum_{k_t=3}^{7} \sum_{k_u=2}^{k_t} \sum_{k_v=1}^{k_u} \left(\left(e^{(\sum_{j=1}^{4} a_{tj}\theta_j s_{tk_t} + c_{tk_t})} \right) \left(e^{(\sum_{j=1}^{4} a_{uj}\theta_j s_{uk_u} + c_{uk_u})} \right) \left(e^{(\sum_{j=1}^{4} a_{vj}\theta_j s_{vk_v} + c_{vk_v})} \right) \right)}{\left(\sum_{k_t} e^{\left(\sum_{j=1}^{4} a_{tj}\theta_j s_{tk_t} + c_{tk_t}\right)} \right) \left(\sum_{k_u} e^{\left(\sum_{j=1}^{4} a_{uj}\theta_j s_{uk_u} + c_{uk_u}\right)} \right) \left(\sum_{k_v} e^{\left(\sum_{j=1}^{4} a_{vj}\theta_j s_{vk_v} + c_{vk_v}\right)} \right)} \right) \quad (5)$$

. This is a convenient property of all divide-by-total component models.

Less conveniently, no dimension reduction was possible; candidate traits had to be estimated in four-dimensional space. The lack of dimension reduction could be a thorny problem in practice. High dimensionality turns fixed-quadrature theta estimation into an exercise in complex (and slow) memory management, limiting the potential of such a model to be used in, for example, live CAT. MHRM (Cai, 2010) and related scoring algorithms do not suffer a penalty for added dimensions, but also may not produce completely deterministic scores for a single response pattern. This is a fairness issue for candidates, in that two candidates can enter the same responses, and yet one passes and one fails. Technical solutions may provide some relief: a consistent seed for the random number generator can make MHRM deterministic; fixed quadrature estimation could be offloaded to the GPU. Still, high-dimensional models are better avoided or reduced when possible.

Using AROS, our 46 original blocks spawn 4-20 triads apiece, for a total of 348. The generated triads derived from the same operational block were commonly dissimilar in measurement properties, even though they used the same situation and shared up to two of three response options. Figure 1 shows an example of the four triads generated from one item, with the response options being compared summarized as single words. The ICSs are neither shaped nor configured the same between triads. They also clearly show that they retain their intrinsic multidimensionality: they are not "lined up."

Simulation

We conducted three simulations in order to compare the performance of an AROS-derived test with a comparably brief form created from whole items. The three simulations each adhered to the same general procedure. First, **Θ** was estimated in four dimensions for each actual candidate (N=1,253). Second, 5,000 simulated response patterns were generated per candidate for each test form: the AROS form, the brief comparison form, and the full form of all 46 items. The AROS form and the brief form were regenerated each time, with new response options selected at random for the AROS form and new whole items selected at random for the brief form, as might be done for exposure control. Third, each response pattern was scored, rationally matching scoring methods across forms as closely as possible. The Item Response Theory model was not used in scoring the response patterns, because it was used to generate them; a test of theta recovery was not intended. Finally, the AROS form was compared to the brief form based upon its recovery of the score computed from the full form. Simple Pearson correlations across the full set of imputed scores were used for clarity, although we did verify the pattern of results using within-candidate averages.

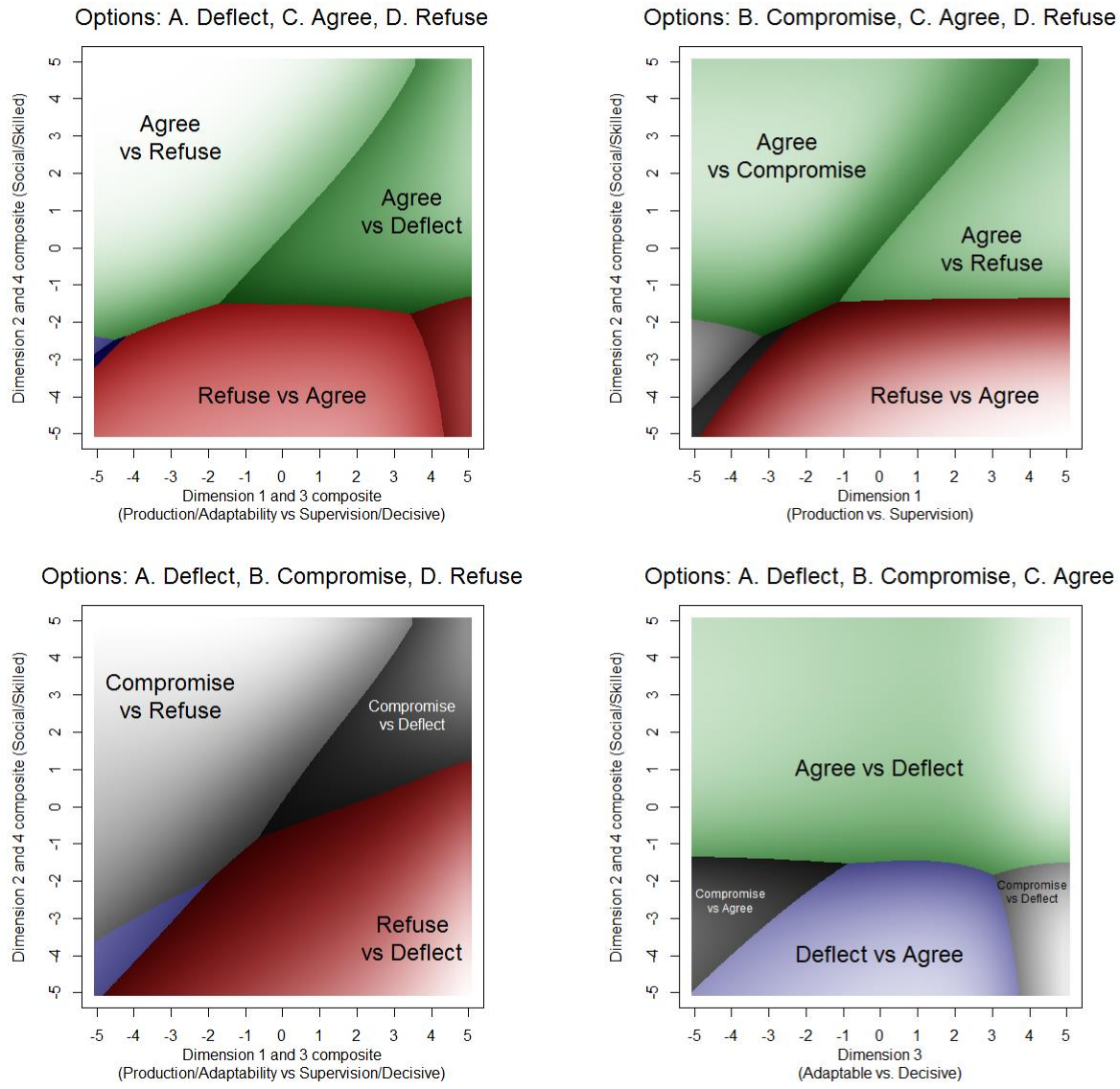
This procedure was somewhat conservative in that the brief form and full form were always scored identically, so that any mismatch of scores due to necessary differences in the scoring method affected only AROS.

Generated responses were always used, even when an actual candidate response was available. There were two reasons for this. First, the matching actual responses would have always applied only to the brief and full forms, and not AROS, because no actual responses were collected under ranking instructions; use of those responses would have tilted the scales more substantially than the scoring method differences mentioned above. Second, the number of actual responses made varied considerably by candidate, and was not independent of the trait estimate. For these reasons, use of actual responses added noise and bias to the system rather than clarifying the results.

Three cases, comprising response instructions and scoring methods, were examined under this general methodology.

Case 1. Most/Highest. Triad response instructions were to select the **most effective response only**. Under these instructions, 46 responses were made by the candidate. The brief comparison form was 10 items long, also averaging 46 responses. Each triad was scored +1 if the expert consensus response was chosen, +0 otherwise. In this condition *only*, the most effective response was always present. Blocks on the full form and brief form were scored +1 if the expert consensus response was rated highest. "Ties" between options were possible, and were awarded in the candidate's favor.

Figure 1. Item 71, operational version, has four response options. The four assembled triads are considerably different in measurement properties. The plots below show a representative two-dimensional slice of the model for each derived triad. The dominant ranking (most vs least effective) is shown for each point in latent trait space. Large dominant regions are labeled; some rare combinations are entirely not visible. The darker borders between "bubbles," where two rankings are similarly popular, mark the regions of high information for each item. Note that, for a better representation of each triad, the horizontal axis varies for the two plots on the right-hand side.



Case 2. High/Low. Triad response instructions were to select the **most effective response and the least effective response**. Under these instructions, 92 responses were made by the candidate. The brief comparison form was 20 items long, also averaging 92 responses.

For this case, the expert consensus responses were supplemented in order to create a full rank ordering of the response options. Where expert ratings were not available or were undifferentiated (tied), the parameters of the item model were used to infer "better and worse" responses -- a method we do not recommend for operational use. Each triad was scored +1 if the best available response option was chosen as "most effective" and an additional +1 if the worst available option was chosen as "least effective." Blocks on the full form and brief form were scored +1 if the best option was rated highest, and an additional +1 if the worst option was rated lowest. No points were given for choices made if the response option was neither the most or least effective. "Ties" between options were again possible, and again were awarded in the candidate's favor.

Case 3. Full Rank. Triad response instructions were to select the **most effective response and the least effective response**. Again, a 20-item brief comparison form was used.

Each triad was scored +1 if the best available response option was chosen as "most effective" and an additional +1 if the worst available option was chosen as "least effective." Blocks on the full form and brief form were scored +1 for every option placed in its correct rank order, such that 4-6 points were available per item. "Ties" between options were again possible, and again were awarded in the candidate's favor.

The triad scoring protocol was identical in Cases 2 and 3. The high/low scoring protocol for rated blocks might be considered more comparable in that only the highest and lowest responses were scored for both response modalities; the scores were on the same scale, with the same number of points possible. However, the triads selected might be composed entirely of middle responses, which are often more difficult to correctly rank. It was the desire to incorporate the middle options that motivated Case 3.

Results

In general, the results were neutral to slightly favorable for AROS. The results are summarized in Table 1. The case the favored AROS the most was Case 1. However, the results for Cases 2 and 3 were nearly identical.

Table 1. Correlations with the full test score. Scores under comparison are based on AROS items and a brief test of comparable length in each case. N=6,265,000 (1,253 x 5,000 imputations). Due to the effective sample size, all differences shown are statistically significant at α=0.05; however, not all are significant in practical terms.

	AROS Triads	Subset of Items
Case 1. Most/Highest.	0.82	0.63
Case 2. High/Low.	0.78	0.77
Case 3. Full Rank.	0.77	0.74

Discussion

With regard to the larger differences between the AROS triads and the subset of items, we observed that candidates for this operational test are better at knowing *what to do* than *what not to do*. Therefore, there was more consistency and better reproducibility among highly-rated response options than low- and middle-rated response options, a factor that may be affecting the Case 2 and 3 comparisons.

Because the response options to the AROS triads, as well as the items in the brief comparison form, were selected at random, the small advantage of AROS triads is probably best interpreted as an advantage of broad coverage. All of the situations were presented, but with less arduous response instructions. By contrast, particularly when only ten items were selected, a simple short form did not always adequately represent the content span of the full test. As many situational judgment tests show low internal consistency reliability, this result can be expected to generalize.

An interesting question follows from the interpretation: how would AROS CAT compare to item-level CAT? A CAT version of AROS could simultaneously adapt at the levels of items *and* response options. In practice, the two-level adaptation is equivalent to entering all combinations of response options as items in the CAT item pool, and constraining each situation to appear no more than once.

No attempt was made to test the recovery of *all four* dimensions measured by the full set of items, for example through use of the item model; however, the pattern of results is expected to be similar. While the candidates' judgments of situational response effectiveness are the usual focus of testing, in some contexts, especially selection, it may be useful to build a multidimensional profile of a candidate, based on estimated latent trait levels, or to integrate the SJIs into a larger measurement model for that purpose.

While the present study addresses only SJIs, AROS as a technology is clearly applicable to other item types. Most obviously, the same system could be applied to MFCs. Additionally, there may be applications in cognitive diagnosis where a dimensional model of misconceptions is appropriate; there may also be a simple variation of the selection and assembly methods that allows for the presence and absence of cognitive skill evidence in each response option to a diagnostic problem.

Summary

In the preceding pages, we have introduced a system of Automatic Response Option Sampling (AROS), and demonstrated its function by creating a system of short forms of an operational Situational Judgment Test.

References

Andrich, D. (1989). A probabilistic IRT model for unfolding preference data. *Applied Psychological Measurement, 13,* 193-296.

Bergman, M. E., Drasgow, F., Donovan, M. A., Henning, J. B., Juraska, S. E. (2006). Scoring Situational Judgment Tests: Once you get the data, your troubles begin. *International Journal of Selection and Assessment, 14,* 223-235.

Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37,* 29–51.

Bolt, D.M. & Johnson, T.R. (2009). Applications of a MIRT model to self-report measures: Addressing score bias and DIF due to individual differences in response style. *Applied Psychological Measurement, 33,* 335-352.

Brown, A. (2010). How Item Response Theory can solve problems of ipsative data. (Doctoral dissertation, Universitat de Barcelona, 2010.) http://www.psychometrics.cam.ac.uk/uploads/documents/Anna_Brown_Bibliography/Tesis_Doctoral_Anna_Brown_2010_distribution.pdf

Brown, A. (2016). Item response models for forced-choice questionnaires: a common framework. *Psychometrika, 81:1,* 135-160.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Munro algorithm. *Psychometrika, 75:1,* 33-57.

Chernyshenko, O.S., Stark, S., Prewett, M.S., Gray, A.A., Stilson, F.R. & Tuttle, M.D. (2009). Normative scoring of multidimensional pairwise preference personality scales using IRT: Empirical comparisons with other formats. *Human Performance, 22:2,* 105-127.

Harvey, R. J. (2016). Scoring SJTs for traits and situational effectiveness. *Industrial and Organizational Psychology*, 9:1, 63-71.

Heggestad, E.D., Morrison, M., Reeve, C.L. & McCloy, R.A. (2006). Forced-choice assessments of personality for selection: Evaluating issues of normative assessment and faking resistance. *Journal of Applied Psychology,* 91:1, 9-24.

Lievens, F. & Motowidlo, S.J. (2016). Situational judgment tests: From measures of situational judgment to measures of general domain knowledge. *Industrial and Organizational Psychology*, 9:1, 3-22.

Luce, R.D. (1959). *Individual choice behavior: a theoretical analysis.* New York, NY: Wiley.

Luce, R.D. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15,* 215-233.

McDaniel, M. A., Hartman, N. S., Whetzel, D. L., & Grubb, W. L., III. (2007). Situational judgment tests, response instructions, and validity: A meta-analysis. *Personnel Psychology, 60*(1), 63-91.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*(4), 730-740.

Motowidlo, S.J., Hooper, A.C. & Jackson, H.L. (2006). A theoretical basis for situational judgment tests. In J.A. Weekley & R.E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 57-82). Mahwah, NJ: Erlbaum.

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

Stark, S., Chernyshenko, O. & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: the multi-unidimensional pairwise preference model. *Applied Psychological Measurement, 29:3,* 184-203.

Stemler, S. E., & Sternberg, R. J. (2006). Using situational judgment tests to measure practical intelligence. In J. A. Weekley & R. E. Ployhart (Eds.), *Situational judgment tests: Theory, measurement, and application* (pp. 107–131). Mahwah, NJ: Erlbaum.

Thissen, D., Cai, L., & Bock, R. D. (2010). The nominal categories item response model. In M. L. Nering & R. Ostini (Eds.), *Handbook of polytomous item response theory models: Development and applications* (pp. 43–75). New York, NY: Taylor & Francis.

Thissen-Roe, A. (2013). Modeling situational judgment items with multiple distractor dimensions. R.E. Millsap et al. (eds.), *New Developments in Quantitative Psychology* (pp. 251-265). New York: Springer.

Zu, J., & Kyllonen, P. (2012, April). Scoring situational judgment tests with item response models. Paper presented at the annual meeting of the National Conference on Measurement in Education (NCME).