Assessing Content Validity Through Correlation and Relevance Tools A Bayesian Randomized Equivalence Experiment

Byron J Gajewski, University of Kansas Medical Center
University of Kansas Medical Center, University of Kansas Medical Center
Diane K Boyle, University of Kansas Medical Center
Marjorie Bott, University of Kansas Medical Center
Larry R Price, Texas State University - San Marcos, et al.

Available at: https://works.bepress.com/byron_gajewski/22/
Assessing Content Validity Through Correlation and Relevance Tools
A Bayesian Randomized Equivalence Experiment

Byron J. Gajewski,1,2 Valorie Coffland,2 Diane K. Boyle,2 Marjorie Bott,2
Larry R. Price,3 Jamie Leopold,2 and Nancy Dunton2

1Department of Biostatistics, University of Kansas School of Medicine, Kansas City, KS, USA,
2University of Kansas School of Nursing, Kansas City, KS, USA, 3College of Education, Texas State University, San Marcos, TX, USA

Abstract. Content validity elicits expert opinion regarding items of a psychometric instrument. Expert opinion can be elicited in many forms: for example, how essential an item is or its relevancy to a domain. This study developed an alternative tool that elicits expert opinion regarding correlations between each item and its respective domain. With 109 Registered Nurse (RN) site coordinators from National Database of Nursing Quality Indicators®, we implemented a randomized Bayesian equivalence trial with coordinators completing “relevance” or “correlation” content tools regarding the RN Job Enjoyment Scale. We confirmed our hypothesis that the two tools would result in equivalent content information. A Bayesian ordered analysis model supported the results, suggesting that evidence for traditional content validity indices can be justified using correlation arguments.

Keywords: elicitation of priors, psychometrics, NDNQI, IACCV, construct validity, content validity

Motivation

Two related steps in the process of instrument validation include accumulating evidence based on the adequacy of item content (content validity) and evidence based on how a collection of items constitute internal structure (an aspect of construct validity) (DeVellis, 2003; Nunnally & Bernstein, 1994). Establishing content validity evidence includes assessing the magnitude of agreement of experts’ evaluations regarding the extent to which items match the definition of the construct being measured. Evidence of internal structure derives from factor analytic studies and an examination of item interrelationships based on data from participants (Wirth & Edwards, 2007).

While one could argue the terms “construct validity” and “content validity” have largely been replaced by Messick’s (1989) “validity as integrated evaluation,” and more recently Kane’s (2006) “validity as argument,” in practice, expert and participant data traditionally are analyzed separately. To achieve more efficient and economical validity evidence, seamless integration of the experts’ and participants’ data – in a method called “Integrated Analysis of Content and Construct Validity (IACCV)” – has been proposed (Gajewski, Coffland, Boyle, & Bott, in press). The challenge of this methodology is in placing both sets of data on the same metric. The framework for accomplishing this goal is via a fully Bayesian model (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Sinharay, Johnson, & Stern, 2006), wherein expert data are treated as a prior distribution updated to a posterior distribution using participants’ data. This approach was applied to an instrument designed to measure nursing home culture change (Gajewski et al., in press). In measuring nursing home culture change, IACCV proved to be useful with the posterior distribution resulting in more stable estimates of usual psychometric parameters, particularly when faced with a small number of participants (Gajewski et al., in press).

The rationale for using Bayesian approach for IACCV is justified in at least two ways. First, Bayesian inference provides a formal framework for specifically including and combining information and outcomes from the literature (or expert opinion) with data from one’s current study in order to update knowledge about the phenomenon under study (Gelman et al., 2004). Second, when expert opinion is available, a Bayesian method has lower mean squared error than traditional analyses (Samaniego & Reneau, 1994). Integrated analysis of content and construct validity uses a Bayesian approach to instrument development and potentially has a lower mean squared error than traditional analyses under various expert opinion conditions – thus more efficient use of information. Among the applications of IACCV is current research on heart failure outcomes and studies with family members of patients requiring life-long, 24-hr, central venous catheter infusions due to orphan
illnesses such as Crohn’s disease. While IACCV shows great potential, expert opinion can only be justified if it is a validated source of information.

The focus of this study is on the content validity aspect of IACCV and its potential as a viable source of prior information. Content validity assessments provide investigators with experts’ view of the validity of items on an instrument. A typical content validity procedure (Grant & Davis, 1997) requires several content experts, where an instrument is given to each expert along with a content review form and definition of an intended measured domain. Typically, reviewers are asked to assess the relevance of each item on a 4-point scale (1 = content is not relevant, 2 = content is somewhat relevant, 3 = content is quite relevant, 4 = content is highly relevant). The content validity index (CVI) for each item is established by calculating the proportion of experts who judge it to be quite relevant or highly relevant, corresponding to a response of either 3 or 4. A minimum item CVI is 0.80 (Grant & Davis, 1997; Lindell & Brandt, 1999; Polit & Beck, 2006). However, there are exceptions as Lynn (1986) advocates for the minimum to vary as a function of the number of content experts.

Although the 0.80 CVI cut-off has good “face validity,” in that it represents situations when at least 80% of experts believe the item is quite relevant or better, it would be advantageous to have improved justification for the CVI cut-point. Another traditional measure used for quantifying content validity is the content validity ratio (CVR) attributable to Lawshe (1975) as described in McIntire and Miller (2007, p. 240) and Cohen and Swerdlik (2005). Like the CVI, the derivation of the CVR is based on the proportion of agreement or lack of agreement among raters or judges regarding how essential an item is in relation to the total scale. Application of the CVR involves calculating a ratio for each item comprising an instrument. The classic CVR (CVI minus 0.5 divided by 0.5) provides guidance on the minimum number of experts based on agreement due to chance, and as such provides various cut-points depending on the number of experts. Note the fixed 80% CVI cut-point for validity would be equivalent to a fixed cut-point of 60% for CVR, if there are 10 experts. Both the CVI and CVR are summaries of the number of experts choosing “quite relevant” or “highly relevant.” Using IACCV’s methodology as a foundation, we explore a new possible interpretation of content validity in this paper: experts’ opinion regarding each item-domain correlation.

**Problem Statement, Purpose, and Outline**

The conceptualization of “construct validity” in IACCV aligns with one aspect of how construct validity was defined traditionally, with a particular focus on “… understanding the relationship between measures and the constructs they represent.” (DeVellis, 2003, p. 14), using factor analysis. A single domain factor analysis can be quantified by the extent each of the items of a proposed instrument is correlated to the domain. For example, consider a single domain that is to be measured by eight proposed items, as shown in Figure 1. An appropriate number of participants (say 80) can be asked to answer the eight items. Once the data are collected, the correlation of the item to its domain (ρ’s) can be estimated using correlation between Item and Item-deleted total score (or corrected item-total score) (Nunnally & Bernstein, 1994). Model-based approaches such as confirmatory factor analysis or Item Response Theory (IRT) are also possible (Wirth & Edwards, 2007). More concretely, let Domain be a latent domain score and Item be a response for item 1. Thus participants provide information about Items which gives us estimates of ρ’s.

Before the items go to participants it is standard practice to assess the representativeness of these items using content experts. Although not obvious at first glance, information regarding experts’ opinion about the correlation of each item to its respective domain may be integrated within the typical 4-point relevancy content response options. For example, the concept for each domain might be stored in the mind of each expert because they are, ideally, familiar with the same literature used to generate the domains. We argue that the content experts provide direct information about the correlation of the item to its domain [ρ1 = corr(f, z1)], when answering the relevancy questions. More specifically, we assume that while experts answered the relevancy questions (typical content validity) these can be converted to the following correlation scale based on Cohen (1988). The scores 1 (not relevant), 2 (somewhat relevant), 3 (quite relevant), and 4 (highly relevant), respectively, can be converted to no correlation [0.0–0.10]; small correlation (0.10–0.30]; medium correlation (0.30–0.50]; and large correlation (0.50–1.00]. Details of the model of content experts’ data are illustrated in Section Model. The measurement model of the experts can be thought of as interval censored realizations of the respective ρ’s. As depicted in Figure 2,
suppose six experts are recruited for content validity. Then, the measurement model depicted in Figure 2 can help investigators discover a prior distribution for all items’ correlation to the domain.

Integrated analysis of content and construct validity uses a framework grounded in a fully Bayesian model where, as in long-standing, empirically verified, Bayesian analyses, experts’ data are treated as prior distributions updated to posterior distributions using participants’ data. Integrated analysis of content and construct validity has demonstrated utility with the posterior distribution, resulting in more stable estimates of usual psychometric parameters than traditional approaches (Gajewski et al., in press). In fact, Gajewski et al. implied that the use of IACCV allows one expert opinion to be equivalent in psychometric testing to data obtained from at least 10 participants. The implications of this result are great, as six content experts provide information equivalent to 60 participants. Thus, integrating experts and participants would increase the equivalent information for the hypothetical factor analysis from 80 to 140 participants.

While the previous case study (Gajewski et al., in press) provides promise for the IACCV approach, it is not clear whether the content data are transferable to a correlation statement; an assumption we wish to explore in this study, using an experiment. The purpose of this study was to validate the assumption that experts view relevance as equivalent to the correlation scale. Relevance is the classical interpretation of content validity data, whereas correlation data can provide a new and fresh approach to interpretation. Thus, we will provide an argument based on the assumption that experts conceptually think of classical content validity as correlations between items and a domain.

Deville (1996) stated, “Since the 1940s, measurement specialists have called for an empirical validation technique that combines content- and construct-related evidence.” Several researchers have attempted to quantify content analysis (Deville, 1996; Rubio, Berg-Weger, Tebb, Lee, & Rauch, 2003; Sireci & Geisinger, 1992, 1995). Classical multivariate methods such as multidimensional scaling and cluster and discriminant analysis were statistical approaches of choice. Our quantification of content analysis expands this literature to include learning about factor analysis parameters (item to domain correlations). In addition, the previous studies had small to moderate number of experts (3–15 experts). Our study uses a much larger number of experts (i.e., recruited 120 experts). This large number of experts was required because we designed and implemented a Bayesian, two-group, completely randomized, equivalency study (Lauzon & Caffo, 2009). Equivalence studies typically require larger sample sizes than classical superiority trials because acceptable differences between groups tend to be small (Kim, Buyon, Petri, Skovron, & Shore, 1999).

**Methods**

This section describes the large database used for this study, including sampling procedures, instruments, and statistical analysis methods.

**Setting**

Because equivalency trials tend to require a large number of subjects, we needed a large number of subject matter experts for our experiment. It was determined that a promising source for such subjects come from a project called the National Database of Nursing Quality Indicators® (NDNQI®). In 1998, NDNQI for acute care hospitals was established by the American Nurses Association to aid nurses in patient safety and quality improvement efforts by providing research-based, national comparative data on nursing care and the relationship of this care to patient outcomes Montalvo, 2007). The NDNQI has over 1,500 participating hospitals, ranging from small rural hospitals to large urban hospitals and academic medical centers, many of which hold Magnet® – designation status (Dupont & Montalvo, 2009). The NDNQI collects two types of data: (1) quarterly data on nursing-sensitive quality indicators (e.g., pressure ulcer rates, fall rates, and nosocomial infections) and (2) annual surveys of Registered Nurse (RN) job satisfaction and nursing work environment. Participating hospitals are provided unit-level comparison reports to state, national, and regional percentile distributions. In addition, participating hospitals have site coordinators to assist with data collection. The site coordinators that are also RNs were the content expert participants in this trial.

**Subjects**

We requested volunteers for the project from the NDNQI site coordinator pool. Site coordinators are NDNQI’s primary point of contact with each hospital, and are responsible for gathering and submitting data on nursing quality indicators. A recruitment e-mail was sent to 1,226 site coordinators. A total of 420 site coordinators responded to the recruitment e-mail. Of those who responded, 397 site coordinators were interested and eligible to participate and 16...
site coordinators were not. An additional seven site coordinators agreed to participate, but did not meet the eligibility requirements, either because they were not an RN or they were going to designate another individual, who was not a site coordinator, to complete the survey. Of the 397 eligible site coordinators, 120 site coordinators were randomly chosen to participate. The 120 NDNQI site coordinators were randomized into two balanced groups, the “relevance” group and the “correlation” group ($n_1 = n_2 = 60$ per group). We oversampled by 22 subjects, in order to improve our chances of obtaining the required 98 completed surveys, as described in the section on sample size calculations (Section Sample Size Calculations Equivalence Trial).

**Procedure**

We implemented a randomized equivalence trial with RN site coordinators randomly assigned to complete the “relevance” or “correlation” content tools, as described in Section Content Tools, regarding the RN Job Enjoyment (JE) Scale. The study was approved by the University of Kansas Medical Center’s Human Subjects Committee and Internal Review Board. The “relevance” and “correlation” tools were created using Survey Monkey (http://www.surveymonkey.com/), a Web-based software (see Appendixes A and B for tools). The participants received an e-mail, with a link to their respective online survey.

**Instruments**

**RN Job Enjoyment Scale**

Based on our experience with the scale, and its established psychometric properties, we identified the RN Job Enjoyment Scale (Taunton et al., 2004) as an appropriate tool for our use. All versions of the annual NDNQI RN surveys contain the JE scale, adapted from Brayfield and Rothe’s questionnaire (Dunton & Montalvo, 2009; Taunton et al., 2004). The scale consists of seven items intended to measure different aspects of JE, such as whether the nurses enjoy their jobs more than the average worker and whether the nurses would never accept another job opportunity. As part of NDNQI’s survey process, hospital-based RNs (non-site coordinators) complete the seven items using a Likert-type scale with possible responses ranging from strongly agree (6) to strongly disagree (1). Previous psychometric analysis of the JE scale has included exploratory factor analysis, confirmatory factor analysis, internal consistency reliability, concurrent validity, and workgroup level reliability studies (Boyle, Miller, Gajewski, Hart, & Dunton, 2006; Taunton et al., 2004).

**Content Tools**

The participants (site coordinators) were randomly assigned to two groups. The first group was named the “relevance” group, and was asked to rate the items on the usual relevance content scale of 1–4. The second group, or correlation group, was asked to rate the items’ content with explicitly defined “correlation” categories. Both the tools for the “relevance” group and the “correlation” group had the RN job enjoyment domain defined to them as “the extent to which nurses like their jobs, in general.”

Both groups were required to respond to a total of 11 items, eight items that needed to be rated on their “relevance” or “correlation” to the domain of JE and three items regarding the participants’ demographic information. The eight items that needed to be rated included the seven items from the JE Scale as well as a non-job enjoyment item that read “Nurses with whom I work would say that they are clinically competent.” Based on empirical evidence, this item should not be as relevant or correlate to the JE domain as the other seven items. “Correlation” was defined for the respective group as: “describing the strength of a relationship between two variables, which can range from 0 to 1.” The possible intervals were presented as: 0–0.10 is no correlation, 0.11–0.30 is a small correlation, 0.31–0.50 is a medium correlation, and 0.51–1.0 is a large correlation. Relevance was presented in the usual way (1 = not relevant; 2 = somewhat relevant; 3 = quite relevant; 4 = highly relevant). The participants then answered how well the eight items were “relevant” or “correlated” to JE.

No personal identifying information was collected from site coordinators. Participation was confidential and anonymous, but some demographic information was collected, including years of experience, how many years they had been employed as an RN on their current hospital unit, how many years they had worked as an RN in the United States, and their highest level of nursing education.

**Data Analysis Method**

The methodology described in this paper is useful for the analysis of psychometric equivalence studies, as well as for study design purposes. Mean comparisons were done between the two groups with different a priori distributions, according to an approximate model and an exact model.

**Model**

In the paragraphs that follow, we describe two models for analyzing the Bayesian equivalency trial. The first, called the exact model, provides an approach for understanding the causal relationship between the two groups on the correlation scale. The second model, the approximate model, places the outcome on equally spaced ordinal values. Although not on the correlation scale, this is a convenient model for design and sample size purposes.

The following is the model of expert opinion that follows the notation in Gajewski et al. (in press) for an equivalency trial with the two groups ($m = 1, 2$), where $m = 1$ is the “relevance” group and $m = 2$ is the “correlation” group. There are eight items assuming that $\rho_{ijkm}$ is the $k$th
expert's opinion for the jth item and define the transformation $g(\rho) = \frac{1}{2} \log \{ (1 + \rho)/(1 - \rho) \}$. This is a convenient alternative to Gajewski et al. (in press) logit transformation because it allows negative correlations (−1 to 1). Combining information from all experts and all items is:

$$g(\rho_{jkm}) = g(\rho_{jm}) + \epsilon_{jkm},$$

where $\epsilon_{jkm}$ is normally distributed with mean 0 and variance $\sigma^2$. The parameter $\rho_{jkm}$ represents the correlation between the item and the domain for expert k, which is a latent quantity in the model related to the elicited ordinal measure of association, $x_{jkm}$, as follows:

$$x_{jkm} = \begin{cases} 
1, & \text{if } 0.00 \leq \rho_{jkm} \leq 0.10 \\
2, & \text{if } 0.10 < \rho_{jkm} \leq 0.30 \\
3, & \text{if } 0.30 < \rho_{jkm} \leq 0.50 \\
4, & \text{if } 0.50 < \rho_{jkm} \leq 1.00.
\end{cases}$$

The parameter $\rho_{jkm}$ is the correlation between the item and domain, pooling over information from all experts within its respective mth group.

The group $m = 1$ is associated with elicitation of opinion on the relevance scale, which is classic content validity and the formulation originally defined by Gajewski et al. (in press). The thresholds of $\rho_{jkm}$ were argued theoretically based on the assumption that the experts learned statistical science from the behavioral sciences where Cohen (1988) suggested small, medium, and large correlations are 0.10 $\leq \rho_{jkm} \leq 0.30$; 0.30 $< \rho_{jkm} \leq 0.50$; and 0.50 $< \rho_{jkm} \leq 1.00$, respectively. Translating the “not relevant” response to less than a small correlation (i.e., 0.00 $< \rho_{jkm} \leq 0.10$) and incrementally increasing to equate the “highly relevant” response to a large correlation gives us thresholds for $\rho_{jkm}$.

The current experiment was designed to test the assumption that the thresholds are correct. That is, are experts actually thinking about relevancy on the correlation level? To test this idea, we asked a group ($m = 2$) for their opinion directly on the correlation level. That is, we defined the levels to be on the scale of Equation 2.

Note that these correlations were not transformed to an ordinal scale. Rather, the data already existed on an ordinal scale of correlation or relevance (each initially scored 1–4). In the correlation group, the experts’ interpretation is directly on the correlation scale. In the relevance group, we used an assumption regarding the experts’ interpretation of correlations to convert the original ordinal scale (relevance) to a meaningful scale (small, medium, and large correlations), and then we applied Cohen’s correlation cutpoints to build a model with interval censoring. The previous step advances assumptions about how experts think of correlations. Subsequently, testing this assumption proceeds by calculating how different the relevance group is from the correlation group—asking experts on the actual correlation scale.

The equations given above formulate the exact model because it properly models the uncertainty in these correlations induced by the interval censoring, similar to other methods developed such as pressure ulcer staging (ordinal 1–4) (Gajewski, Hart, Bergquist, & Dunton, 2007), audiology (Gajewski, Sedwick, & Antonelli, 2004; Gajewski, Nicholson, & Widen, 2009), and instrument development measuring nursing home quality of life (Gajewski, Boyle, & Thompson, 2010).

For the purposes of design, we also defined an approximate model, modeling the x’s from Equation 2 on the ordinal scale 1–4. More specifically:

$$x_{jkm} = \mu_{jm} + \epsilon_{jkm},$$

where $\mu_{jm}$ represents the mean response for the jth item from the kth expert in the mth group and $\epsilon_{jkm}$ is normally distributed with mean 0 and variance $\sigma^2$. The key question is if these means are within a certain value to deem them as equivalent.

For the approximate model we would like to demonstrate that

$$H_{ij} : |\mu_1 - \mu_2| < 0.5,$$

stating that for each of the items, the difference in the two means is within 0.5 on the (1–4) scale. Practically, this means that the two averages will always be within 0.25 correlation from one another, because it is half the largest gap from level 4 (0.50 $< \rho_{jkm} \leq 1.00$). Thus, for the exact model, the hypotheses will be written so as to demonstrate

$$H_{ij} : |\mu_1 - \mu_2| < 0.25,$$

placing the hypothesis on the exact scale, and that the correlations are exactly within 0.25 or less.

Posterior Calculation

The posterior calculation for the approximate model is a straightforward application of a Bayesian two-sample comparison (e.g., Gönen, Johnson, Lu, & Westfall, 2005). Using conjugate priors, this analysis can be done using spreadsheets such as Microsoft Excel.

The posterior calculation for the exact model is complicated by the fact that cut-points are modeled on the transformed scale with inference based on the untransformed scale. Therefore, Markov chain Monte Carlo (Gelman et al., 2004) is used to calculate the posterior distribution of $\rho_{jkm}$. We used WinBUGS (Gilks, Thomas, & Spiegelhalter, 1994) to perform the Bayesian modeling calculations. To allow a simulation estimate for the posterior distribution of $\rho_{jkm}$, a burn-in sample of 1,000 draws was used with the next 10,000 iterations for inference.

Priors

Prior distributions are shown in Table 1. For the approximate model we assume a normal distribution for $\mu_{jm}$. The mean of this normal distribution is the midpoint (2.5) and standard deviation equivalent to information from $n_{km} = 4$ subjects in each of the two groups. We set the standard deviation a priori to be 1 (based on data from Gajewski et al., in press). One can see that both of the mean parameters for each group, as well as the difference between them, covered a broad range of values. In fact, the a priori probability that
Table 1. Prior distributions for parameters for both the exact and the approximate models along with 95% Credible Intervals (CrI) and the probability of equivalency hypothesis being correct. Note that $\sigma_i = 1$ for the purposes of this summary but later is set to the observed standard deviation once data are collected.

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Distribution</th>
<th>Median (95% CrI)</th>
<th>Pr</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_{jm}$</td>
<td>$N(2.5, \sigma_j/\sqrt{4})$</td>
<td>2.5 (1.52, 3.48)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>$\mu_j - \mu_{j2}$</td>
<td>$N(0, \sigma_j/\sqrt{2})$</td>
<td>0.0 (-1.39, 1.39)</td>
<td>0.52</td>
</tr>
<tr>
<td>Approximate</td>
<td>$\sigma$</td>
<td>$1/\sigma^2 \sim U(0.01,100)$</td>
<td>0.50 (0.03, 0.98)</td>
<td>-</td>
</tr>
<tr>
<td>Exact</td>
<td>$\rho_{jm}$</td>
<td>$g(\rho_{jm}) \sim N(0.5493, 1/\sqrt{8})$</td>
<td>0.50 (-0.15, 0.85)</td>
<td>-</td>
</tr>
<tr>
<td>Exact</td>
<td>$\rho_j - \rho_{j2}$</td>
<td>Simulation</td>
<td>0.00 (-0.74, 0.74)</td>
<td>0.52</td>
</tr>
</tbody>
</table>

$H^*_j$ is true was equal to 0.52, reflecting indifference that the hypothesis is true.

The exact model’s prior distributions were set up as follows. Wilks (1962) shows that the large sample properties of the sampling distribution of correlation on the transformation $g(.)$ are unbiased with a variance of $1/n$. Therefore, we used this mathematical relationship to assist in formulation of the priors. We assumed that the prior information on this parameter is $n_{0m} = 8$ participants. As in the approximate model, this vague prior also corresponds to the probability that $H^*_j$ is true to be 0.52.

### Equivalence Analysis and Summary

After defining the prior distributions and collecting the data, the posterior calculations from WinBUGS result in $\Pr(H^*_j)$ and $\Pr(H_j)$ for the approximate and exact models, respectively. These calculations provide probabilities that the content groups are equivalent for each of the items. For both models we also calculate the posterior median (50 percentile) and 95% Credible Intervals (CrI) using the 2.5 percentile and 97.5 percentile for each of the parameters in the model.

### Sample Size Calculations Equivalence Trial

For the purposes of sample size calculations, we adhere to the notation and ideas described in Gajewski and Mayo (2006) for designing clinical trials, but adapted here for an equivalency trial. For each of the items in our clinical trial, we observed an approximately normal random variable $\bar{x}_{j1} - \bar{x}_{j2}$ with mean $\mu_j - \mu_{j2}$ variance $\sigma_j^2 (1/n_1 + 1/n_2)$. Because equivalency is assessed with Bayesian methods, there is a prior distribution for $\mu_j - \mu_{j2}$. In this trial we assume this prior is normal with mean 0 and variance $\sigma_j^2 (1/4 + 1/4)$, equivalent to a prior sample size of four in each group.

We needed to choose a sample size ($n_1 + n_2$) sufficient for a reasonable chance of finding a result such that $H^*_j$ (Equation 4) is true. It is assumed that the content methods are equivalent if the posterior probability that the $|\mu_j - \mu_{j2}| < 0.5$ is bigger than some level $\lambda$. In mathematical notation, content methods for item $j$ are equivalent if:

$$P_{\mu_j - \mu_{j2}} \left( |\mu_j - \mu_{j2}| < 0.5 \mid \bar{x}_{j1}, \bar{x}_{j2}, n_1, n_2, \sigma_j^2, 4, 4 \right) > \lambda.$$  (6)

We set the parameters on the right to be $A = (\bar{x}_{j1}, \bar{x}_{j2}, n_1, n_2, \sigma_j^2, 4, 4)$.

As in Gajewski and Mayo (2006) we questioned what value to choose for $\lambda$. A reasonable default value is 0.90 which is in the same spirit as usual power at 90%. We opted for this value in the current trial.

In order to calculate the sample size, we set $\bar{x}_{j1} - \bar{x}_{j2}$ equal to some value $<0.5$ and define $\sigma_j^2$. For the purposes here we set $\bar{x}_{j1} - \bar{x}_{j2} = -0.25$ and $\sigma_j^2 = 1$. The former was because we believe the differences are zero, but we allowed our answer to deviate halfway between 0 and the allowed bound (0.5), the latter coming from a prior content study (Gajewski et al., in press). Next, we needed to find the minimal integer $n = n_1 = n_2$ which satisfies Equation 6:

$$\min \left\{ n \mid P_{\mu_j - \mu_{j2}} \left( |\mu_j - \mu_{j2}| < 0.5 \mid \bar{x}_{j1} - \bar{x}_{j2} = -0.25, n, A \right) > .90 \right\}. \quad (7)$$

In order to facilitate the calculation of Equation 7 we note the posterior distribution $|\mu_j - \mu_{j2}| \bar{x}_{j1} - \bar{x}_{j2} = -0.25, n, A \sim N\left(-0.25, \sqrt{2}/(4 + n)\right)$. Therefore, the solution to Equation 7 was found by plotting the probability as a function of $\lambda$ (Figure 3) that indicates that $n = 49$ is the solution. To account for possible dropout we proposed to recruit $n = n_1 = n_2 = 60$.

### Sensitivity Analysis and Secondary Hypothesis

Equation 2 describes the Cohen-motivated unequally spaced values, which were presented to the correlation group. However, as pointed out by reviewer, the use of four category scales might represent equally spaced values. To test the sensitivity of this assumption, we will re-fit the exact model using equally spaced values:

$$x_{jkm} = \begin{cases} 1, & \text{if } 0.00 \leq \rho_{jkm} \leq 0.25 \\ 2, & \text{if } 0.25 < \rho_{jkm} \leq 0.50 \\ 3, & \text{if } 0.50 < \rho_{jkm} \leq 0.75 \\ 4, & \text{if } 0.75 < \rho_{jkm} \leq 1.00 \end{cases}. \quad (8)$$

Then, we will calculate $\Pr(H_j)$ to test the sensitivity of this assumption.

One reviewer questioned whether the exact model correlations predicted actual correlations. The JE questions have been collected on over 70,000 participants over one year (Gajewski, Boyle, Miller, Oberhelman, & Dunton, 2010).
Using this large dataset, we were able to calculate item-domain correlations and call them “actual” correlations ($q_{j0}$) because Wilks (1962) standard deviation estimate is quite small as it is $1/\sqrt{70000} = 0.0038$. Using these data we tested the hypothesis $H_2: |\rho_{jm} - \rho_{j0}| < 0.25$.

The hypothesis states that correlations from either the unequally spaced or equally spaced exact models are within acceptable ranges. We will calculate $Pr(H_2)$ to test the equivalence of predicted versus actual.

**Results**

**Summary Statistics and Demographics**

All of the site coordinators (participants) were RNs. The site coordinators had a wide variety of experience in their current job, with the plurality having been in their current job one to five years (40.4%) or greater than twenty years (24.8%). A vast majority (70%) of the site coordinators had been practicing as an RN in the United States for more than 20 years. Most of them were prepared with either a baccalaureate (37%) or master’s degree (47%). The others were either diploma/associate (9%), doctorate (6%), or did not respond (1%).

There were 59 RN site coordinators who responded out of 60 site coordinators who initially agreed for the “relevance” group ($m = 1$), and 50 out of 60 site coordinators for the correlation group ($m = 2$). Using a two-sample Beta-Binomial distribution with uniform priors, the posterior probability that the response rates for “relevance” group were larger than the correlation group was 0.9979 and 95% CrI in difference was (0.05, 0.26), strongly suggesting the correlation group had a significantly smaller response rate than the relevance group.

Table 2 provides the summary statistics by item for each of the two groups. The primary analysis was analyzed using the approximate model, which assumes the observed ordinal responses came from a normal distribution. However, of note in Table 2 is that many of the items are skewed to the left. This further argues for the validity of the exact model because the cut-point approach supports skewed distributions.

### Approximate Model Results

The results for all of the items analyzed using the approximate model are shown in Table 3. All of the items had posterior probabilities $Pr(H_2)$ greater than 0.90 except for items 5 and 7 which were just below the cut-point for equivalence (0.90). Thus, we may conclude that when treating the ordinal values as continuous variables using the relevance or the correlation content models will provide equivalent information.

### Exact Model Results

It was projected that the inconclusive nature of any two variables in the approximate model would be addressed when analyzing using the exact model. This was indeed the case.

Table 2. Summary statistics for content validity for both the “relevance” group ($m = 1$) and the “correlation” group ($m = 2$). The “relevance” group had $n_1 = 59$ subjects on the scale 1 = not relevant to 4 = highly relevant and the “correlation” group had $n_2 = 50$ subjects on the scale 1 = 0.00–0.10 to 4 = 0.50–1.00.

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>$m = 1$</th>
<th>$m = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1 (%)</td>
<td>2 (%)</td>
</tr>
<tr>
<td>1</td>
<td>Are fairly well satisfied with their jobs</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>Would not consider taking another job</td>
<td>3</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>Have to force themselves to work much of the time</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Are enthusiastic about their work almost every day</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>Like their jobs better than the average worker does</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>6</td>
<td>Are clinically competent</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>Feel that each day on their job will never end</td>
<td>7</td>
<td>25</td>
</tr>
<tr>
<td>8</td>
<td>Find real enjoyment in their work</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
as shown in Table 4, where Pr(H1) > 0.99 for all of the items under both unequally and equally spaced models. Thus, for the purposes of our definition of equivalent (within 0.25), both the relevance and the correlation groups provide the same information. However, the unequally and equally spaced models diverged when testing equivalence of these models to actual correlations (ρm). For the unequally spaced model, the posterior distributions were not equivalent to actual (most Pr(H2) < 0.26). However, for the equally spaced model the correlations are equivalent (Pr(H2) > 0.99).

These results can be further explored by inspecting the posterior distributions of the correlations shown in Figures 4 and 5, unequally and equally spaced models, respectively. While all of the items are equivalent, one can see that most of the separation did, in fact, take place for items 5 and 7. The 95% CrIs in Table 4 indicate the upper CrI is around 0.15 for items 5 and 7, respectively, for both models. Indeed, item 6 also did not fit with JE validity and is reflected by its relatively low posterior distribution. Also indicated in both graphs further illustrates the divergence of the unequally spaced model from the actual model and the convergence of equally spaced model toward the actual.

Table 4. Results for the exact model – using both unequally spaced and equally spaced assumptions – for the relevance group (m = 1) versus the correlation group (m = 2) using informative prior – Equation 4. Pr(H1) = Pr(|μ1 − μ2| < 0.25). Note that the median (95% CrI) for σ was 0.24 (0.22–0.26). ρ0 is the actual correlation. Pr(H2) = Pr(|μm − μp| < 0.25).

<table>
<thead>
<tr>
<th>#</th>
<th>E(ρ1)</th>
<th>sd(ρ1)</th>
<th>E(ρ2)</th>
<th>sd(ρ2)</th>
<th>E(ρ1 − ρ2)</th>
<th>2.5% percentile</th>
<th>97.5% percentile</th>
<th>Pr(H1)</th>
<th>Pr(H2)</th>
<th>ρ0</th>
<th>m = 1</th>
<th>m = 2</th>
</tr>
</thead>
</table>
| Unequally spaced
| 1 | 0.56 | 0.03 | 0.58 | 0.03 | 0.02 | −0.05 | 0.09 | 1.00 | 0.83 | 0.03 | 0.03 |
| 2 | 0.45 | 0.03 | 0.47 | 0.03 | 0.02 | −0.06 | 0.10 | 1.00 | 0.69 | 0.26 | 0.26 |
| 3 | 0.49 | 0.03 | 0.52 | 0.03 | 0.03 | −0.04 | 0.11 | 1.00 | 0.70 | 0.94 | 0.95 |
| 4 | 0.54 | 0.03 | 0.57 | 0.03 | 0.03 | −0.04 | 0.10 | 1.00 | 0.83 | 0.00 | 0.00 |
| 5 | 0.42 | 0.03 | 0.50 | 0.03 | 0.07 | 0.00 | 0.15 | 1.00 | 0.84 | 0.00 | 0.00 |
| 6 | 0.35 | 0.03 | 0.36 | 0.03 | 0.00 | −0.08 | 0.09 | 1.00 | 0.43 | 1.00 | 1.00 |
| 7 | 0.44 | 0.03 | 0.50 | 0.03 | 0.06 | −0.01 | 0.14 | 1.00 | 0.65 | 0.91 | 0.91 |
| 8 | 0.62 | 0.03 | 0.61 | 0.03 | 0.00 | −0.08 | 0.07 | 1.00 | 0.81 | 0.77 | 0.77 |
| Equally spaced
| 1 | 0.81 | 0.02 | 0.82 | 0.02 | 0.03 | −0.05 | 0.08 | 1.00 | 0.83 | 1.00 | 1.00 |
| 2 | 0.70 | 0.03 | 0.72 | 0.02 | 0.04 | −0.06 | 0.10 | 1.00 | 0.69 | 1.00 | 1.00 |
| 3 | 0.74 | 0.03 | 0.77 | 0.03 | 0.04 | −0.04 | 0.11 | 1.00 | 0.70 | 1.00 | 1.00 |
| 4 | 0.79 | 0.02 | 0.82 | 0.03 | 0.03 | −0.04 | 0.09 | 1.00 | 0.83 | 1.00 | 1.00 |
| 5 | 0.67 | 0.03 | 0.75 | 0.08 | 0.04 | 0.00 | 0.16 | 1.00 | 0.84 | 0.99 | 0.99 |
| 6 | 0.59 | 0.04 | 0.59 | 0.00 | 0.05 | −0.10 | 0.11 | 1.00 | 0.43 | 1.00 | 1.00 |
| 7 | 0.69 | 0.03 | 0.75 | 0.07 | 0.04 | −0.01 | 0.15 | 1.00 | 0.65 | 1.00 | 1.00 |
| 8 | 0.86 | 0.02 | 0.85 | 0.00 | 0.03 | −0.06 | 0.05 | 1.00 | 0.81 | 1.00 | 1.00 |
model was straightforward to implement for analysis. The exact model was advantageous because of more precise estimates on the correlation scale – the scale in which we believe investigators have the most interest. An advantage of the approximate model is that it appears to be a conservative means for calculating sample sizes. Thus, if a study sample size is justified from the approximate model, investigators will have large enough sample sizes for the experiment using the exact model.

The assumption in IACCV that the scores 1 (not relevant), 2 (somewhat relevant), 3 (quite relevant), and 4 (highly relevant), respectively, correspond to no correlation \([0.0–0.10]\); small correlation \([0.10–0.30]\); medium correlation \([0.30–0.50]\); and large correlation \([0.50–1.00]\) was assessed by this equivalency trial. The sensitivity of the unequally spaced assumption was tested by also fitting the equally spaced model. Under either assumption (e.g., unequal or equal), the two groups were found equivalent. Unexpectedly, the equally spaced model was closer to actual correlations than the unequally spaced model, indicating experts might not be paying close attention to the response options. Notably, the results clearly indicate promise that an equally spaced assumption on the correlations is substantiated. This evidence provides us with an indication that a new interpretation of the CVI is validated. From the proportion of experts choosing “relevant,” we can infer the proportion of experts who score items as having a correlation. In fact, if the appropriate experts are chosen, we now can use traditional content validity to tell us the “... number of discriminating items” (Nunnally & Bernstein, 1994, p. 305) before we give the instrument to participants. Of course, this content-based estimate could also be updated using Bayesian techniques such as IACCV (Gajewski et al., in press).

Figure 4. Results for the unequally spaced assumption exact model for the relevance group \((m = 1)\) versus the correlation group \((m = 2)\) using informative prior – Equation 4. The spikes are actual correlations from 70,000+ participants (Gajewski et al., 2010).
The results of the current study indicate the site coordinators carefully considered relevancy and correlation as equivalent concepts, and so traditional content validity tools not only enjoy the “relevance” interpretation but can also be interpreted as a correlation. However, it appears that the site coordinators had a downward bias relative to the “truth.” In fact, the item expected to exhibit the lowest CVR, and did, was item 6, since it is not a valid measure of JE, which is reflected by CVR = 0.05 and 0.20 (relevance and correlation groups, respectively) (Table 5).

For thoroughness we used the approximate model and the exact model to calculate model-based CVR, called CVR\(_a\) and CVR\(_e\), respectively, using the models’ respective posterior predictive distributions (Gelman et al., 2004). Consistent with the results of the two models and the ordinal-skewed nature, the exact model reflects the observed data (CVR) much more closely (Table 5).

Our work also has implications for the Bayesian literature. O’Hagan et al. (2006) indicate, “elicitation of prior information is accepted as having a fundamental role in Bayesian statistics.” Specifically, there are some consistencies between our work and the work by Clemen, Fischer, & Winkler (2000). Clemen et al. (2000) compared the success rates of eliciting opinion regarding the correlation between height and weight, where they compared six different approaches ranging from Likert scale, two of which were direct specification of correlation and conditional distributions. Our elicitations of correlations were on a Likert-type scale. Additionally, our work was complicated in that we were asking about several correlations, although we requested correlations within one domain.

The idea of using content information to elicit correlations might be useful beyond instrument development. For example, suppose one is eliciting a general multivariate correlation.

Figure 5. Results for the equally spaced assumption exact model for the relevance group \(m = 1\) versus the correlation group \(m = 2\) using informative prior – Equation 4. The spikes are actual correlations from 70,000 + participants (Gajewski et al., 2010).
A reasonable approach is to model the multivariate correlation through a single domain, and then elicit variable-domain correlations as pairs at a time. An approach such as the one used here is less burdensome than eliciting a general correlation matrix, which requires 28 different correlations and is susceptible to computational issues such as a non-determinant correlation matrix. In our single domain case one could elicit, for example, two correlations \( \rho_j \) and \( \rho_r \) and multiply them to gain an estimate of the correlation between \( j \) and \( r \), reducing the 28 pairs down to 8. Requiring a small “prior” sample size would minimize problems with this approach being too structured for a general correlation matrix.

The results have implications for both Bayesian psychometric analysis and content validity. Generally speaking, Bayesian analysis has shown good statistical properties and can be more efficient than classical analysis. However, this only works if the priors are “valid.” Our analysis suggests content validity can be used to construct a valid prior. Therefore, Bayesian psychometrics using expert opinions shows great promise. Our original intent of this study was to substantiate the claim that content validity is more quantitative than once believed. Previous studies (Deville, 1996; Sireci, & Geisinger, 1992, 1995) indicated experts and participants do indeed cluster items similarly. We take this a step further and argue for evidence that experts and participants also supply information about correlations in a similar fashion.

However, common sense trumps our interpretation of correlation when performing content validity. As an example from reviewer, consider a mastery testing situation, where the goal is to determine whether or not key concepts have been acquired. Content validity experts could consider a basic item on photosynthesis to be “highly relevant” to a botany test. However, the correlation between this item and domain could be very low since nearly all students could answer the item correctly. For some items the correlation depends in part on the population of students; the relevance of the content does not. This argues for always collecting participant data.

There are limitations to our study that have implications for future research. First, item three, “Have to force themselves to work much of the time” and item seven, “Feel that each day on their job will never end” are “negatively” worded items. Items with negative wording have the potential to confuse experts. Such confusion is especially relevant for those in the correlation group who may have been searching for a negative correlation. Specifically, we received a comment from one expert stating this possible misunderstanding. To clarify such misunderstandings in future studies we included a line in the instructions stating “When giving your answers, place them on a positive scale. For example, if you believe a correlation is −.50, select the answer that corresponds to .50.”

A second possible limitation was that the “correlation” group had significantly fewer responses than the “relevance” group, indicating the “correlation” group of participants likely felt uncomfortable about giving an opinion regarding statistical measures such as a correlation. Another possible interpretation is the correlation group may have indicated the lack of understanding in interpretation of the correlation and the instrument. However, this effect size was not strong.

A third possible limitation is that the two tools may agree for spurious reasons. For instance, raters may have ignored the category labels on the rating scales making the rating task much the same for both groups. Dunham and Davison (1991, p. 23) discuss evidence that raters sometimes pay little attention to labels. To address this problem, a future experiment could be designed in which a third group is
added without labels to see if the results change. If results do not change, then the group is not paying attention to the labels.

A fourth possible limitation is that we provided very little training for “experts” to verify that the correlation group actually understood the interpretation of a correlation coefficient. The only instruction given to the correlation group was to define a correlation as “describing the strength of a relationship between two variables, which can range from 0 to 1.” The next set of experiments will focus on verifying the experts’ understanding of a correlation, providing more training, or both.

Conclusion

We have support for our hypothesis that the two tools are equivalent. A Bayesian ordered analysis model supported the results, suggesting traditional content validity can be justified using correlation arguments and providing direct information regarding construct validity. We anticipate replicating this experiment using other psychometric instruments and/or experts with more statistical training or more knowledge about the content area. A major caveat to this conclusion is the possibility of spurious results, which can be explored in future experiments using the design and analysis methods presented in this paper.

Our study suggests great implications on psychometric modeling and the use of content validity in CTT, IRT, and factor analysis. However, the study limitations suggest the need for further work. For example, while both the Cohen-based (unequal spaced) correlations and equally spaced correlations are equal across groups, evidence suggests that equal spaced is closer to the truth. This was a secondary finding in the current study. Additional work to confirm this finding could be based, for example, on a set of prospective experiments with equally spaced correlations among experts, using instruments with more domains and more items. Results from these studies could demonstrate the potential benefits of Bayesian-based methods, such as IACCV, to psychometric testing.

Using the statistical tools for experimental design in equivalence studies presented in this paper, we provide a methodological springboard for future research. Our results suggest content validation is much more informative than originally thought — in that it actually is providing construct information. However, this conclusion should be viewed in light of possible limitations of our experiment. Nevertheless, the design and analysis methods presented provide a clear path for future experiments across a wide range of experts and situations. Globally, this research program is very promising in establishing new and efficient methods for determining content and construct validity.

Acknowledgments

We thank Kim Boyle for the recruitment of NDNQI site coordinators. Partial support of the work from Gajewski, Boyle, Bott, and Dunton came from a contract with the American Nurses Association (PI Dunton). Coffland’s support was provided through a Bridging Grant awarded by the University of Kansas Research Institute (PI Gajewski). J. Leopold used experience on this study as partial fulfillment of “research experience” as a graduate student in KU School of Nursing. We thank K Reeder for editorial review and comment of first revision of this paper.

References


Received May 4, 2010
Accepted February 9, 2011
Published online August 5, 2011

Byron J. Gajewski

The University of Kansas Medical Center
Department of Biostatistics
Mail Stop 1026
3901 Rainbow Blvd.
Kansas City
KS 66160
USA
Tel. +1 913 588-1603
Fax +1 913 588-0252
E-mail bgajewski@kumc.edu
Appendix A
Relevance Group

Job Enjoyment Scale Validity- Relevance

Job Enjoyment Scale Validity Survey Instructions

1. Read and review the definition of Job Enjoyment.

2. Make note of the response options for each item on the Job Enjoyment Scale and consider how they play into the relationship of the item to Job Enjoyment. The item response choices are: Strongly Agree, Agree, Tend to Agree, Tend to Disagree, Disagree, Strongly Disagree

3. Read each item and indicate how relevant you believe the item is to Job Enjoyment. The rating scale ranges from "is not relevant" to "is highly relevant" for each item. Select the corresponding answer to your choice in the right hand column.

Definition of Job Enjoyment:
The extent to which nurses like their jobs in general.

Consider the associated item response choices when evaluating each item:
Strongly Agree, Agree, Tend to Agree, Tend to Disagree, Disagree, Strongly Disagree.

**Please remember, we are not asking for your response to the item, but rather your rating of the item’s relevance to Job Enjoyment.**

1. Nurses with whom I work would say that they:

<table>
<thead>
<tr>
<th>Item is not relevant to Job Enjoyment</th>
<th>Item is somewhat relevant to Job Enjoyment</th>
<th>Item is quite relevant to Job Enjoyment</th>
<th>Item is highly relevant to Job Enjoyment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are fairly well satisfied with their jobs.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would not consider taking another job.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have to force themselves to work much of the time.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are enthusiastic about their work almost every day.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Like their jobs better than the average worker does.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are clinically competent.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feel that each day on their job will never end.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find real enjoyment in their work.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B

Correlation Group

Job Enjoyment Scale Validity- Correlation

Job Enjoyment Scale Validity Survey Instructions

1. Read and review the definitions of Job Enjoyment and Correlation.

2. Make note of the response options for each item on the Job Enjoyment Scale and consider how they play into the relationship of the item to Job Enjoyment. The item response choices are: Strongly Agree, Agree, Tend to Agree, Tend to Disagree, Disagree, Strongly Disagree

3. Read each item and indicate how correlated you believe the item is with Job Enjoyment. The rating scale ranges from "is not correlated" to "has a large correlation" for each item. Select the corresponding answer to your choice in the right hand column.

Definition of Job Enjoyment:
The extent to which nurses like their jobs in general.

Definition of a Correlation:
A correlation describes the strength of a relationship between two variables and can range from 0 to 1.
The strength of the relationship is based on these cutoffs:
0-10 = no correlation
.11-.30 = small correlation
.31-.50 = medium correlation
.51-1.0 = large correlation

Consider the associated item response choices when evaluating each item:
Strongly Agree, Agree, Tend to Agree, Tend to Disagree, Disagree, Strongly Disagree.

**Please remember, we are not asking for your response to the item, but rather your rating of the item’s correlation with Job Enjoyment.**

Job Enjoyment Scale Validity- Correlation

<table>
<thead>
<tr>
<th>Item is not correlated with Job Enjoyment (0-10)</th>
<th>Item has a small correlation with Job Enjoyment (.11-.30)</th>
<th>Item has a medium correlation with Job Enjoyment (.31-.50)</th>
<th>Item has a large correlation with Job Enjoyment (.51-1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Are fairly well satisfied with their jobs.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Would not consider taking another job.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Have to force themselves to work much of the time.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are enthusiastic about their work almost every day.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Like their jobs better than the average worker does.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are clinically competent.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Feel that each day on their job will never end.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Find real enjoyment in their work.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix C

Example WinBUGs Code for Exact Model

model
{
  for (i in 1:8)
  {
    for (j in 1:59)
    {
      Trho[i,j] ~ dnorm(mu1[i], invsige2) I(TrhoL[i,j], TrhoU[i,j])
    }
    TrhoP1[i] ~ dnorm(mu1[i], invsige2)
    mu1[i] ~ dnorm(0, 0.0001)
    rho1[i] <- (exp(2*mu1[i])-1)/(exp(2*mu1[i])+1)
    rhoP1[i] <- (exp(2*TrhoP1[i])-1)/(exp(2*TrhoP1[i])+1)
    CV1[i] <- step(-.3+rhoP1[i])
    CVR1[i] <- (CV1[i] -.5)/.5
  }
  for (i in 1:8)
  {
    for (j in 60:109)
    {
      Trho[i,j] ~ dnorm(mu2[i], invsige2) I(TrhoL[i,j], TrhoU[i,j])
    }
    TrhoP2[i] ~ dnorm(mu2[i], invsige2)
    mu2[i] ~ dnorm(0, 0.0001)
    rho2[i] <- (exp(2*mu2[i])-1)/(exp(2*mu2[i])+1)
    rhoP2[i] <- (exp(2*TrhoP2[i])-1)/(exp(2*TrhoP2[i])+1)
    CV2[i] <- step(-.3+rhoP2[i])
    CVR2[i] <- (CV2[i] -.5)/.5
  }
  invsige2 ~ dgamma(0.001, 0.001)
  sige <- 1/sqrt(invsige2)
  invsige26 <- 6*invsige2
  for (i in 1:8)
  {
    Diff[i] <- rho2[i] - rho1[i]
    P1[i] <- step(-abs(Diff[i])+.25)
    Diff1[i] <- rho1[i] - rhoTRUTH[i]
    P12[i] <- step(-abs(Diff1[i])+.25)
    Diff2[i] <- rho2[i] - rhoTRUTH[i]
    P22[i] <- step(-abs(Diff2[i])+.25)
  }
}

# Initial Values
list(
  invsige2 = .2,
  Trho = structure(.Data = c(<place values on transformed scale>), .Dim = c(8,109)),
)

# Data
list(rhoTRUTH = c(0.8640, 0.7230, 0.6980, 0.8880, 0.9000, 0.4500, 0.6540, 0.8570),
     TrhoL = structure(.Data = c(<place lower values on transformed scale>), .Dim = c(8,109)),
     TrhoU = structure(.Data = c(<place upper values on transformed scale>), .Dim = c(8,109))
)