21st International Congress on Modelling and Simulation, Gold Coast, Australia, 29 Nov to 4 Dec 2015
www.mssanz.org.au/modsim2015

Group Assessment of Interview Ready Model Reliability

Andrew Coutts

Defence Science and Technology Group
Email: Andrew.coutts@dsto.defence.gov.au

Abstract: Developing decision-aiding models to support national security decision making is challenging due to constrained access to high profile subject matter experts (SME). In such cases the necessary model building interviews with SMEs are rare and time critical. Interview ready models (IRM), that is, preliminary models that convey an initial view of the problem, provide a powerful means to extract the maximum benefit from these interviews. However there are risks associated with this strategy. Of greatest concern is that an unreliable model could lead, constrain or distract the interviewee, thereby negating its benefits and undermining the validity of the resulting decision-aiding model. Consequently, there is a need to extend the concept of model validity to the construction of the IRM. In the case of interview ready causal models, possible causal associations between concepts (model nodes) may need to be identified via textual sources using a subjective coding scheme such as content analysis. The reliability of the coding scheme essentially governs the logical validity of the resulting IRM. The process of assessing coding reliability is further complicated when a single analyst/model builder conducts the coding and analysis used to construct the IRM. This paper examines the argument for employing an IRM, reviews the literature regarding model reliability for such models, proposes an approach to assess the reliability of an IRM constructed using content analysis that balances adequacy and feasibility and applies this approach to a case study.

Keywords: *Interview ready model, decision-aiding models, model validity, content analysis*

1. INTRODUCTION

In some decision support applications, decision-aiding models (Gass, 1983) must be constructed via interviews with high profile stakeholder subject matter experts (SME) with limited time and no immediate decision problem to solve. For example, due to the crisis nature of such situations, decision-aiding models for political strategic decisions may need to be constructed ahead of some future decision based on a detailed understanding of how the decision system has previously operated (Coutts, 2013). Such modelling problems are at best poorly defined and therefore these interviews are critical to the model building process to confirm model structure, key variables and relationships. Consequently, increasing the efficiency and effectiveness of these time-constrained interviews is critical to the success of model building under these conditions.

A number of problems combine in limiting the effectiveness of model building interviews with time constrained stakeholders. Firstly, with complex policy problems, such as political-strategic decision making, not everyone can know the full system (J. A. Vennix et al., 1988). Thus, data elicitation must involve multiple stakeholders holding specialist knowledge in different areas. However, such stove-piped knowledge elicitation can lead to biases in model development (J. A. Vennix et al., 1988, p. 420). While eliciting data widely from many stakeholders may mitigate these biases it can cause difficulties in structuring the acquired knowledge (J. A. Vennix et al., 1988, p. 421) resulting in incoherency in model building. Secondly, even experienced stakeholders' knowledge of a complex decision system can be tacit rather than explicit (Ford and Sterman, 1998). Building formal models requires explicit knowledge of the system and hence the modeller needs a means to capture this explicit information from a stakeholder's tacit mental models. Such elicitation needs to deliver detailed descriptions of key factors, variables and relationships between them. However, eliciting such data 'from scratch' can lead to a 'blank page' effect (Joyce, 2009) in which a lack of boundaries constrains rather than liberates creativity and hence data elicitation. Joyce claims that "a moderate degree of constraint could help turn the blank page into a tractable creative challenge" (Joyce, 2009, p. 75).

One promising method in the literature to address these problems and improve the effectiveness of model building interviews is through the use of a preliminary model (Ford et al., 1998) constructed as an interview-ready model (IRM) and based on available knowledge without the benefit of direct input from key stakeholders. An IRM provides a robust start point for the interview, helps the interviewee to understand the problem the analyst is trying to solve and assists the interviewer to focus on key areas of uncertainty and model structure. An effective IRM must therefore be descriptive while also containing sufficient detail and knowledge of the problem to both demonstrate the interviewer's credibility (Hartley, 1969) by their preparation and to engage the interviewee's interest and comment (Ford et al., 1998). Such models will likely be constructed subjectively based on the model builder's understanding of the situation and interpretation of available information sources with a view to improving the model through exposure to SMEs and by formal validation. They capture a basic understanding of the system of interest, provide an indicative structure of the desired model and indicate initial boundaries of the problem. This in turn provides a start point, rather than a 'blank page', for stakeholders to rapidly engage with the model building task, presents a common, if speculative, view of the problem area and clarifies key areas of knowledge requiring review, modification or comment. The IRM exists to encourage a response from interviewed stakeholders leading to more detailed information, modification and possibly complete rejection of particular areas of the preliminary model.

Despite these benefits, there are concerns over the use of an IRM to structure interviews. Of greatest concern is that a poor (or invalid) model could lead (J. Vennix, 1999), constrain or distract the interviewee, thereby negating its' benefits and undermining the validity of the resulting formal decision-aiding model. These concerns are heightened where an IRM is constructed through subjective processes by a single modeller rather than a group. This raises IRM validity as an issue separate to the validity of the final decision-aiding model. For IRM constructed using textual sources, this validity is strongly linked to coding reliability.

This paper considers the problem of managing the validity of an IRM constructed via textual coding, establishes an approach for assessing coding reliability and applies it to a case study involving an IRM constructed by a single modeller. A preferred IRM validation approach should be both adequate for the modelling purpose and feasible within the available resources for the modelling task (Hossain et al., 2013).

2. VALIDATION OF DECISION-AIDING MODELS

A number of authors address validity for decision-aiding models. Gass (1983) identified decision-aiding models as the core scientific contribution of operations research (OR) and provided guidance for their validation. Landry et al. (1983) discussed the principles of building and validating OR models. Barlas (1996) identified principles for the validation of decision-aiding system dynamics models, claiming that any judgement of a model's validity implies a judgement of its purpose which is essentially a subjective and qualitative process. He categorises decision-aiding models as either 'causal-descriptive' models, described as theory-like 'white-box' models, or purely correlational 'black-box' models. The IRMs discussed in this paper are a form of 'white-box' model as they embody a series of causal statements about how real world processes actually work. 'White-box' model validity requires an assessment of internal structure in addition to model output. This recognition of the importance of balancing model representativeness - the extent to which the model matches the actual system in structure and mechanism - and usefulness is central to the five stages of model validation proposed by Landry et al. (1983): Conceptual Validation; Logical Validation; Experimental Validation; Operational Validation; and Data Validation.

2.1. IRM Validation Requirements

Barlas' (1996) observations on model validity and the validity framework proposed by Landry et al. (1983) are generally applicable to the IRM problem. From Barlas (1996) it holds that IRM validation should be cognisant of its purpose and seen as a part of the wider modelling process towards creating a valid formal model and hence not all validation stages are necessary. Two of the validation stages outlined by Landry et al. (1983) appear relevant for IRM validity and to mitigate the risks identified previously in their use: conceptual and logical validation. That is, if we establish a level of confidence that the IRM portrays the problem and main variables of interest from a perspective acceptable to a knowledgeable SME (conceptual validation) and that the relationships between variables within the model are consistent with the available evidence (logical validation) we accept that the model is interview ready[1]. To achieve conceptual validation, a means is required to confirm the plausibility of how the problem situation is represented in the conceptual model from the perspective of someone with sufficient relevant knowledge and/or experience. In some situations where access to key stakeholders is limited, this may require identifying a surrogate stakeholder.

Like verification (Landry et al., 1983), logical validation confirms the translation or encoding of relationships identified in the conceptual model as they are implemented in the IRM. Borrowing a term from qualitative research, this essentially establishes the reliability of the IRM and involves a structured review of the available evidence regarding variable relationships and how they were incorporated into the model. For IRM based on subjective interpretation of data (e.g. relationships identified in textual sources), reliability requires comparison with other interpretations of the same data. Where a single modeller (as opposed to a group) conducts these subjective interpretations, this review process is more challenging[2]. Such situations require some form of external peer review preferably using multiple reviewers to establish a level of confidence in the process used to form the IRM. The following case study examines IRM validation in such a situation.

3. IRM VALIDATION CASE STUDY

The National Political-Strategic Decision (NPSD) study (Coutts, 2013) investigated the factors that have influenced Australian political-strategic decisions on the foreign deployment of Australian military force to inform the construction of a decision-aiding model to support future decision making. An IRM was deemed necessary as part of the model building process due to limited interview time available with key stakeholders and the need to maximise the benefits of these interviews. The preliminary model structure was developed through analysis of relevant decision maker statements in textual sources made at the time of the decisions using directed content analysis (Hsieh and Shannon, 2005). The initial concepts (model variables) required for a directed content analysis were based on decision maker objectives and influences identified for a range of similar strategic decisions (Coutts, 2013). The way in which directed content analysis was applied to link variables via causal statements in the IRM is similar to the method detailed by Nadkarni and Shenoy (2004).

From Section 2, IRM validation requires both conceptual and logical validation (Figure 1). Conceptual validation confirms that, in the consulted surrogate SME's opinion, the model reasonably represents the problem situation and that the model structure adequately captures key model variables. The approach used for the NPSD study (Coutts, 2013a) is referred to as "face validation" (Pace and Sheehan, 2002) and uses a suitably knowledgeable academic SME in place of a stakeholder SME. Logical validation for a model structure based on content analysis can be equated to the reliability of the coding process. Reliability in turn has two components: stability and reproducibility (Stemler, 2001). Stability here is the ability of a single coder to produce the same results in recoding the same text while reproducibility refers to the level to which different coders code the same texts in the same way. Before describing how reliability testing was conducted in the NPSD study, it is necessary to briefly explain the directed content analysis process as applied therein.

[1] Aspects of data validity (mental/written sources) are subsumed into conceptual and logical validity.

[2] Such modelling, based on a form of thematic or content analysis, is often conducted in a group setting.

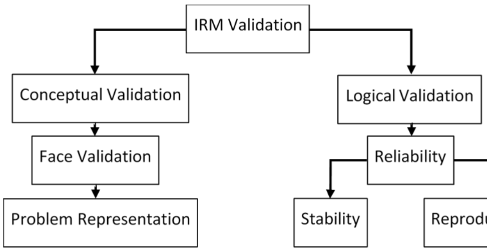

Figure 1. IRM validation requirements for model structured by content analysis

3.1. Directed Content Analysis Process in the NPSD Study

Figure 3 details the way in which directed content analysis was implemented in the NPSD study. Key elements in this diagram are explained in the following sub-paragraphs.

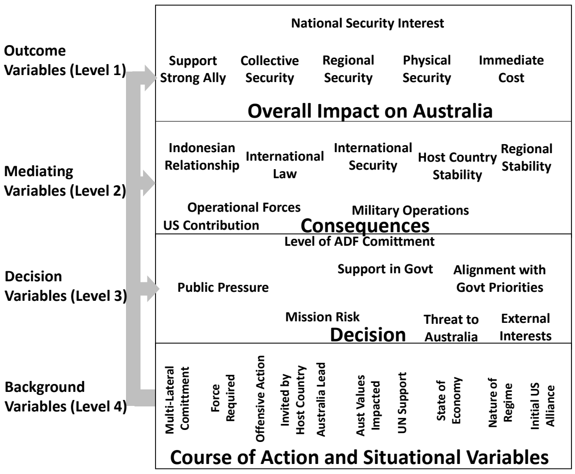

Figure 2. NPSD Study Conceptual Mode

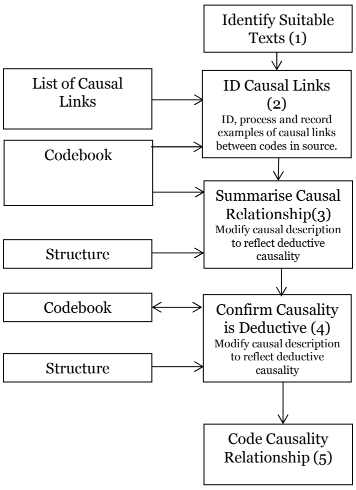
Identify Suitable Texts (1)
List of Causal Links
ID Causal Links (2)
ID, process and record examples of causal links between codes in source.
Codebook
Summarise Causal Relationship(3)
Modify causal description to reflect deductive causality
Structure
Codebook
Confirm Causality is Deductive (4)
Modify causal description to reflect deductive causality
Structure
Code Causality Relationship (5)

Figure 3. Directed Content Analysis Process

a. **List of Causal Links**: This is a list of possible textual causal connectors that will be used to identify causal connections in a text (Nadkarni and Shenoy, 2004), including phrases and single words that imply a causal relation between two concepts such as: "leads to" or "impacts on".

b. **Codebook**: A 'codebook' (MacQueen, et al., 1998) contains concepts, definitions and rules that allow codes to be mapped to and represent large portions of text through the process of content analysis.

c. **Structure/Conceptual Model**: A high level model structure both represents what is known about the problem and the form the model should take. Four hierarchical variable categories (Figure 2) were identified for the study based on a framework proposed by Kjaerulff and Madsen (2008) : Background; Decision; Mediating; and Outcome. Added variables are assigned to one of these categories.

Following the process outlined in Figure 3, the analyst reviews selected texts searching for passages that contain a code (or synonym) associated with a causal link (steps 1 and 2 in Figure 3). The code may be a cause (or parent) for some effect (or child) variable or the reverse. The text is summarised to align it with existing codes and causal links (step 3). A new causal link and/or variable may be discovered through this process and added and defined in the associated record. For example, in reviewing a text on Australia's 1999 decision on East Timor from the NPSD study below, the analyst identified the following phrase of interest.

> "They are going at the request of the United Nations" (Commonwealth of Australia, 1999, p. 10025)

Using the process outlined in Figure 3, this was coded as:

Causal Phrase	**Causal Connector**	**Effect Phrase**
United Nations	Request	ADF to Deploy to East Timor

This process results in a set of causal relationships between variables within the IRM. The purpose of IRM logical validation is therefore to confirm the stability and repeatability of this process.

3.2. Assessing the Stability of the Coding Process

The NPSD study was conducted as part of a doctoral thesis by a single analyst who was solely responsible for establishing and conducting the coding scheme and analysing its results. As such, assessing the consistency of how this analyst would code the same text multiple times was unlikely to provide a true independent assessment. Instead, stability was interpreted for the NPSD study to mean the level to which independent observers agreed with how well the coding rules were applied over a series of assessments.

A workshop was conducted with 10 OR practitioners as participants, none of whom were associated with the NPSD study, to determine the level of peer agreement with a set of existing analyst modelling decisions. Participants were presented with the process and codes (codebook) used, provided with training and examples and asked to assess their level agreement with 22 examples of NPSD coding. Agreement with the conduct of the process was assessed on a scale of 5 (complete agreement) to 1 (complete disagreement) using an approach similar to Ceniccola et al. (2014). Employing a relevant success criteria (Ceniccola et al., 2014) a favourable assessment of consistent process application – and hence stability – was deemed to occur if 80% or more of assessments scored 4 or greater for each coding example and for all coding examples combined.

3.3. Assessing the Reproducibility of the Coding Process

Instead of simply measuring the percentage agreement between coders, Stemler (2001) recommends Cohen's Kappa Coefficient (Cohen, 1960) which accounts for chance in calculating agreement (Equation 1). The Kappa coefficient is widely used and accepted in the literature in scientific disciplines as diverse as: medical science (Sim and Wright, 2005; Viera and Garrett, 2005) and Ecology (Monserud and Leemans, 1992)[3].

In Equation 1, K is the Kappa coefficient, Pa is the relative observed agreement between coders and Pc is the probability of a chance agreement. The value of K will vary from 1 (perfect agreement) down to 0 which indicates that any agreement is likely due only to chance. K can produce negative values which would indicate agreement less than that expected through chance and "this might indicate some systematic disagreement between observers" (Viera et al., 2005, p. 361). The more coding variables in the assessment the more likely that a lower value of K will result (Sim et al., 2005). Weighted forms of the Kappa coefficient have also been proposed for situations where coders are selecting between ordinal values.

Calculating the probability of chance agreement (Pc) in this situation is complicated by a coding rule requiring that a cause variable must be selected from the same or lower category as a given effect variable. As a result there is a changing number of possible variables that the cause could be selected from for a given effect. For example, in the NPSD study, for an effect belonging to the highest level category, there are 36 possible variables from which the cause could be selected (i.e. all variables are possible). While, for an effect variable belonging to the lowest category the cause variable could only be selected from one of 13 (variables from the lowest category). This total is further reduced by 1 as a variable cannot be a cause for itself.

The reproducibility of the NPSD coding was assessed as part of the workshop discussed in Section 3.2. Participants were provided with textual fragments, previously identified in the NPSD study, and asked to apply the coding process to map them to causal concepts from the code book. This was essentially equivalent to steps 3 and 4 in Figure 3, with the exception that the effect phrase had been pre-coded. The coding task was limited in this way so as to align with standard measures of repeatability. Participants' coding results were then compared with equivalent NPSD coding using Table 1.

4. RESULTS

The stability and reproducibility results, less one non-compliant participant's scores, are outlined below.

Stability. The distribution of Likert scores from each participant for each of the 22 code examples assessed is displayed in Figure 4. A total of 88% of the assessments had Likert scores of 4 or 5 and, based on the success criteria identified in Section 3.2, the coding process is accepted as being stable. While all of the participants were trained as OR analysts, only two analysts had previous experience with content analysis.

Reproducibility. The distribution of Kappa scores for each participant performing 22 coding examples is displayed in Figure 5. Despite the limited training provided to the participants, the second activity resulted in 1 slight, 3 fair, 3 moderate and 2 substantial inter-rater agreements using Landis and Koch's (1977)

[3] There are criticisms of Cohen's Kappa and other standard reliability coefficients in the literature – for example (Krippendorff, 2004) – however Kappa remains a widely used measure for content analysis.

commonly accepted criteria (Table 1). Two participants with content analysis experience both assessed achieved a substantial agreement, suggesting that pre-existing skills may be an important factor.

Having established individual measures of reproducibility, the remaining question is how these scores should be combined to establish an overall measure of reliability. A common approach, assuming that all coders coded the same data set, is to average the reliability scores (Kassarjian, 1977). In this case the average score of 0.44 equates to a moderate agreement. The median and mode of the scores provide arguably a more general measure of the distribution and in this case both were calculated as 0.37, equating to a fair agreement.

Table 1. Kappa Interpretation Scale as proposed by Landis and Koch (1977)

Kappa Value	Strength of Agreement
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost Perfect

$$K = \frac{P_a - P_c}{1 - P_c}$$

Equation 1 Cohen's Kappa Coefficient

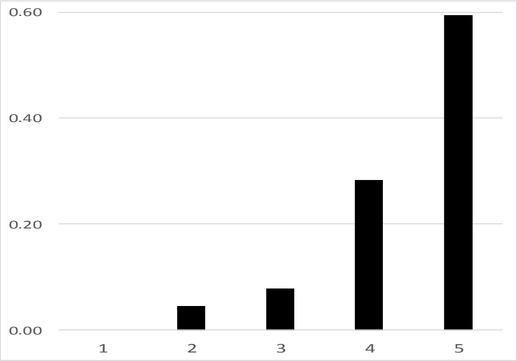

Figure 3. Percentage of assessments per Likert Score

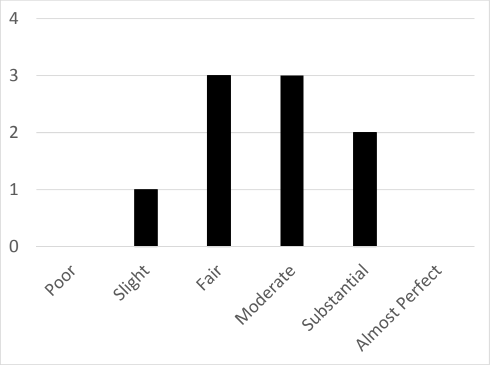

Figure 4. Distribution of Kappa Reproducibility Scores for Workshop Participants

5. DISCUSSION AND CONCLUSION

This paper set out to propose an adequate and feasible approach to validate an IRM, within the wider context of formal model validation, and demonstrate that approach through the NPSD case study, the central question being "what validation measures are necessary to mitigate the risks of using IRMs while realising their benefits in decision-aiding model interviews?". While other forms of IRM are possible, this paper focused on an IRM constructed via formal coding of textual sources using methods such as content or thematic analysis. As with any validation process, the aim is to ensure the resulting model is fit-for-purpose. Hence the aim of IRM validation is to ensure that they are sufficiently reliable to aid, rather than undermine, the model building interviews. Within the broader framework of formal model validation, IRM validation includes conceptual and logical validation, but not data, experimental and operational validation (Landry et al., 1983). It is part of the model building process that leads to a validated formal model. The intent is not to discard the IRM after the model building interviews, but instead to evolve it into the formal model.

Conceptual validation requirements for an IRM formed by textual coding are essentially the same for 'white-box' models in general (Barlas, 1996) and are discussed separately (Coutts, 2013). Logical validation requirements are more situation dependent and aligned to the method used to construct the IRM. In the NPSD case the requirements of logical validation were equated with content analysis coding reliability, which in turn is determined by assessing coding stability and reproducibility. The question is whether the proposed reliability assessments are both feasible and adequate for the purpose of IRM validation. With regards to adequacy, it is useful to consider what we might say of an IRM that emerges from these validation stages. Specifically we could say that we have a level of confidence that the model effectively represents the decision problem, includes key variables of interest and reliably incorporates textual evidence on causal relationships. Arguably this is sufficient for the purpose of providing a structured and common start point for model building interviews that enables stakeholder SME to rapidly understand the problem and quickly engage in useful model building activity without overly constraining or distracting them. Consequently, the NPSD case-study can claim to have achieved an adequate level of validation for its purpose.

Feasibility here is more concerned with access to the necessary resources and ease with which model builders can undertake the required validation activities. For reliability testing within a team of model builders, data for reproducibility testing can be collected as part of the process, providing that different analysts code some of the same texts. Stability testing could be conducted within the group by conducting controlled tests requiring individual analysts to re-code previously coded texts. It is more challenging to conduct controlled and independent testing in situations where there is a single coder, such as the NPSD study, hence the method proposed in the case study, qualitative group assessment. While it can be challenging to form a group of independent peers for a reliability workshop, it is less onerous than requiring that *all* coding in a study only be conducted by a group. Overall the approach seems adequate and feasible for the intended role as shown, however further work is required to confirm the utility of this approach for other modelling situations.

REFERENCES

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System dynamics review, 12*(3), 183-210.

Ceniccola, G. D., Araújo, W. M. C., and Akutsu, R. (2014). Development of a tool for quality control audits in hospital enteral nutrition. *Nutr Hosp, 29*(1), 102-120.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37-46.

Coutts, A. (2013). *Balancing the Validity and Viability of Bayesian Belief Networks for the Study of National Strategic Decisions*. Paper presented at the 22nd National Conference of the Australian Operations Research Society (ASOR 2013), Adelaide, South Australia, Australia.

Ford, D. N., and Sterman, J. D. (1998). Expert knowledge elicitation to improve formal and mental models. *System Dynamics Review, 14*(4), 309-340.

Gass, S. I. (1983). Decision-aiding models: validation, assessment, and related issues for policy analysis. *Operations Research, 31*(4), 603-631.

Hartley, D. L. (1969). Perceived counselor credibility as a function of the effects of counseling interaction. *Journal of Counseling Psychology, 16*(1), 63.

Hossain, A., Moon, T., and Curtis, N. J. (2013). An assurance process for the exchange of operations research software models used for military simulation. *Journal of Simulation, 7*(1), 38-49.

Hsieh, H. F., and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative health research, 15*(9), 1277-1288.

Joyce, C. K. (2009). The blank page: effects of constraint on creativity. *Available at SSRN 1552835*.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of consumer research*, 8-18.

Kjaerulff, U. B., and Madsen, A. L. (2008). *Bayesian Networks and Influence Diagrams: A Guide to Construction and Analysis*: Springer.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research, 30*(3), 411-433.

Landry, M., Malouin, J.-L., and Oral, M. (1983). Model validation in operations research. *European Journal of Operational Research, 14*(3), 207-220.

Monserud, R. A., and Leemans, R. (1992). Comparing global vegetation maps with the Kappa statistic. *Ecological modelling, 62*(4), 275-293.

Nadkarni, S., and Shenoy, P. P. (2004). A causal mapping approach to constructing Bayesian networks. *Decision Support Systems, 38*(2), 259-281.

Pace, D., and Sheehan, J. (2002). Subject matter expert (SME)/peer use in M&S V&V. *Proc. of the Foundations*.

Sim, J., and Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy, 85*(3), 257-268.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation, 7*(17). Retrieved from http://PAREonline.net/getvn.asp?v=7&n=17 website:

Vennix, J. (1999). Group model-building: tackling messy problems. *System Dynamics Review, 15*(4), 379-401.

Vennix, J. A., Gubbels, J., Post, D., and Poppen, H. (1988). *A structured approach to knowledge acquisition in model development.* Paper presented at the International Conference of the Systems Dynamics Society, La Jolla, CA.

Viera, A., and Garrett, J. (2005). Understanding interobserver agreement: the kappa statistic. *Fam Med, 37*(5), 360-363.