JOURNAL OF APPLIED MEASUREMENT, *16*(3),

Measuring Psychosocial Impact of CBRN Incidents by the Rasch Model

Stef van Buuren
Netherlands Organization for Applied Scientific Research TNO,
Delft, The Netherlands
University of Utrecht, The Netherlands

Diederik J. D. Wijnmalen
Netherlands Organization for Applied Scientific Research TNO,
Delft, The Netherlands

An effective response to chemical, biological, radiological and nuclear (CBRN) incidents requires capability planning based upon an assessment of risks in which all types of possible consequences of such incidents have been taken into account. CBRN incidents can have a wide range of consequences of which psychological and social effects (possibly leading to societal unrest) are often pointed out as very likely to occur. The goal of our research was to establish an objective measurement of psychosocial impact of CBRN incidents with the use of the Rasch model. We created a list of eleven items, each of which tapped into an aspect of psychosocial impact of incidents. Eleven judges scored ten CBRN scenarios on this list of items. Two items needed to be removed due to misfit. The resulting nine-items test fitted the Rasch model well. Three items showed mild forms of differential item functioning, but were retained in the test. The reliability of the instrument was 0.83. The scale can be used to quantitatively measure the inherently qualitative nature of psychosocial impact of CBRN incident scenarios in order to better compare this type of impact with quantitative impact types such as number of casualties, costs, etc. Administration of the scale is simple and takes about one minute per scenario. We recommend wider use of the Rasch model for improving the quality of total impact measurement in case of being faced with both qualitative and quantitative types of impact.

Requests for reprints should be sent to Stef van Buuren, Netherlands Organization for Applied Scientific Research TNO, Postbus 3005, 2301 DA Leiden, The Netherlands, e-mail: stefvanbuuren@tno.nl.

Introduction

An effective response to chemical, biological, radiological and nuclear (CBRN) incidents requires capability planning based upon an assessment of risks in which all types of possible consequences of such incidents have been taken into account. CBRN incidents (man-made or accidental) can have a wide range of consequences: inaccessibility to parts of the territory, casualties, financial consequences, environmental damage, disruption to critical infrastructures or institutions, societal fear and unrest. The reduction of adverse consequences of CBRN incidents requires careful capability planning in terms of prevention, preparation, response, aftercare, etc. An adequate assessment of CBRN incident risks can provide a basis for prioritisation of capability enhancement. One common approach in national risk assessments is to compare different types of CBRN incidents by plotting the impact against its likelihood of occurring. Such a risk chart helps to bring focus to those scenarios that require risk mitigation capabilities and enhance an effective use of resources (Bergmans, van der Horst, Janssen, Pruyt, Veldheer, Wijnmalen et al., 2009; Pruyt and Wijnmalen, 2010).

Some types of consequences of CBRN incidents are measured on quantitative scales, like the amount of area flooded, number of casualties, monetary loss, and so on. CBRN incidents can also have huge impacts of psychological and social nature (Hall, Norwood, Ursano and Fullerton, 2003; Lemyre, Clement, Corneil, Craig, Boutette, Tyshenko et al., 2005), but it is less obvious how such impacts can be quantified. The root causes of psychosocial effects are inherently qualitative in nature. In this paper we investigate how to establish an objective, quantitative measurement scale for psychosocial impact resulting from CBRN incidents.

Current practice in risk assessment is to design a system of rules that lead to intensities on some ordinal scale, to treat them separately from quantitative impact, or even to ignore such types of impact entirely and focus on casualties and costs. The hope is that measurements of psychosocial impact made on a quantitative scale will aid planners in balancing the different type of impacts, including consequences of psychological and social nature.

This paper proposes a simple scale to measure psychosocial impact of incidents, including man-made incidents resulting from attacks with CBRN agents. The article will address only the quantification of qualitative impact, and not the way in which qualitative and quantitative impact assessments are combined in an overall risk assessment methodology.

Method

Theoretical concepts

Psychosocial impact of an incident refers to the reaction of the population being characterized by feelings and expressions of negative emotions, such as fear, anger, dissatisfaction, sorrow, disappointment or aversion. The population consists not only of victims, but also includes those not directly or immediately affected. The definition of psychosocial impact covers emotions, whether they are explicitly expressed or not.

Expressions of fear could include flight and avoidance behaviour, deviations from the usual patterns of daily life, and taking apparently irrational decisions. Expressions of anger could include, for example, protests, demonstrations, disturbances in public life, vandalism, and calls for protest fuelled by feelings of dissatisfaction. Fear and anger that is mixed with grief and horror might lead to chaos and mass hysteria.

The goal of our research was to establish an objective measurement of psychosocial impact of CBRN incidents. The primary scoring mechanism is based on various drivers of negative emotions (used as ‘indicators’), as well as the intensity with which these apply. In addition, we use the extent of the observable manifestations as a possibly reinforcing or mitigating factor when establishing a score.

The indicators of psychosocial impact mentioned above fall into three groups, each of which is considered equally important (Bergmans et al., 2009; Pruyt et al., 2010):

a. Perception of the incident to the victims or the rest of the population:
   - Unfamiliarity with the nature or source of the risk;
   - Uncertainty about the degree of threat or danger and the possibility that one may be personally affected by it;
   - Degree of unnaturalness of (the cause of) the incident;
   - Extent to which vulnerable groups - such as children, elderly, sick, needy – are or will be disproportionately affected.
b. Trust in the act of (government) agencies and/or companies to the victims and the rest of the population:
   - Degree of perceived culpability (failure) of relevant businesses and (government) agencies in the occurrence of the incident or the occurrence of undesirable effects (related to prevention);
   - Degree of loss of confidence in the public information actions of the government authorities and the companies involved and other bodies (Note: not emergency) on the one hand, and the management of the incident and other disclosures about the incident and its causes (related to preparation and initial response) on the other;
   - Degree of loss of confidence in the actions of the emergency services in the management of the incident, for example in case of exceeding emergency times, lack of capacity, inadequate / incorrect operations etc. (related to preparation and initial response).
c. Operational perspective for those affected by the incident:
   - Degree of ignorance and / or inexperience with possible forms of self-reliance in the specific situation;
   - Degree of personal inability to manage their own situation.

Table 1 lists a total of eleven items (questions) designed to cover the above-mentioned

Table 1

Candidate items of the psychosocial impact scale for CBRN incidents. All items have three response categories: 0 = no, 1 = to some extent, 2 = to a large extent.

Item	Question
a1	Are people (whether or not victims) unfamiliar with the nature or cause of the incident?
a2	Are people (whether or not victims) uncertain about the extent of the threat or danger?
a3	Are people uncertain as to whether they will be personally affected?
a4	Do people (whether or not victims) think that the incident's cause is unnatural?
a5	Do people (whether or not victims) think that specific social groups (as to age, social position, cultural background, disabled, etc.) are or will be disproportionately affected?
b1	Do people (whether or not victims) think that companies or (governmental) institutions can be faulted for causing the incident and its consequences?
b2	Have people (whether or not victims) lost their trust in the risk mitigation management capabilities of relevant organisations or (governmental) institutions with respect to the incident and its consequences?
b3	Have people (whether or not victims) lost their trust in the information capabilities of relevant organisations or (governmental) institutions regarding the incident and its consequences?
b4	Have people (whether or not victims) lost their trust in the emergency response capabilities of relevant rescue organisations?
c1	Are people who are affected unfamiliar or inexperienced with ways of self-reliance in the specific situation?
c2	Are people who are affected (and know what to do) unable to manage their individual situation?

indicators. The response categories were: *no* (no logical relationship with the incident and its causes), *to some extent* (the indicator is present to a small degree), and *to a large extent* (the indicator is clearly present).

Data collection

The items in Table 1 were used to create a score sheet. We set up a small experiment where eleven different judges scored the impact of ten different CBRN scenarios. The scenarios described a deliberate disruption of daily life situations in various intensities and under various circumstances using various CBRN agents. The judges were recruited among researchers working at TNO, colleagues of the authors.

Two incomplete questionnaire forms were removed prior to analysis. The resulting dataset consisted of a matrix of 108 rows (= 11 judges * 10 incidents - 2) and 11 columns (number of items). The unit of analysis is the combination of judge and scenario.

Measurement model

The data were analysed with the Rasch model (Rasch, 1960). The model relates the test data to a latent construct β, here the degree of psychosocial impact. The Rasch model specifies the probability of passing a test at a given β by a series of logistic curves, one curve for each category transition. The logistic curves can only differ in their location, not in their slope. Suppose that the data are coded as $X_{ni} \in \{0, 1, 2\}$ where $n = 1,\ldots, N$ indexes the N rows in the data, and where $i = 1, \ldots, P$ indexes the P indicators. The degree of psychosocial impact for the n'th unit is written as β_n, and the number of categories of indicator i is denoted by $m_i + 1$. For an indicator with $m_i + 1$ categories, we define m_i initially unknown threshold parameters δ_{ij} ($j = 1,\ldots, m_i$). The thresholds are points on the latent scale at which the probabilities of responses in one of the two adjacent categories are equal. According to the polytomous Rasch model, the probability of observing response x for unit n on item i can be written as (Wright and Masters, 1982)

$$\Pi_{nix} = \frac{\exp \sum_{j=0}^{x} (\beta_n - \delta_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^{k} (\beta_n - \delta_{ij})}, \quad (1)$$

where $\delta_{i0} \equiv 0$ for convenience. Equation (1) specifies the probability for item i. The model assumes that the scoring probabilities of another item $i' \neq i$ are conditionally independent given β. A special property of the Rasch model is that, apart from arbitrary rescaling, the threshold parameters do not depend on the distribution of β in the sample. Hence, any sample of units can be used for threshold estimation, though some will be more efficient than others. The separation index expresses the portion of the total variation that is attributable to differences in ability, and hence is a measure of reliability, akin to Cronbach's alpha.

Rasch originally developed the model to measure the reading ability of pupils in schools. The typical test (exams) in such settings consists of a series of questions (items) whose answers can be either right or wrong. Andrich (1978) and Wright and Masters (1982) developed extensions to the polytomous case. The Rasch model had a great influence on measurement in the social sciences, and is now transforming measurement in the medical field. Apart from a few studies (Fischer, Frewer and Nauta, 2006; Li, Liu, Liu, Feng and Cai, 2011; Weller, Dieckmann, Tusler, Mertz, Burns and Peters, 2013), the model is rarely applied in the context of risk assessment.

The Rasch model essentially looks for consistency among the item responses. For example, if an incident has a high score on one item, it is also expected to score higher on the other items. An important assumption is that all items measure the same construct, here 'psychosocial impact.' When the Rasch model holds, it embodies many desirable properties of a measurement scale. For example, the simple sum score contains all relevant information about the qualitative construct 'psychosocial impact.' The psychosocial impact of each incident scenario and the item difficulties are additive measures on the same latent variable, so measurements are made on an interval scale. The precision of each scenario's psychosocial impact and item difficulty is known, so we can determine in advance how well a test classifies a scenario.

Statistical analysis

We used the RUMM 2020 software to fit and evaluate the model. (RUMM Laboratories, 2003) We applied the usual item selection strategy. Items were removed from the analysis until the remaining subset of items fitted the Rasch model, as evidenced by a non-significant item-trait interaction χ^2-test at 0.05 level. We also investigated whether different judges used the items in different ways, a phenomenon known as differential item functioning (DIF). (Holland and Wainer, 1993). We split the data into eleven subsets (one for each judge), re-estimated the item parameters per judge, examined the differences between sets graphically, and tested for uniform and interaction DIF by ANOVA.

Results

Table 2 contains the frequency distribution per item. Note that all categories are reasonably filled, except for category 2 in items *a5* and *b4*.

The eleven-items scale does not fit the Rasch model ($\chi^2 = 48.4$, $DF = 22$, $P < 0.001$). Inspection of the item fit residuals revealed two items with large residuals (*a5* and *c1*) that misfit the Rasch model. Item *a5* ("specific groups affected") failed to discriminate different levels of psychosocial impact, whereas item *c1* ("victims are helpless") was sensitive only on a specific part of the scale. Removal of both items considerably improved the solution, and in fact resulted in a fitting Rasch model on the remaining nine items ($\chi^2 = 14.6$, $DF = 14$, $P = 0.69$). The reliability of the nine-items scale as measured by the proportion of true score variance accounted for by the sum score is equal to 0.83.

Figure 1 is the person-item threshold distribution of the nine-items scale. The histogram in the upper part of the display is the distribution of the estimated psychosocial impact for ten CBRN incidents using the sum score on nine items. The smoothed line ('test information') indicates the locations on the scale where the test (i.e., the sum score over the nine items) is sensitive. In this case, the match is good, so the nine-items test is able to discriminate between the ten CNRN scenarios.

We tested for uniform and interaction DIF. We found two items for which some of the judges scored uniformly higher than other judges: b1 ("failed rescue") and c2 ("victims are unable"). Such items may potentially be problematic if different judges score different scenarios (which was not the case in this experiment). Item b4 ("lost confidence in help") ($F = 1.89$, $p = 0.03$) showed mild signs of undesirable interaction DIF, but as the evidence was not strong and as the item fitted the model, we decided to keep it in the test.

Table 2

Frequencies per category of the psychosocial impact items.

Item	Description	No	Some extent	Large extent
a1	Cause unknown	22	49	37
a2	Uncertain about threat	15	55	38
a3	Uncertain whether affected	17	63	27
a4	Unnatural cause	30	31	47
a5	Specific groups affected	77	26	4
b1	Failed rescue	32	52	23
b2	Lost confidence in management	57	39	10
b3	Lost confidence in information	56	40	10
b4	Lost confidence in help	75	28	3
c1	Victims are helpless	35	46	25
c2	Victims are unable	50	36	20

Figure 2 plots the sum scores over nine items assigned to each of the ten CBRN scenarios by the eleven judges. The test created from nine items provides a scale with scores in the range 0 to 18. An interval-scaled measure can be derived from this raw score by a unique non-linear, S-shaped transformation. Typically, the raw score itself is taken for further calculation, as the practical differences between the two are minor, especially in the middle part of the scale.

Figure 3 shows the item map of the nine items ordered in difficulty. The colours delineate the different areas where each category has the largest probability of being observed. The spread of the categories is balanced over the latent variable. No reversed thresholds were found, which

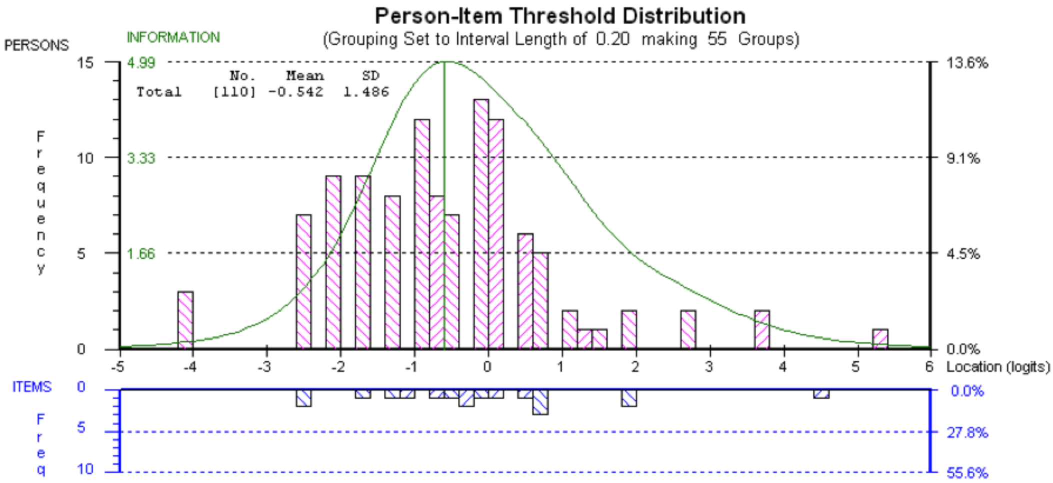

Figure 1. Person-item threshold distribution. The histogram in the upper part is the distribution of the estimated psychosocial impact for 10 CBRN incidents using the sum score on nine items. The bottom part visualizes the locations of the estimated category threshold parameters. The smoothed line is the test information.

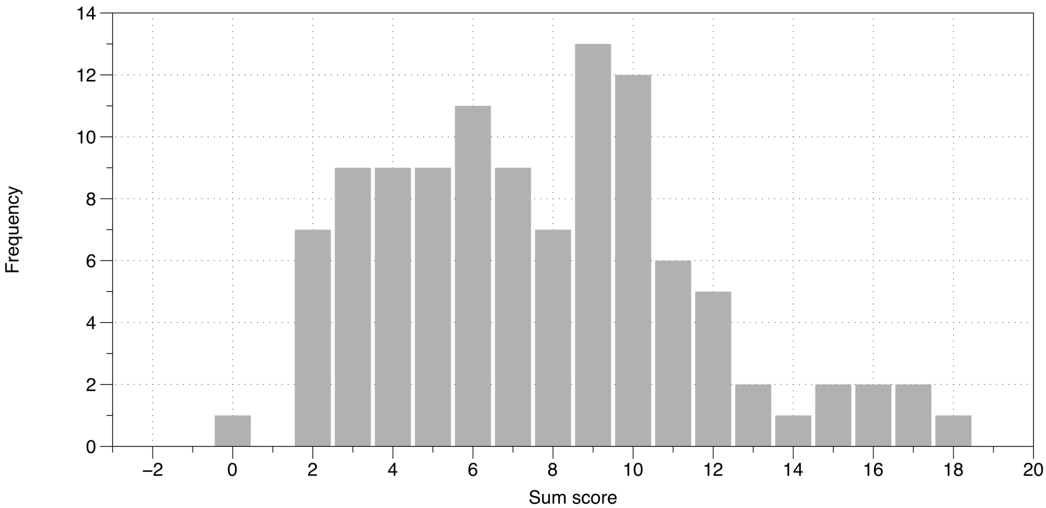

Figure 2. Frequency distribution of the raw sum score of the nine items test for ten different CBRN incidents as evaluated by eleven judges.

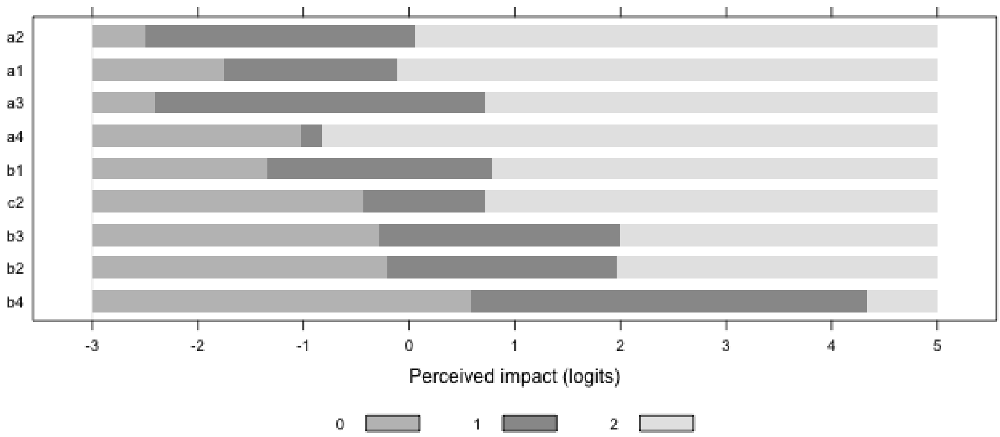

Figure 3. Item map displaying the threshold estimates of the 9-item scale.

is an indication that the implied category ordering works as intended.

In practice, one may calculate an average raw score over the eleven judges to obtain one summary score of psychosocial impact for the incident. In general, if multiple judges score the same scenarios, it is possible to calculate a single-number summary for a scenario as the average of the individual impact scores. Note that this will increase the reliability of the measurement beyond 0.83. According to the Spearman-Brown prophecy formula, two independent judges will produce a reliability of 0.91, with three judges 0.94, and with ten judges 0.98, which is almost perfect. Thus, in practice, averaging the scores over three to five judges provides a quantitative score of psychosocial impact that has excellent reliability.

As this score measures the amount of psychosocial impact of an incident, it does not tell us how it should be combined with other impact measures, like territory lost and financial loss. It does however give a reliable quantification of psychosocial impact on a well-defined scale that can now be balanced against other quantitative measures. The combination of this measure with all other impact measures has been done in the actual case study (using multi-criteria analysis) but is beyond the scope of this paper.

Conclusion

Some incidents, especially CBRN incidents and among these particularly malicious CBRN incidents (misuse, attacks) can have considerable psychosocial impact on the population. Until now, however, it was difficult to quantify that type of impact. We have presented a list of nine items (in the form of 'questions'), based on the Rasch model, whose sum score can be regarded as a quantitative impact measure. The reliability of the nine-items test was good (0.83). This measure can be used as a quantitative assessment of psychosocial impact when evaluating the consequences and risks of incidents on the population. For example, the impact score can be combined with measures of other, inherently quantitative, impact types to an overall impact value and thus be used in a risk chart of incident scenarios. It hence contributes to cost-effective resource allocation.

The time to fill out the nine items is short, typically less than one minute per scenario once the judge has become familiar with the scenario. If multiple judges score the same scenarios, it is possibly to calculate a single-number summary for a scenario as the average of the individual impact scores.

This paper used the classic Rasch model, which represents the data by two sets of parameters, one for the columns (items) and one for the row units (judge-scenario combinations). The editor pointed out that our data would also enable a facet approach. The many-facet Rasch model splits the parameters of the judge-scenario combinations into two sets, one for the judges and one for the scenarios (Linacre, 1994). The facets model assumes that the effects of judges and the scenarios are additive, and hence, when it fits, provides a more parsimonious representation of the data. We did not use the facet approach here because our primary interest was to investigate whether the items formed a scale. The facet model is more restrictive to the Rasch model, so it might fail to detect the case where items are scalable according to the Rasch model (our goal), but where judges and scenarios are not additive. Second, in order to preserve confidentiality of the scenarios, future data collection schemes may not administer multiple scenarios to the same judge. Notwithstanding these practical conditions, the facet approach provides an elegant approach to the problem of integrating judgements across judges. Our work should therefore be seen as only a first step to solving the issues surrounding the objective measurement of psychosocial impact of CBRN incidents.

In the experiment, we not only studied psychosocial impact, but also rated three other types of impact of incidents: infringement of the international position/prestige, violation of the democratic system, and loss of cultural heritage. We attempted to create measurement scales for these outcomes as well using in a similar methodology. However, these attempts were not successful. The reason was the fact that the developed scenarios were not described detailed enough with respect to these types of impact. This resulted in the score of zero almost everywhere, and hence these tests were 'too difficult' to see any differences between the incidents. Of course, this does not preclude the existence of such scales. Future work may follow two complementary strategies to create Rasch scales for the other outcomes. The first is to set up an experiment with new scenarios that will have a higher impact on international position, violation of the democratic system, and loss of cultural heritage. The second is to adapt and reformulate the current items so that they will become easier and thus more informative. A combination of both is likely to be needed to achieve the best results.

Our proposed psychosocial impact scale represents an improvement over current practice, but it is not a definitive answer. It would be useful to validate the results to other measures of psychosocial impact. Further refinement is likely to be possible using other scenarios and other judges. However, we believe that the methodology as used here is on target. It aids in carving out fair and scientifically valid impact scales for including inherently qualitative consequences of incidents in a quantitative risk analysis. This enables a more thorough comparison of different types of consequences across incident scenarios.

References

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561-573.

Andrich, D., Sheridan, B. S., and Luo, G. RUMM 2020. Rasch unidimensional measurement models [Computer software]. Perth, WA, Australia: www.rummlab.com.au. 2005.

Bergmans, H., van der Horst, J., Janssen, L., Pruyt, E., Veldheer, V., Wijnmalen, D., et al. (2009). *Working with scenarios, risk assessment and capabilities in the national safety and security strategy of the Netherlands*. The Hague, The Netherlands: Ministry of the Interior and Kingdom Relations.

Fischer, A. R. H., Frewer, L. J., and Nauta, M. J. (2006). Toward improving food safety in the domestic environment: A multi-item Rasch scale for the measurement of the safety efficacy of domestic food-handling practices. *Risk Analysis*, *26*, 1323-1338.

Hall, M. J., Norwood, A. E., Ursano, R. J., and Fullerton, C. S. (2003). The psychological impacts of bioterrorism. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *1*, 139-144.

Holland, P. W., and Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Lemyre, L., Clement, M., Corneil, W., Craig, L., Boutette, P., Tyshenko, M., et al. (2005). A psychosocial risk assessment and management framework to enhance response to CBRN terrorism threats and attacks. *Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science*, *3*, 316-330.

Li, J., Liu, H.J., Liu, H., Feng, T., and Cai, Y. (2011). Psychometric assessment of HIV/STI sexual risk scale among MSM: A Rasch model approach. *BMC Public Health*, *11*, 763.

Linacre, J. M. (1994). *Many-facet Rasch analysis*. Chicago, IL: MESA Press.

Pruyt, E., and Wijnmalen, D. J. D. (2010). National risk assessment in the Netherlands – A multi-criteria decision analysis approach. In M. Ehrgott, B. Naujoks, T. Stewart, and J. Wallenius (Eds.), *Multiple criteria decision making for sustainable energy and transportation systems* (pp. 133-143). New York, NY: Springer Physica Verlag.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J. and Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*, 198-212.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.