COGNITIVE RADIO TESTING USING PSYCHOMETRIC APPROACHES

Carl B. Dietrich (Bradley Department of Electrical Engineering, Virginia Tech, Blacksburg, VA, USA, cdietric@vt.edu), Edward W. Wolfe (Pearson, Iowa City, IA, USA, ed.wolfe@pearson.com), Garrett Vanhoy (University of Arizona, Tucson, AZ, USA, gvanhoy@email.arizona.edu)

ABSTRACT

Cognitive radios promise efficient spectrum use and other performance improvements through use of machine learning to adapt the radios' operational parameters to optimize performance; however, their flexibility complicates evaluation of cognitve radios' performance. We propose to improve cognitive radio development and evaluation using approaches developed for efficiently measuring and testing human cognitive characteristics. Cognitive radio performance evaluation requirements and applicable psychometric approaches are described. Finally, a proof-of-concept application of a psychometric measurement technique to evaluate cognitive engine performance is presented for simulated channel conditions for multiple prioritizations of optimization goals.

1. INTRODUCTION

Cognitive radios (CRs) can improve communications through learning and adaptation in applications that include dynamic spectrum access (DSA) and power management. A CR is controlled by a cognitive engine (CE) implemented in software, that controls frequency, power, modulation, and coding to meet designated goals. CRs employ observation, orientation, memory, learning, planning, decision making, and action [1] to achieve goals such as maximizing data rate while minimizing interference to other spectrum users. CR capabilities, while limited, roughly parallel those of human cognition.

Effective methods for evaluating CR performance are needed to refine CR capabilities in the research and development stages, verify CR compliance with regulations, and compare CR products. However, measuring performance of CRs is difficult; their flexibility allows a variety of behaviors. Psychological measurement models applied to humans may improve efficiency and effectiveness of CR testing. These models depict interaction of a person with a testing context mathematically via an equation—the measurement model—and a set of associated assumptions used to derive useful parameter estimates and diagnostic indices. In a typical psychological testing situation, a person responds to a series of test items designed to be indicators of an underlying latent trait, referred to as a construct. Those responses are coded into nominal (categorical) or ordinal values through a scoring process. The codes serve as data upon which the measurement model's parameters are estimated. Applications of these models can be extended to estimate parameters that depict the quality of the decision making model implemented by a CE.

Measurement models employed in education and psychology, referred to as item response models (IRMs), offer considerable flexibility in how one scales a variety of types of test data, compensate for difficulties that often arise in testing situations (e.g., missing data, multiple forms of a test), and allow for efficient delivery of tests (e.g., adaptive tests that are tailored to provide the most relevant and precise information about each individual). Potential benefits of IRMs for CR performance evaluation include: efficient, adaptive testing; conjoint measurement of CE quality and test case difficulty; flexibility in modeling different types of events; and generation of diagnostic statistics to identify testing anomalies. CR performance evaluation requirements, relevant psychometric approaches, and a plan for applying these approaches are presented. CR and CE performance evaluation is referred to as testing in this paper although it can include system-level evaluation that does not involve direct testing of software or hardware components.

2. COGNITIVE RADIO TESTING REQUIREMENTS AND APPROACHES

This section describes required capabilities, and distinct characteristics and requirements of each type of testing

2.1 Testing Requirements

Requirements include: (a) development and selection of worst case and typical scenarios; (b) development and measurement of CR performance figures of merit; (c) ability to relate variables under control of designers to figures of merit; and (d) efficient, possibly dynamic, selection of relevant, representative test scenarios from among myriad alternatives.

Research and development engineers must evaluate CR designs and relate resulting figures of merit to factors under their control. This allows optimization of CR architectures

and CE structures, algorithms, and algorithm parameters. The internal state of the CE is available, allowing investigation of CE parameter value effects.

Regulatory compliance tests enable type acceptance and product development; however, it is difficult to verify that a CR will not violate regulations in operation. Regulations can specify performance, e.g., interference avoidance capabilities, as well as frequency and power limitations. Regulators must compare each radio's performance to an objective standard, and IRMs can be applied to establish worst-case tests.

Users must compare performance of CR products to make informed purchasing decisions, and CR producers must compare their products to those of competitors. Internal CR or CE states may not be accessible; performance is observed by external monitoring. Rapid, efficient testing is desired. Both a figure or figures of merit and knowledge of worst-case performance are of interest. Standard and adaptively administered test cases are desirable and IRMs can determine these cases efficiently.

2.2 Proposed CR Performance Metrics

Haykin [2] identifies quality of service and spectrum efficiency as major goals of CRs. Power consumption is also important. Performance indicators that express achievement of these goals include data throughput, latency, and several others [3]. Zhao, et al [4] provide a comprehensive review and examples of goals, metrics, and utility functions at node, network, and application levels. Some metrics are inherently numerical (e.g., dynamic range and SINR) and others can be treated as categorical (e.g., Mobility and trajectory awareness, Distributed or centralized decision making). Metrics are grouped into four categories: cognitive functions, overall performance, complexity, and technical maturity. Measurement of some indicators may require a proprietary interface.

2.3 Cognitive Radio Testing Approaches

CR behaviors may be impossible to fully anticipate and characterize. Current approaches to radio testing include manual or automated testing using multiple test instruments [5] or multifunction test instruments [6]. In addition, several cognitive radio testbeds as well as measurement campaigns and techniques are reviewed in [7].

Alternative approaches to CR testing lie on a continuum from a fixed battery of tests through automated, adaptive testing using a software controlled testbed, to "cognitive testing" that mirrors the CR's own capabilities.

Adaptive testing provides needed flexibility, using test cases chosen and administered dynamically based on established theory. Test cases are optimized to learn most about each radio quickly, and adaptation can follow a standard algorithm. Thus, adaptive testing is more efficient and flexible than fixed testing but is less complex and more theoretically tractable than cognitive testing.

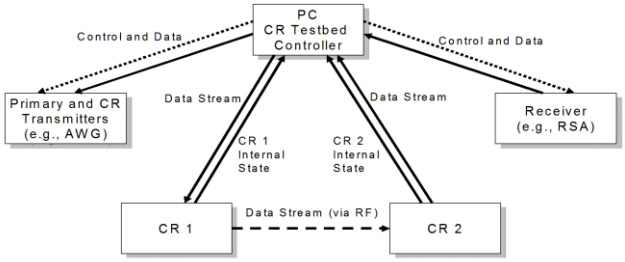

Figure 1. Cognitive radio test configuration for research and development.

2.4 Proposed Cognitive Radio Test Configuration

Figure 1 shows a test configuration for CR development testing that could be achieved in a controlled laboratory environment or in a testbed such as described in [8]. For clarity, a simplex or unidirectional link between two CRs is shown, although duplex communications and multiple CRs could be used. A computer controls an arbitrary waveform generator (AWG) and real time spectrum analyzer (RSA) to generate RF scenarios and measure the CR's responses, respectively. This extends the work of [5]. Transmitted and received data streams are compared to measure error rate and latency. Internal states of the CR, if available, are monitored. Similar configurations enable regulatory and product comparison testing, except internal CR states may be unavailable. The following sections describe relevant psychometric models and CR testing applications.

3. PSYCHOLOGICAL MEASUREMENT AND ITEM RESPONSE MODELS

3.1 Benefits of IRMs

IRMs have been applied to measurement of numerous aspects of human performance. IRMs depict how latent traits govern behaviors so that observed behaviors can be used to estimate levels of those traits. To elicit observable behaviors, test items that allow for the expression of behaviors indicative of different levels of the latent trait are administered in a standardized manner. Responses are recorded and assigned categorical or ordinal codes that depict the level of the latent trait that the observed behaviors imply. These codes serve as data from which IRM parameters are estimated.

Application of IRMs to CR testing could simplify, make more efficient, and introduce flexibility into CR testbed procedures. First, IRMs allow for estimation of CE quality without requiring the same tests to be performed on all CRs. Thus, efficient testbed procedures could be created, allowing for adaptive testing based on the successes and failures of a particular CR on previous tests. As a simple example, a cognitive radio that fails to detect and react to a strong interfering signal may not need to be tested for its ability to react to a weaker interfering signal.

Second, IRMs allow for conjoint measurement of the quality of CEs associated with different CRs and the difficulty associated with individual testing items. Such information can provide insights into areas in which classes of CEs may be deficient. For example, suppose a particular CEs exhibits erratic performance on a class of testing items that manipulate a common parameter of the testing environment. CR developers can use this information to improve the decision making rules employed by that CR.

Third, due to the variety of IRMs available, considerable flexibility would be afforded to those engaged in CR testing. For example, IRMs can model a wide variety of outcomes including dichotomous, polytomous, binomial, and Poisson events—not all test results need to be simplified into pass/fail distinctions [9]. In addition, multilevel IRMs [10] can take into account dependencies that may arise due to the testing process, such as correlated errors due to performing multiple tests on a specific piece of testing equipment or correlated errors across instruments within a particular type of test. Also, Multidimensional IRMs [11] can allow for the measurement of multiple dimensions of a CR's performance by taking into account the different types of tests and exploiting the covariance between responses to those tests in order to create more precise measures of the CR's efficiency.

Fourth, IRMs generate useful diagnostic statistics, which provide information that can be used to identify testing-process anomalies. Data-to-model fit indices are used in educational and psychological testing to identify cheating, inattention, poorly written test items, miskeyed test items, and potentially biased test items. Applied to CR testing, these indices may provide useful information concerning the function of individual CR units (e.g., unit failure or variations in manufacturing quality), differential performance of groups of CRs across testing tasks (e.g., strengths and weaknesses of particular models or software versions), or functioning of test equipment (e.g., malfunctions or calibrations requirements).

3.2 Overview of IRMs

Rasch [12] specified one of the first IRMs, which indicates that the natural log of the odds of observing a particular outcome (coded as $X = 1$, the probability of which is $\pi_{X=1}$) versus the converse of that outcome (coded as $X = 0$, the probability of which is $\pi_{X=0}$) is a linear function of two parameters—one that represents the level of the latent trait being expressed by the person (θ_n) and one that represents the level of the latent trait depicted by the indictor (δ_i), $\ln(\pi_{X=1} / \pi_{X=0}) = \theta_n - \delta_i$. Applied to a CR context, this model would depict the log of the odds of an effective versus an ineffective adaptation by a particular CR ($\pi_{X=1} / \pi_{X=0}$) as a linear function of the adequacy of the decision making model used by that CR's CE (θ_n) and the complexity of the decision making tasks to which the CE's rules have been applied (δ_i).

The CE parameter, θ_n, is typically scaled so that negative values indicate poor performance while positive values indicate superior performance. The task parameter, δ_i, is typically scaled so that negative values indicate less demanding tasks while positive values indicate demanding tasks. Hence, when a task is perfectly matched to a CE's performance (i.e. $\theta_n = \delta_i$ or $\theta_n - \delta_i = 0$), the CE has a .50 probability of making an effective decision in response to the task in question. The values of these parameters can be estimated from the responses of multiple CEs to multiple test tasks using maximum likelihood procedures, and several commercially available computer programs are available to perform these estimation routines.

The precision of the IRM's parameter estimates are dependent on several features of the testing context. Generally, more precise estimates are obtained when a CE is exposed to a large number of tests—typically, a minimum of 20 tests is recommended. More precise estimates are also obtained when test tasks are well matched to the CE, increasing the amount of variability in the vector of scored responses. Generally, smaller standard error values are associated with more precise measures, and these standard errors are also computed routinely by commercially available computer programs.

An important use of the standard errors of the IRM parameter estimates is their application to computing the reliability of the CE measures. Conceptually, reliability refers to the expected correlation between two independent tests of the same CE. In practice, this correlation is estimated as a function of the average standard error of parameter estimates for a set of CEs (SE_θ) and the variance of the estimates themselves. Specifically, reliability = 1 – [Mean(SE_θ) / Variance(θ_n)]. In applications of IRMs to humans, reliabilities greater than .80 are considered to be adequate for most purposes.

Another important feature of IRMs is the fact that they can be used to predict the success or failure of a particular CE on a set of testing tasks. That is, the model expresses an expected outcome for each CE on each task (E_{ni}), given what is known about the CE (i.e. its estimated θ_n) and the task (i.e. its estimated δ_i). These model-based expectations are used in practice to evaluate the model-data fit for a

particular CE or a particular test task. Specifically, a residual can be computed for each observed response as the difference between the observed and expected outcome ($X_{ni} - E_{ni}$). These residuals are typically standardized and squared and are then averaged across CEs for a particular test task to obtain a measure of fit for each task. The residuals can also be averaged across items for a particular CE to obtain a measure of fit for each CE. Fit values near zero indicate a high level of model-data agreement while extreme positive values indicate a high degree of model-data misfit. These fit indices are useful for detecting potentially problematic testing issues such as inappropriate test tasks, data recording errors, erratic CE functioning, or task-specific strengths and weaknesses of a particular CE.

4. PROOF OF CONCEPT METHOD

As a first step toward a proof of concept, Rasch's model for polytomous scores was applied to simulated CE data. This section describes the simulated data generation and Rasch analysis methods.

4.1 CE Data Simulation

CEs use one of two general approaches to adapt a radio's operational parameters to its environment. In the first approach, a CE is given a prototype of its operating environment in the form of an objective function. This function attempts to fully describe relationships between the spectral environment, operational parameters, and link quality. A CE using this form of adaptation first senses its environment, then optimizes the hard-coded function to produce operational parameters. In the second approach, a CE is given no prototype of its environment. To adapt to its environment, it gathers information about the environment, then it optimizes its link parameters by strategically guessing with the goal of selecting successively better operational parameters.

Success of both of these methods depends on the underlying algorithm for adaptation. Theoretical relationships between several transmission parameters and channel properties are well-known. Newman [13] explores how to aggregate these relationships into a single objective function so that a CE can generate a solution given a specific priority.

For the experiments presented in this paper, each CE was given the task of optimizing a trade off between bit-error-rate, power, and good throughput. These trade-offs are expressed through weights (see Table 1). Each CE could change four transmission parameters within a specified upper and lower bound (see Table 2).

Table 1: Objective weights for each task or item

Item Number	BER	Power	Good Throughput
1	1/3	1/3	1/3
2	2/3	1/6	1/6
3	1/6	2/3	1/6
4	1/6	1/6	2/3
5	2/5	2/5	1/5
6	2/5	1/5	2/5
7	1/5	2/5	2/5
8	1/2	1/4	1/4
9	1/4	1/2	1/4
10	1/4	1/4	1/2
11	4/5	1/10	1/10
12	1/10	4/5	1/10
13	1/10	1/10	4/5
14	1/20	9/10	1/20
15	1/20	1/20	9/10
16	4/7	2/7	1/7
17	2/7	4/7	1/7
18	2/7	1/7	4/7
19	1/7	4/7	2/7
20	1/7	2/7	4/7

Table 2: Parameters used in the simulation

Parameter	Minimum	Maximum	Step
Signal-to-Noise Ratio (SNR)*	0.1 (-10 dB)	10 (+10dB)	1/2048
Modulation Index	2	256	2^n
Payload Length (bytes)	94	1504	10
Symbol Rate (Kbps)	62.5	1000	62.5

*SNR at receiver controlled by adjusting transmitter power

The algorithms tested in this paper are found in the standard MATLAB global optimization toolbox. The algorithms tested were limited to those that could havea specified amount of function evaluations. The Generic Pattern Search (GPS), Mesh-Adapative Direct Search (MADS), Genetic Algorithm (GA), and MATLAB's fminsearch, which employs the Nelder-Mead simplex direct search algorithm, were all tested with a subset of possible parameters for each one.

Table 3. Parameter estimate statistics

Index	Mean	SD	Minimum	Maximum
CE Measures	0.03	3.01	-5.58	6.30
SE(CE)	0.66	0.61	0.29	1.06
Task Measures	0.00	0.84	-1.16	1.88
SE(Task)	0.17	0.01	0.16	0.29

The objective function used in these simulations is the weighted sum of multiple objective functions as in [8]:

$$f_{objective} = w_1 f_{minBER} + w_2 f_{maxthroughput} + w_3 f_{minpower}$$

where f_{minBER} , $f_{maxthroughput}$, and $f_{minpower}$ are given by equations (5.11), (5.14), and (5.18) of [8], respectively.

Each of the algorithms was limited to 30 function evaluations and tested with 20 different sets of weights that correspond to different priorities. For each test, a key, assumed to closely approximate the optimal solution, was generated using MATLAB's fminsearch with no specified maximum number of function evaluations.

The quality of each CE's performance was evaluated by calculating the Euclidean distance between the parameters of the key and CE's solution. Each parameter is normalized so that 0 corresponds to the lower bound of that parameter and 1 corresponds to the upper bound of that parameter. Given that there were four parameters to change, the maximum distance from the correct answer is 2.

4.2 Rasch Analysis

The simulated Euclidean distances were transformed into four ordinal categories ranging from 0 to 3 (scores, X_{ni}), and parameters for a polytomous version of the Rasch model were estimated based on these scored data using the *Winsteps* computer program [14]. This application produced a measure and associated standard error for each test task (δ_i) and each CE (θ_n). In addition, several additional indices were computed—a reliability coefficient for the CE measures, test task fit statistics, and CE fit statistics.

5. PROOF OF CONCEPT RESULTS

Table 3 displays the mean and standard deviation (SD) of the CE and test task measures and their standard errors (SE). The mean of the CE measures equals 0.03, and the mean of the task measures equals 0.00, suggesting that the simulated tasks were well targeted to the simulated CEs (i.e. the tasks were not overly difficult nor easy for the CEs to solve). In addition, the amount of error in the CE measures is fairly small compared to the variability of the CE measures. That is, the mean standard error of the CE measures is only about 22% of the size of the CE measure standard deviation. This

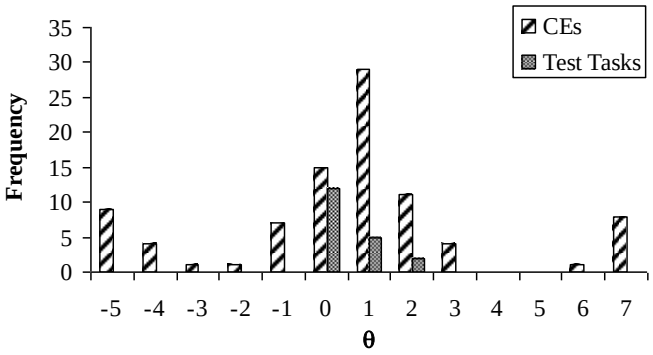

Figure 2: CE and Task Measures

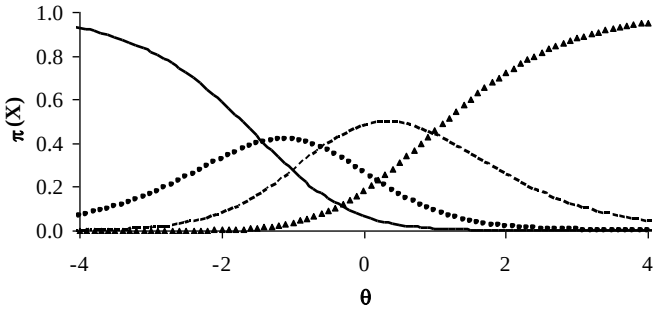

Figure 3: Probabilty Curves for Test Task 1

is supported by the fact that the reliability coefficient of the CE measures is fairly high (reliability = .91). However, it is also worth noting that there is considerable variability in the magnitudes of the standard errors of the individual CE measures as demonstrated by the minimum and maximum values of those measures. This suggests that, while some CEs were measured fairly precisely, some engines were not measured precisely at all by this battery of test tasks.

Figure 2, which displays the joint distributions of the CE and task measures, provides some explanation for the wide range of CE standard errors. This figure illustrates that although the tasks were not overly difficult or easy for the CEs to solve, the task measures cover only a narrow range of the of the performance capabilities of the CEs. In fact, 8 of the 91 engines received scores of 3 on all of the 20 tasks, and 10 of the CEs received scores of 0 on all of the tasks. Because the test tasks were not well suited for measuring these two relatively extreme groups of CEs, the standard errors for those CEs were large when compared to the remaining CEs.

Figure 3 displays the predicted probability of a CE receiving a score in each of the four score categories as a function of the CE's measure. On the leftmost section of the figure, the model-based probability that a CE will receive a score of 0 (solid curve) approaches 1.00, although the probability of a score of 1 (dotted curve) is not remote.

Table 4. Parameter estimate statistics

Index	Mean	SD	Minimum	Maximum
Task UMS	1.04	0.40	0.62	2.38
CE UMS	1.04	0.76	0.28	5.46

As CE measures increase in value, the probability of scores of 0 and 1 decreases as the probability of a scores of 2 (dashed curve) and 3 (triangle curve) increases. An important feature of this figure is the fact that each rating category is the most probable category within a range of CE measures, meaning that the rating scale applied to the Euclidean distance measures preserved the meaning contained in those original data values.

These model-based probabilities can be averaged at any particular value of θ_n to compute the expected value (E_{ni}) upon which CE and test task fit statistics are based. For example, the difference between the observed score for each CE and the conditional model-based expected value (i.e. the model-based residual) can be computed and then standardized. The average of these squared standardized residuals is referred to as the *unweighted mean square fit statistic* (UMS_i) for the task in question.

Table 4 summarizes the task and CE UMS values. The UMS values for the 20 items range from a low of 0.62 to a high of 2.38. Historically, rule-of-thumb lower and upper limits for UMS values have been set to 0.60 and 1.40 [15]. In the simulated data, two tasks exhibit misfit (i.e. UMS > 1.40)—Tasks 14 and 15. Task 14 (which weighted BER 5%, Power 90%, and Good Throughput 5%) and Task 15 (BER 5%, Power 5%,, Good Throughput 90%) were two of the most difficult test tasks, perhaps because of the highly unequal weightings. It may be that a few unexpected successes by a few of the low-performing CEs caused large residuals that inflated this test task's fit statistic.

The CE fit statistics have a larger range than do the test task fit statistics. In all, 13 of the 91 simulated CEs exhibited inflated levels of misfit (i.e. UMS > 1.40). It is interesting to note, however, that the majority of these CEs utilized MADS.

In fact, these particular CEs do stand out as being substantively different from the remaining CEs. A subsequent principal component analysis (PCA) of the residuals revealed that most of the MADS CEs not only exhibit misfit from the model, but they also exhibit consistency between engines in the patterns of their residuals. That is, although the patterns of scores associated with these CEs do not behave in a manner that is consistent with other CEs across the test tasks, these CEs exhibit patterns of scores that are consistent with one another. When such violations of the local independence assumption around which most IRMs are built occur, a common explanation is that the measures are multidimensional, meaning that multiple latent traits are necessary to fully explain the performance of the CEs.

6. CONCLUSION

Application of psychometric methods, particularly use of item response models, to testing of cognitive radios has potential to improve testing efficiency and effectiveness. Extensive investigation is needed to determine how best to apply these approaches and to assess their effectiveness in a variety of CR testing scenarios. As a first step towards this goal, this paper has presented a preliminary proof-of-concept study on application of psychometric methods to evaluation of cognitive engines.

Although this study is simulation based, it demonstrates that application of IRMs may hold promise as a medium for evaluating CEs. IRMs not only provide a mathematical model that describes how the entity being tested is likely to interact with the testing apparatus, but in doing so, IRMs provide diagnostically useful information to those who conduct the testing. Specifically, examination of the joint distributions of the CEs and test tasks allows testers to identify the types of testing tasks that are easiest and most difficult for CEs to respond to effectively. This type of information may also help developers of CEs identify the areas for improvement in a particular CE. Similarly, examination of model-data fit statistics may provide insights to those conducting tests and those developing CEs concerning how groups of algorithms behave differently in testing environments. Future work should extend the methods developed here to real CEs as well as to cases in which entire CRs are tested and the CEs' internal states may not be directly available, for the purpose of determining what potential barriers may exist to bridging the gap between simulation and reality.

The results of our simulation illustrate several useful pieces of information that IRMs can provide to those testing CEs. First, the fact that CE and test task measures are conjoint allows those who test CEs to determine the suitability of the test tasks chosen for a particular application. In our example, it is clear that the test tasks are fairly homogeneous in terms of difficulty and that several of the CEs responded very well or very poorly to all of those tests as demonstrated by the clusters of CE measures in the two tails of the θ_n distribution. In future tests of these CEs, it would be useful to add more variability to the battery of test tasks in order to make the difficulty of those tasks more variable, allowing for more precise measurement of the CEs in these two groups.

Second, the rating scale that we created based on the Euclidean distance measures provides useful information for differentiating the levels of performance of the CEs. In Figure 3, each of the four ordinal score categories is the most probable outcome for a range on the underlying CE

measure continuum. In some applications, one or more of these curves fails to be the most likely outcome at any point on the continuum—a result that calls into question the usefulness of the chosen scoring algorithm. In fact, we could have applied the IRM to the original Euclidean measures. However, it would have been nearly impossible to estimate the model's parameters due to the small number of observations that would have been associated with any particular score category. Hence, the chosen scoring algorithm achieved a useful balance between having too many categories to provide useful parameter estimates and too few categories to provide useful information about the CEs.

Third, the model-based expectations and the associated fit statistics pointed to substantively interesting differences between the performance of the CEs. Specifically, the mean square fit statistics pointed to two tasks that elicited CE responses that were not quite consistent with the responses that the remaining tasks elicited. In other words, the rank ordering of the performances of the CEs was different for these two test tasks than it was for the remaining tasks. Again, such differences point out potentially important features of the tasks chosen to test the CEs—differences that can guide future CE development, refinement, and testing efforts.

Fourth, in a similar vein, the model-based fit evidence points to potentially important differences between groups of the simulated CEs. Specifically, the MADS engines exhibited inflated misfit values. That is, the model-based residuals for these CEs were larger than were the residuals for the remaining CEs. This fact, in isolation, suggests only that these CEs are not measured in the same way as the remaining CEs by this set of test tasks. However, further evidence, in the form of correlations among these model-based residuals for these CEs, suggests that the MADS actually perform similarly across all of the test tasks. That is, while most of the CEs provide one rank ordering of the difficulties of the test tasks, the MADS **e**ngines jointly provide a different rank ordering of the difficulties of those test tasks. Commonly, results such as these imply that there are multiple characteristics being measured by the test tasks (i.e. that the latent trait is multidimensional, rather than unidimensional). This fact is informative to those who develop CEs and those who test CEs because it suggests that CE refinement and evaluation may need to focus on more than one CE characteristic in order to more accurately depict the quality of a particular engine's performance.

ACKNOWLEDGMENT

Thanks to Cecile Dietrich for suggesting this collaboration.

This work was supported in part by the National Science Foundation under Grant 0851400. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Authors Dietrich and Wolfe share first authorship.

REFERENCES

[1] J. Mitola III, *Cognitive Radio: An Integrated Agent Architecture for Software Defined Radio* [Dissertation], Royal Institute of Technology (KTH), Stockholm, Sweden, May 8, 2000.

[2] S. Haykin, "Cognitive radio: brain-empowered wireless communications," *IEEE JSAC*, vol. 23, no. 2, pp. 201-220, Feb. 2005.

[3] S. Soliman, (2004, November 1). Cognitive radio: key performance indicators. *BWRC Cognitive Radio Workshop* [Online]. Available: http://bwrc.eecs.berkeley.edu/Research/MCMA/CR%20Workshop/ssoliman_BWRC_CR_workshop.pdf

[4] Y. Zhao, S. Mao, J.O. Neel, and J.H. Reed, "Performance Evaluation of Cognitive Radios: Metrics, Utility Functions, and Methodology, Proceedings of the IEEE, Vol. 97, No. 4, pp. 642-659, April 2009.

[5] *Application Note: Testing Modern Radios: solutions for designing software defined radios that employ legacy and modern modulation schemes with frequency hopping techniques* [Online]. Tektronix. Available: http://www2.tek.com/cmsreplive/tirep/12622/37W_21488_1_HR_2010.12.30.12.48.13_12622_EN.pdf

[6] *Application Note: Installed Radio Testing with the 3500* [Online]. Aeroflex. Available: http://www.aeroflex.com/ats/products/prodfiles/appnotes/3500RT.pdf

[7] J. Riihijärvi and R. Agustí, Eds., *Flexible and Spectrum-Aware Radio Access through Measurements and Modelling in Cognitive Radio Systems, FARAMIR Document Number D2.1: State of the Art Review, April 30, 2010.* Available: http://www.ict-faramir.eu/fileadmin/user_upload/deliverables/FARAMIR-D2.1-Final.pdf

[8] T.R. Newman, S.M.S. Hasan, D. DePoy, T. Bose, J.H. Reed, "Designing and deploying a building-wide cognitive radio network testbed," *IEEE Communications Magazine*, Sept. 2010.

[9] B.D. Wright and G.N. Masters, *Rating Scale Analysis: Rasch measurement*. MESA, Chicago, 1982.

[10] P. DeBoeck and M. Wilson, *Explanatory Item Response Models*, Springer, New York, 2004.

[11] D.C. Briggs and M. Wilson, "An Introduction to Multidimensional measurement using Rasch models," *Journal of Applied Measurement*, vol. 4, no. 1, pp. 87-100, 2003.

[12] G. Rasch, *Probabilistic Models for some Intelligence and Attainment Tests*, University of Chicago, Chicago, IL, 1980.

[13] T.R. Newman, *Multiple Objective Fitness Functions for Cognitive Radio Adaptation*, Ph.D. dissertation, University of Kansas, 2008.

[14] J.M. Linacre, *WINSTEPS Rasch measurement computer program* (Version 3.71.0). Winsteps.com, Chicago, 2011.

[15] B.D. Wright and M. Linacre, Reasonable mean-square fit values. *Rasch Measurement Transactions*, vol. 8, p. 370, 1994.