

Spring 2-18-2005

## Item Analysis of the Interpreter Competency Exam- Mental Health (ICE-MH)

Melinda Pearson

Follow this and additional works at: <https://digitalcommons.georgefox.edu/psyd>



Part of the [Clinical Psychology Commons](#)

---

Item Analysis of the Interpreter Competency Exam – Mental Health (ICE-MH)

by

Melinda Pearson

Presented to the faculty of the

Graduate Department of Clinical Psychology

George Fox University

in partial fulfillment

of the requirements for the degree of

Doctor of Psychology

in Clinical Psychology

Newberg, Oregon

February 18, 2005

Item Analysis of the Interpreter Competency Exam – Mental Health (ICE-MH).

by

Melinda Pearson

has been approved

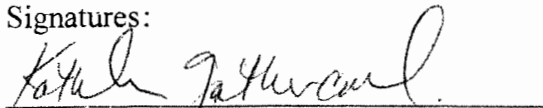
at the

Graduate School of Clinical Psychology

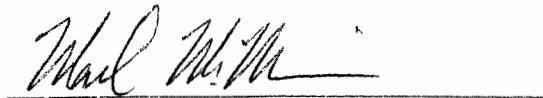
George Fox University

as a Dissertation for the PsyD. degree.

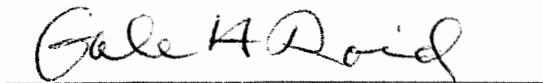
Signatures:



Kathleen Gathercoal, Ph.D., Chair



Mark McMinn, PhD



Gale Roid, PhD

Date: 3.17.08

Item Analysis of the Interpreter Competency Exam – Mental Health (ICE-MH)

Melinda Pearson

Graduate Student of Clinical Psychology at

George Fox University

Newberg, Oregon

Abstract

Interpreter services are becoming increasingly prevalent in mental health settings. The competence of mental health interpreters can be assessed using a new tool, the Interpreter Competency Exam – Mental Health (ICE-MH). Unfortunately, the current form of the ICE-MH is long. This study used archival data from the ICE-MH to assess (a) the ability of the items to discriminate the examinees' level of knowledge in mental health areas, (b) the difficulty of each item, measured by percentage of correct answers, (c) the internal consistency of each part of the test to verify the uniformity of the subtest constructs being measured, and (d) the composite score reliability index for the test as a whole. The item analysis resulted in a reduction in the number of items from 442 to 98. The composite score reliability index of the resulting measure was high (.96). Suggestions for further analyzing and refining the instrument include a review of remaining items for clarity and ease of understanding (e.g., possibly removing "all of the above" questions), conducting validity studies (e.g., samples of interpreters who are known to have the desired skills could be used as well as coupling this assessment with the results of periodic performance reviews), using a Rasch analysis on a larger sample of responses, and finally, performing a cross-validation of the test on new samples.

## Table of Contents

Approval.....	ii
Abstract .....	iii
Table of Tables.....	v
Chapter 1 Introduction.....	1
Chapter 2 Methods.....	7
Participants .....	7
Instruments .....	8
Interpreter Competency Exam – Mental Health .....	8
Procedure .....	8
Chapter 3 Results.....	9
Item-Discrimination Index.....	9
Item-Difficulty Index .....	10
Internal Consistency Reliability.....	11
Composite Score Reliability Index .....	11
Chapter 4 Discussion.....	14
Recommendations for Modifying the ICE-MH.....	14
Limitations of the Present Study.....	15
Recommendations for Future Research .....	15
References .....	18
Appendix A Results of Item Pruning in the Definitions Subtest.....	20
Appendix B Results of Item Pruning in the Concepts Subtest.....	30
Appendix C Results of Item Pruning in the Ethics Subtest.....	33
Appendix D Examples of Well-Worded and Poorly-Worded Items from the ICE-MH.....	36
Appendix E Curriculum Vita.....	38

Table of Tables

Table 1	Beginning internal consistency analyses (based on 442 total items in the test) and ending analyses (98 total items) using Cronbach's Alpha per subtest.....	11
Table 2	Mean scores on the three subtests of the ICE-MH.....	12
Table 3	Correlations between the final versions of the subtests of the ICE-MH used in the calculation of composite reliability.....	13

## Chapter 1

### Introduction

The United States Census Bureau (as cited in Dysart-Gale, 2007) estimates a population of 21 million in the United States to have limited English proficiency. This poses a significant barrier to accessing medical and mental health treatment. Resolving this language gap is essential to the provision of medical and mental health services and the use of interpreters is one method of increasing accessibility (Dysart-Gale, 2007).

In the field of clinical psychology, there has been increased attention in the last 30 years to the provision of services tailored to diverse cultures. Stanley Sue began his crusade for cultural competency in the 1970s and noted in 2003 that, although research still has methodological limitations, there is substantial evidence for cultural competency among psychologists. This same year the American Psychological Association formalized its policies regarding providing culturally competent mental health services when it published “Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists.” These guidelines encourage psychologists to respect their client’s language preference, obtain informed consent, and to use interpreters who are neither family members nor community leaders and who are skilled in the mental health field (APA, 2002, 2003).

Finding interpreters who meet these recommendations is a challenge. Wong et al. (2006) identified access to interpretative services as one of the structural barriers for Cambodian

refugees seeking mental health services. The Department of Health in the United Kingdom reports interpreters having been known to “omit, add, condense, simplify, and substitute facts” (as cited in Meddings & Haith-Cooper, 2008). Flaskerud (2007) adds normalizing client’s thoughts and attempting to prevent ethnic stigma by answering for the client as other treatment-compromising behaviors in which some interpreters engage.

For the mental health provider, the lack of competent interpreters is worrisome. It is important for mental health interpreters to be familiar with the specialized language and concepts important in mental health services. The APA Ethical Code of Conduct also notes a responsibility of the health care provider to insure that an interpreter provides competent services and follows ethical guidelines such as confidentiality (APA, 2002). The interpreter also has a code of ethics to join with that of the mental health provider (Thompson & Blasquez, as cited in Lehna, 2005). Miller, Martell, Pazdirek, Caruth, and Lopez’s (2005) study recommended a model of training for interpreters that focused on mental health practice and theory prior to working in those settings. The mental health provider needs to be confident that information is being conveyed correctly to the client, as the effectiveness of treatment is very dependent on the skills and knowledge of the interpreter. As of this time there is no published assessment tool to help assure a mental health provider of an interpreter’s ability to provide effective services within the proposed guidelines above. The development of such an assessment tool is necessary and Barsukov (2005) undertook the development of a tool for Russian-language interpreters.

Gregory (2004) outlined standards for developing psychometrically-sound assessment tools. The first step is to define the test so that there is a clear understanding of its purpose and measurement. This includes an appropriate domain of content including material chosen and



target population (Nunnally, 1994). A new test should fill a need in the field it is intended to service. Once the purpose is clearly outlined, the scaling of the test and the type of item construction should be determined (Gregory, 2004). A large number of items are then generated for the pilot administration in order to increase the reliability of the test (Nunnally, 1994).

Barsukov's test, the *Interpreter Competency Exam – Mental Health* (ICE-MH) was a paper-and-pencil test developed to assess the competency of Russian-language interpreters for mental health services (Barsukov, 2005). In this test, participants were asked to complete, in writing, the assembled booklet of questions written in English. The rational scaling method was chosen and items were constructed with various difficulties, following three types of questions: multiple choice, short answer and true-false. The initial version of the ICE-MH includes subtests to assess for translation of terminology and concepts as well as a set of scenarios with questions to demonstrate an understanding of required ethics.

The first section of the test required the participant to translate 359 words commonly used in mental health treatment from English into the target language of Russian. When words were not translatable literally, the examinees were allowed to provide a brief explanation of the term. The second section included 40 multiple choice questions concerning key mental health concepts. This section required the interpreter to choose the best answer among five choices provided. The final section of the test consisted of ten short scenarios, each involving an ethical dilemma. After reading through the paragraph, the examinee was asked to answer yes or no to 4-5 questions regarding what he or she would do in that particular situation.

It is clear that ICE-MH showed promise as tool in the mental health field. The initial study administration resulted in coefficient alphas ranging from .82 to .98 (Barsukov, 2005). Subject matter experts were used to evaluate the content validity of items prior to inclusion of

the assessment tool and criterion validity analysis showed high correlations between the participant's verbal translations and the definitions and concepts subtests. The correlation with the ethics subtest was low as the content is conceptually dissimilar.

Gregory (2004) also suggested that after a tool with sufficient purpose, length, and promising initial reliability and validity has been pilot tested, later analysis should determine which items might be most effective for the purposes of the tool. After the first administration is completed, analyses can be run on the resulting data to determine the usefulness and necessity of individual test items (Gregory, 2004). Then the tool is revised, and the resulting set of items is administered to a new sample and the analysis can continue through several iterations until the tool is sufficiently refined (Nunnally, 1994).

Currently, the ICE-MH is long; it includes 442 items which requires a significant amount of time to complete. The next step in the development of this assessment tool is to determine the value of each item in measuring the competence of the potential interpreter. The goal of this examination is to make recommendations regarding the elimination of items which contribute little or nothing to the ICE-MH results. Item analyses will include (a) item discrimination, (b) item difficulty indexing, (c) calculation of internal consistency of the test, and (d) composite score reliability index.

The item-discrimination index describes how efficient an item is in identifying those translators who perform poorly on a test and those who do well. The product moment item-total correlation can provide discrimination by comparing the score on a specific item achieved by participants who did well overall with those who did not do well overall. Using the item-discrimination index, one can determine the role each item plays in determining subjects' ability to score well on this test of translation skill. By analyzing the data in SPSS, the item can be

correlated with the sum of the remaining test items to reduce overlap in the item discrimination (Nunnally, 1994).

The item-difficulty index identifies unproductive items that are either too difficult or too easy and, therefore, are psychometrically unnecessary in the test. When choosing the appropriate level of item difficulty, it is necessary to decide how discriminating the interpreter selection process should be. The multiple-choice section of the test also should be analyzed for the role of chance in the results due to guessing. Item-difficulty is examined by the mean scores for each item. Gregory (2004) recommends difficulty levels between 0.3 and 0.7 to effectively discriminate among examinees. A few items of low and high difficulty will be retained for building testing rapport and minimizing ceiling effect (Roid, personal communication, February 28, 2007)

The internal consistency reliability (Gregory, 2004) of the ICE-MH should also be studied. The reliability index, developed and described by Cronbach (1951), was used to analyze the role of each item in the internal consistency of the test. Results were compared to professionally acceptable levels of reliability, such as .70 for research instruments, and levels of .80 and higher for measures used to make decisions about individuals (Roid, personal communication, January 31, 2007). Nunnally (1967) recommends not striving for reliabilities above 0.80 as the effort needed to increase the reliability would be excessively time consuming with only a minimal change to the measurement error (p. 226). Roid notes this cut-off as a standard of practice in test development for measures such as statewide achievement tests (personal communication, January 31, 2007).

The composite score reliability index is a way of measuring an assessment tool's reliability despite the fact that each subtest measures a different trait or construct. If the

reliability of the three subtests were simply added, the resulting estimate by coefficient alpha would be misleading. By using the composite score reliability index, or reliability of linear combinations, a more accurate reliability can be calculated (Nunnally, 1967).

It is clear that ICE-MH shows promise as a useful test to assess mental health interpreter competence. However, personal observation of participants in previous research suggests that the length of the test and the amount of time required to finish it are frustrating and would prohibit the use of the test as well as decrease performance (Barsukov, 2005). The current study is designed to refine the test by undertaking traditional psychometric analyses to suggest items that may be removed because they are too easy, do not correlate with other items or criteria, and do not discriminate between high and low performers.

## Chapter 2

### Methods

#### *Participants*

Archival records from the initial administration of the ICE-MH served as the data for this study. The test was given to applicants who were recruited through an advertisement in local Russian-language newspapers. The initial administration of the ICE-MH was completed by a convenient sample of 31 experienced native-Russian-speaking interpreters as part of an interview process for a private company. Four additional applicants failed to complete the ICE-MH as part of the interview process and their partial data were excluded from the analysis. The participants were tested on three dates and, because date of administration, did not significantly affect the outcome, the data were collapsed across days. Of those who completed the test, 25% were male with the sample's ages ranging from 21 to 56 years ( $M = 34$ ). The interpreters' employment ranged from working for interpretation companies (41%), working in psychiatric hospitals and mental health institutions (34%), and those who work as independent contractors (25%). Together they had a mean of 2.3 ( $SD = 2.2$ ) years of experience serving as an interpreter.

### *Instruments*

*Interpreter Competency Exam – Mental Health* (Barsukov, 2005) is a paper-and-pencil test developed to assess the competency of Russian language interpreters for mental health services. The ICE-MH has three sub-sections. In the first, interpreters are asked to translate 359 words commonly used in mental health treatment. In the second section, respondents answer 40 multiple choice questions concerning key mental health concepts. And in the third section, respondents answer 4-5 yes-no items for each of ten vignettes describing an ethical dilemma. The ICE-MH yields three subscale scores and a total score. As stated above, reliability was calculated for the three subtests individually and yielded coefficient alphas ranging from .92 to .98. A comparison of the subtest scores and expert evaluation of the required verbal translation shows a strong relationship between the definitions and concepts scores and verbal translation ability.

### *Procedure*

After giving their informed consent, the participants individually provided an oral translation of approximately one and a half typewritten pages before a panel of two raters, who scored them on scale from 1 (very inaccurate) to 5 (very accurate). After the translation task was completed, the instructions to the ICE-MH were read to each participant. Then each participant was taken into a quiet room with other participants to complete the written exam. Participants were given as long as they needed to complete the test. Upon completion, they were allowed to hand in the test, verify their contact information and leave.

## Chapter 3

### Results

This chapter presents the results of the four analyses: (a) item discrimination, (b) item difficulty, (c) internal consistency, and (d) composite reliability. The item-discrimination was analyzed by correlating each of the test items with total and subscale scores. Item difficulty was analyzed by calculating the pass rate for each item and identifying which items fall between the 30-70% rates. For item internal consistency, each item was correlated with total and subscale scores using Cronbach's alpha. For composite reliability, formulas from Nunnally (1967) were used to calculate the reliability of the combination of the three subtests, taking into account the intercorrelations among the subtests.

#### *Item-Discrimination Index*

The item-discrimination index describes how efficient each item is in identifying persons who perform poorly on a test and those who do well. Discrimination analysis of the ICE-MH was conducted by correlating each of the test items with total and subscale scores. The corrected item-total correlation provides information about how much the individual item and the measure correlate (Murphy, 2001).

Each of the three subtests was examined using the SPSS scale reliability analysis routine that calculates discrimination indexes for each item. The items of the definitions,

concepts, and ethics subtests as well as the results of the pruning process appear in Appendices A, B, and C. The definitions subtest originally contained 359 items. Discrimination analysis led to the removal of all items with a corrected item-total correlation of 0.55 and less, thus leaving 38 items. This new list was examined again to verify the discrimination quality of the remaining items, which were subsequently retained. The next subtest examined was concepts. The initial list of 40 items was pared down to 28 items when those with an item-total analysis of less than 0.2 were removed (Roid, 1995). The item-total correlation was run on the new list and two additional items with correlations below 0.2 were removed. After one final discrimination analysis, one more item with the lowest score was removed leaving a convenient number of 25 items. The ethics subtest began with 43 items. Discrimination analysis led to the removal of eight items with corrected item correlations of .22 or less, leaving a convenient number of 35 items.

#### *Item-Difficulty Index*

After the subtests were pruned using discrimination analysis, the remaining items were examined for difficulty. After items with low discrimination were removed, the difficulty of the remaining items was analyzed making sure the means were primarily scattered between .3 and .7 with a few outliers for testing warm up and ceiling effects. On the definition subtest, the means of only three items fell outside of the 0.1 to 0.9 margin. These three items were retained to use as warm-up items since the definitions subtest falls at the beginning of the ICE-MH. The difficulty of the 25 remaining concept items and the 35 remaining ethics items fell within 0.1 and 0.9, so no further items were dropped.



*Internal Consistency Reliability*

Next each subtest was examined for internal consistency to verify the measurement of the construct. The definitions subtest initially had a Cronbach's Alpha of 0.98. Reducing the number of items lowered the internal consistency to 0.95, which is still well within acceptable range. The Cronbach's Alpha for the concepts and the definitions subtests both increased in internal consistency (see Table 1).

Table 1

*Beginning internal consistency analyses (based on 442 total items in the test) and ending analyses (98 total items) using Cronbach's Alpha per subtest*

Subtest	Initial Cronbach's Alpha	Resulting Cronbach's Alpha
Definitions	0.978	0.954
Concepts	0.824	0.841
Ethics	0.919	0.929

*Composite Score Reliability Index*

Final versions of the shortened subtests were scored for all subjects. The number of items, means, and standard deviations for the final versions of the subtests are shown in Table 2.

When product moment correlations were calculated among the three subtests, they were found to have correlations ranging from .41 to .64 (see table 3). However given the diversity of content between the three subtests, we would not expect to have a high correlation.

Calculation of Cronbach's alpha for all of the items in all three subtests would not supply an accurate picture of the test's overall reliability. The fact that each subtest is measuring a different

Table 2

*Mean scores on the three subtests of the ICE-MH*

Subtest	Univariate Summary		
	<u>Number of Items</u>	<u>M</u>	<u>SD</u>
Definitions	38	23.40	10.78
Concepts	25	15.04	5.48
Ethics	35	23.36	6.18

*Note.*  $N = 31$ . Subtest scores are calculated as the total number of items correctly answered (1 point for each question).

construct must be taken into account in the statistical analysis (Nunnally, 1967). A more accurate reliability of the three subtests as a group resulted when an analysis of the reliability was based upon a linear combination of the three subtest scores. Nunnally (1967) describes this statistical process takes into account the diversity of content of subtests and the intercorrelations among the subtest scores in estimating the combined reliability of the test components. Roid (1991) developed a FORTRAN program to assist with this statistical analysis, implementing the formulas provided in Nunnally (1967). Input of this data produced an overall composite reliability index of 0.96.

Table 3

*Correlations between the final versions of the subtests of the ICE-MH used in the calculation of composite reliability*

	Definitions	Concepts	Ethics
Definitions	1.00		
Concepts	.41*	1.00	
Ethics	.64**	.59*	1.00

*Note.*  $N = 31$ .

\*  $p < .05$  \*\*  $p < .01$

## Chapter 4

### Discussion

The resulting data from the pilot administration of the ICE-MH were analyzed for item discrimination, item difficulty, internal consistency of each subtest, and composite reliability of the measure. Each subtest was analyzed individually, using accepted standards of practice and items which were too difficult, too easy, or unhelpful to the outcome of the assessment tool were eliminated thereby reducing the number of items from 442 to 98. (See Appendices A, B, and C for the list of 442 original items by subtest and how they fared in the pruning process.) The remaining items were then analyzed for reliability of linear combinations with a resulting reliability index of 0.96.

#### *Recommendations for Modifying the ICE-MH*

The first recommendation from this study is to reduce the length of the ICE-MH measure by removing the suggested items from the original item pool. The shortened version will have acceptable reliability, be easier to score and is less likely to cause testing fatigue.

A next recommendation for refining the test is to examine the remaining items to see if they may need to be reworded for clarity and to reduce cognitive load. For example, this might include omitting “all of the above” response options and checking for limiting psychological language that is unnecessary in the purpose of the measure. In his *Process of Test*

*Development* chapter, Murphy (2001) provides examples of poorly and well-written items.

Examples of poorly and well-written items from the ICE-MH are provided in Appendix D.

### *Limitations of the Present Study*

The pilot administration of the ICE-MH resulted in only 31 usable sets of data. This limits the generalizability of research findings as well as the types of analyses that can be run on the data. An increased sample size would lend more credibility to the use of the ICE-MH as an effective assessment tool.

The sample for the present study consisted of persons who responded to an advertisement in a local Russian-speaking newspaper. They were not required to have previous experience with interpretation or translation. The findings of the psychometric properties of the ICE-MH could be more finely tuned if the theoretical population and the sample frame were more similar.

While the results of the pilot administration were compared to the oral translation of standard paragraphs before experts, it would be helpful if there were better measures of interpreters' "on-the-job" translation abilities for validation of the test. Additionally, the length of the test (359 items plus the oral translation component) led to examinee fatigue and frustration, likely impairing the accuracy of the results.

### *Recommendations for Future Research*

The previous administration of the ICE-MH used a small sample ( $N = 31$ ) of people who were looking for work as interpreters, however demonstrated skill and experience with interpretation or translation was not required. A future study might use a larger sample, only including interpreters who have established a good reputation within the community of their practice.

For this measure to be useful in hiring, it should also be true that a person who

performs well on this measure should perform well in his or her interpreter role. Using periodic performance evaluations to follow-up new hires screened by the ICE-MH would provide information about the predictive validity of the measure. A performance review would likely take into account more than just interpretation ability (e.g., the ability to interact well with superiors and subordinates). Miller et al. (2005) pointed out that multi-linguistic understanding was only one skill of a competent interpreter. Because the interpreter changes the dyadic format to a triadic relationship, their role in the room is not limited to merely interpretation. These researchers suggest that good mental health interpreters have similar characteristics to effective psychotherapists such as empathy, interpersonal skills, and psychological mindedness. If these characteristics are to be used as criteria for hiring, it may be beneficial to add a component to the ICE-MH to evaluate these characteristics.

The current study found the definitions and concepts subtests to have more consistent inter-reliability with each other than with the ethics subtest. This is attributed to the conceptually dissimilar content of the ethics test. It may be more useful to develop the ethics subtest as an additional screening tool outside of the core ICE-MH test. Additionally, as reported previously, interpreters who perform well in mental health settings have similar personal characteristics to mental health professionals. A screening tool for these characteristics might be helpful in hiring decisions.

Prior to publication, the measure should be studied further to verify the reliability and validity of the ICE-MH in other samples. If future studies included prediction of the interpreter's actual performance on the job, the ICE-MH scores (particularly Definitions and Concepts) could be used as predictor variables. The process of evaluating the predictive validity of the test could include "cross validation" where the regression equation from one

sample is applied to another new sample to verify the test's predictive dependability (Gregory, 2004). When the measure in the second sample is less predictive of the criterion, an event known as validity shrinkage occurs. To minimize this, it is recommended that larger samples be used when testing and cross-validating the shortened version of the ICE-MH (Gregory, 2004).

Rasch analysis, based on item-response theory (IRT; Embretson & Reise, 2000; Rasch, 1980) also is recommended for future evaluation for several reasons. Rasch analysis describes the attributes of the individual being assessed independent of the specific items selected for use because estimates of ability can be obtained from any subset of items, as is done in computer-adaptive testing (Embretson & Reise, 2000). WINSTEPS (Wright & Linacre, 1999), a Rasch analysis program, for example, provides a wide range of item information including item difficulty (calculated to be more sample-independent than conventional percentage-correct difficulty) and various statistics showing the fit of the item data to an ideal pattern relating item performance to the attribute-level of the respondents (Embretson & Reise, 2000). Also, Rasch analysis allows each item of the scale to be displayed showing which items are at high, medium, and low levels of the measured attribute. In other words, the items of the scale contributing most to interpreter expertise would be shown at the highest level and vice versa. The Rasch analysis requires a minimum sample of 200 participants, again making a larger sample a recommendation for future research.

## References

- American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved March 1, 2007 from <http://www.apa.org/ethics/code2002.html>
- American Psychological Association. (2003). *Guidelines on multicultural education, training, research, practice, and organizational change for psychologists*. Retrieved March 1, 2007 from <http://www.apa.org/pi/multiculturalguidelines.pdf>
- Barsukov, S. (2005) *The Interpreter Competency Exam – Mental Health (ICE-MH)*.  
Unpublished test and manual; George Fox University, OR.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Dysart-Gale, D. (2007). Clinicians and medical interpreters: Negotiating culturally appropriate care for patients with limited English ability. *Family Community Health*, 30 (3), 237-246.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Routledge/Erlbaum.
- Flaskerud, J. H. (2007). Cultural competence column: What else is necessary? *Issues in Mental Health Nursing*, 28 (2), 219-222. Retrieved June 17, 2007 from EBSCOhost.
- Gregory, R. (2004). *Psychological testing: History, principles, and applications* (4<sup>th</sup> ed.). Boston: Allyn and Bacon.
- Lehna, C. (2005). Interpreter services in pediatric nursing. *Pediatric Nursing*, 31 (4), 292-296. Retrieved March 1, 2007 from EBSCOhost.
- Meddings, F. & Haith-Cooper, M. (2008). Culture and communication in ethically appropriate care. *Nursing Ethics*, 15 (1), 52-61. Retrieved January 28, 2008 from EBSCOhost.



- Miller, K. E., Martell, Z. L., Pazdirek, L., Caruth, M., & Lopez, D. (2005). The role of interpreters in psychotherapy with refugees: An exploratory study. *American Journal of Orthopsychiatry*, 75 (1), 27-39. Retrieved March 1, 2007 from EBSCOhost.
- Murphy, K. R. & Davidshofer, C. O. (2001). *Psychological testing: Principles and applications* (5<sup>th</sup> ed.). Upper Saddle River, NJ: Prentice Hall.
- Nunnally, J. (1967). *Psychometric Theory*. New York, NY: McGraw-Hill.
- Nunnally, J. (1994). *Psychometric Theory*. 3<sup>rd</sup> ed. New York, NY: McGraw-Hill.
- Rasch, G. (1980). Some probabilistic models for intelligence and attainment tests. Chicago, IL: University of Chicago Press.
- Roid, G. H. (1991). A FORTRAN program for calculating composite reliability [Computer software]. Unpublished program, Southern Methodist University, School of Education.
- Roid, G. H. (1995). Measurement and test-development issues in the Oregon Statewide Assessment of reading and mathematics (Grades 3, 5, 8, and 11): Final report. Salem, OR: Assessment Research.
- Sue, S. (2003). In defense of cultural competency in psychotherapy and research. *American Psychologist*, 964-970. Retrieved March 1, 2007 from PsycARTICLES.
- Wong, E. C., Marshall, G. N., Schell, T. L., Elliott, M. N., Hambarsoomians, K., Chun, C., & et al. (2006). Barriers to mental health care utilization for U.S. Cambodian refugees. *Journal of Consulting & Clinical Psychology*, 74 (6), 1116-1120. Retrieved March 1, 2007 from PsycARTICLES.
- Wright, B. D., & Linacre, J. M. (1999). WINSTEPS: Rasch analysis for Windows manual [Computer software]. Chicago, IL: MESA Press.

Appendix A

Results of Item Pruning in the Definitions Subtest

## Appendix A

Results of pruning items in the Definitions subtest as a result of discrimination analysis and item difficulty analysis.

Item #	Discrimination Analysis	2nd Discrimination Analysis	Item Difficulty
1	0.161		
2	0.057		
3	0.381		
4	0.358		
5	0.421		
6	0.509		
7	0.235		
8	0.318		
9	0.007		
10	0.599	0.511	0.84
11	0.078		
12	0.466		
13	0.567	0.577	0.66
14	0.321		
15	0.291		
16	0.182		
17	0.264		
18	0.313		
19	0.400		
20	0.249		
21	0.466		
22	0.249		
23	0.452		
24	0.576	0.589	0.84
25	0.153		
26	0.000		
27	0.141		
28	0.501		
29	0.352		
30	0.149		
31	0.260		
32	0.135		
33	0.081		

34	-0.312		
35	0.457		
36	0.086		
37	-0.062		
38	0.139		
39	0.132		
40	0.052		
41	0.216		
42	0.561	0.550	0.88
43	0.260		
44	0.127		
45	0.000		
46	0.417		
47	0.000		
48	0.503		
49	0.229		
50	0.219		
51	0.158		
52	0.377		
53	0.000		
54	0.000		
55	0.337		
56	0.362		
57	0.160		
58	0.181		
59	0.172		
60	0.502		
61	0.358		
62	0.077		
63	0.226		
64	0.392		
65	0.412		
66	0.013		
67	0.159		
68	0.266		
69	0.287		
70	0.334		
71	0.407		
72	0.305		
73	0.457		
74	0.103		

75	0.127		
76	0.312		
77	0.285		
78	0.309		
79	0.345		
80	0.223		
81	0.327		
82	0.536		
83	0.587	0.581	0.91
84	0.321		
85	0.102		
86	0.032		
87	0.370		
88	0.171		
89	0.475		
90	0.708	0.707	0.56
91	0.120		
92	0.143		
93	0.557	0.565	0.59
94	-0.092		
95	0.276		
96	0.267		
97	0.276		
98	0.301		
99	0.538		
100	0.467		
101	0.311		
102	0.311		
103	0.632	0.565	0.81
104	0.649	0.591	0.66
105	0.274		
106	0.295		
107	0.137		
108	0.388		
109	0.529		
110	0.137		
111	0.000		
112	0.394		
113	0.456		
114	0.508		
115	0.393		

116	0.000		
117	0.047		
118	0.298		
119	0.610	0.622	0.63
120	0.427		
121	0.501		
122	0.000		
123	0.301		
124	0.310		
125	0.127		
126	0.345		
127	0.345		
128	0.377		
129	0.299		
130	0.411		
131	0.714	0.722	0.78
132	-0.312		
133	0.168		
134	0.390		
135	0.407		
136	0.237		
137	0.047		
138	0.279		
139	0.467		
140	0.618	0.642	0.53
141	0.460		
142	0.407		
143	0.143		
144	0.364		
145	0.319		
146	-0.029		
147	0.351		
148	0.357		
149	0.135		
150	0.098		
151	0.370		
152	0.221		
153	0.455		
154	0.202		
155	0.350		
156	0.643	0.600	0.50

157	0.350		
158	-0.081		
159	0.212		
160	0.459		
161	0.320		
162	0.351		
163	0.507		
164	0.000		
165	0.287		
166	0.153		
167	0.467		
168	0.000		
169	0.564	0.506	0.94
170	0.281		
171	0.467		
172	0.455		
173	0.625	0.581	0.28
174	0.475		
175	0.469		
176	-0.020		
177	0.458		
178	0.458		
179	0.199		
180	0.303		
181	0.170		
182	0.355		
183	0.466		
184	0.294		
185	0.417		
186	0.413		
187	0.257		
188	0.579	0.489	0.66
189	0.401		
190	0.413		
191	0.258		
192	0.602	0.616	0.69
193	0.367		
194	0.535		
195	0.291		
196	0.578	0.604	0.59
197	0.280		

198	0.548		
199	0.449		
200	0.603	0.668	0.72
201	0.038		
202	0.533		
203	0.546		
204	0.444		
205	0.246		
206	0.237		
207	0.309		
208	0.467		
209	0.413		
210	0.212		
211	0.412		
212	0.292		
213	0.107		
214	0.339		
215	0.248		
216	0.528		
217	0.469		
218	0.315		
219	0.162		
220	0.213		
221	0.379		
222	0.529		
223	0.370		
224	0.585	0.561	0.50
225	0.298		
226	0.567	0.579	0.94
227	0.509		
228	0.615	0.577	0.38
229	-0.032		
230	0.630	0.575	0.34
231	0.318		
232	0.000		
233	0.089		
234	0.200		
235	0.513		
236	0.352		
237	0.598	0.557	0.81
238	0.587	0.637	0.59



239	0.339		
240	0.321		
241	0.339		
242	0.170		
243	0.235		
244	0.558	0.582	0.41
245	0.518		
246	0.272		
247	0.228		
248	0.627	0.577	0.53
249	0.467		
250	0.531		
251	0.454		
252	0.457		
253	0.315		
254	0.073		
255	0.556	0.574	0.81
256	0.432		
257	0.521		
258	0.593	0.524	0.28
259	0.390		
260	0.311		
261	0.301		
262	0.508		
263	0.161		
264	-0.200		
265	0.422		
266	0.303		
267	0.259		
268	0.398		
269	0.596	0.595	0.41
270	0.259		
271	0.116		
272	0.154		
273	0.467		
274	0.522		
275	0.383		
276	0.532		
277	0.406		
278	0.494		
279	0.560	0.574	0.69

## Item Analysis of the ICE-MH 28

280	0.475		
281	0.574	0.580	0.84
282	0.271		
283	0.269		
284	0.270		
285	0.109		
286	0.413		
287	0.287		
288	0.137		
289	0.000		
290	0.319		
291	0.608	0.603	0.44
292	0.538		
293	0.144		
294	0.435		
295	0.349		
296	0.560	0.517	0.28
297	0.137		
298	0.320		
299	0.541		
300	0.600	0.622	0.63
301	0.388		
302	0.471		
303	0.414		
304	0.407		
305	0.465		
306	0.371		
307	0.220		
308	0.370		
309	0.428		
310	0.522		
311	0.321		
312	0.505		
313	0.173		
314	0.358		
315	0.467		
316	0.260		
317	0.315		
318	0.146		
319	0.264		
320	-0.212		

## Item Analysis of the ICE-MH 29

321	0.564	0.498	0.75
322	0.047		
323	0.317		
324	-0.230		
325	0.188		
326	0.159		
327	0.111		
328	0.393		
329	0.119		
330	0.212		
331	0.346		
332	0.601	0.535	0.34
333	0.394		
334	0.486		
335	0.435		
336	0.430		
337	0.459		
338	0.000		
339	0.093		
340	0.319		
341	0.170		
342	0.436		
343	0.078		
344	0.107		
345	0.082		
346	0.371		
347	0.299		
348	0.352		
349	0.345		
350	0.450		
351	0.361		
352	0.076		
353	0.707	0.728	0.75
354	0.411		
355	-0.051		
356	0.482		
357	0.525		
358	0.295		
359	0.386		

## Appendix B

### Results of Item Pruning in the Concepts Subtest

## Appendix B

## Results of Item Pruning in the Concepts Subtest

Item #	Discrimination Analysis	2nd Discrimination Analysis	3rd Discrimination Analysis	Item Difficulty
1	0.309	0.283	0.311	0.77
2	0.045			
3	0.262	0.320	0.353	0.35
4	0.275	0.280	0.251	0.90
5	0.394	0.359	0.382	0.42
6	0.292	0.279	0.308	0.81
7	0.552	0.533	0.509	0.35
8	0.284	0.288	0.305	0.39
9	0.159			
10	0.156			
11	0.241	0.187		
12	0.350	0.349	0.321	0.65
13	0.461	0.490	0.485	0.74
14	0.538	0.553	0.519	0.77
15	0.479	0.532	0.563	0.81
16	0.259	0.254	0.233	0.77
17	0.383	0.381	0.354	0.71
18	0.136			
19	-0.098			
20	0.145			
21	0.192			
22	0.036			
23	0.400	0.446	0.439	0.74
24	0.274	0.245	0.197	0.52
25	0.163			
26	0.273	0.198		
27	0.218	0.173		
28	0.222	0.220	0.262	0.52
29	0.442	0.481	0.518	0.61
30	0.405	0.372	0.396	0.42
31	0.387	0.435	0.453	0.16
32	0.118			
33	0.550	0.553	0.569	0.77
34	0.107			

Item Analysis of the ICE-MH 32

35	0.427	0.423	0.410	0.42
36	0.150			
37	0.353	0.304	0.292	0.32
38	0.547	0.523	0.464	0.58
39	0.475	0.468	0.476	0.61
40	0.224	0.243	0.279	0.68

**Appendix C**

**Results of Item Pruning in the Ethics Subtest**

## Appendix C

## Results of Item Pruning in the Ethics Subtest

Item #	Discrimination Analysis	2nd Discrimination Analysis	Item Difficulty
1	0.486	0.494	0.74
2	0.152		
3	0.738	0.749	0.68
4	0.188		
5	0.538	0.510	0.77
6	0.272	0.309	0.77
7	0.304	0.291	0.61
8	0.455	0.437	0.45
9	0.738	0.756	0.74
10	0.316	0.328	0.48
11	0.335	0.300	0.45
12	0.220		
13	0.242	0.262	0.29
14	0.820	0.837	0.71
15	0.747	0.744	0.77
16	0.525	0.512	0.81
17	0.356	0.381	0.81
18	0.199		
19	0.557	0.528	0.68
20	0.825	0.821	0.84
21	0.812	0.800	0.71
22	0.154		
23	-0.032		
24	0.231	0.215	0.13
25	0.449	0.480	0.55
26	0.377	0.385	0.71
27	0.281	0.275	0.90
28	0.572	0.568	0.74
29	0.314	0.326	0.52
30	0.663	0.669	0.68
31	0.393	0.391	0.81
32	0.364	0.340	0.45
33	0.563	0.574	0.81
34	0.854	0.856	0.68





Appendix D

Examples of Well-Worded and Poorly-Worded Items from the ICE-MH

Examples of well-worded and poorly-worded items from the ICE-MH

A well written item is concise with the stem grammatically matching the choices. The stem should have all the necessary information so that it need not be repeated in the choices.

Answers such as “all of the above” or “none of the above should be used sparingly as it adds a compounding factor to the question because it does not allow the examinee to select the item which is truest (Nunnally, 1994)

Poorly written item from ICE-MH:

Which of the following is a typical effect of caffeine on the body?

- A. Restlessness
- B. Psychomotor agitation
- C. Stimulates nervous system
- D. All of the above
- E. A & B

Well written item from ICE-MH:

People who have depression:

- A. Do not like doctors
- B. Are unable to express what they want
- C. Cannot breath due to compression of lungs
- D. May be socially isolated
- E. Do not believe in God