An Evaluation of Substantive and Statistical Methods in Cross-Lingual Test Item Adaptation: The Case of a High Stakes Test in Kyrgyzstan

A Report for American Councils for International Education/Research Scholars Program

Todd W. Drummond
PhD. Candidate
Michigan State University

Introduction

This report provides highlights of the accomplishments of a research program carried out in the summer of 2010 in Bishkek, Kyrgyzstan. The research was carried out with financial support from the *American Councils for International Education's* Research Scholars Program. After a brief overview of the central research questions, a detailed account of the data collection tasks completed this summer follows. Then, a bibliography along with data collection protocols are provided as appendices. At the time of this update, the investigator is completing the overall data analysis.

The Research

Since the collapse of the Soviet Union many nations throughout Eurasia have dramatically altered their higher education admissions systems. The primary rationale for change has been to overcome corrupt practices that have plagued admissions in the post-Soviet era (Blau, 2004; Osipian, 2007; Heyneman et al., 2008). In the case of Kyrgyzstan for example, change has entailed the replacement of traditional oral admissions

examinations with written, standardized tests.[1] The new admissions tests in Kyrgyzstan are conducted in the Kyrgyz, Russian, and Uzbek languages (Drummond & De Young, 2004).

The introduction of standardized admissions testing to the Eurasian region merits scholarly attention for several reasons. First, admissions policies have high stakes distributive effects as they shape the allocation of the valuable resource of higher education. Second, selection testing for tertiary admissions can impact secondary education as administrators, teachers, and students adjust to new incentives created by what is assess on high stakes tests (Yeh, 2005). Finally, selection inferences in the Eurasian countries are now based on results of cross-lingual tests, the validity of which must be carefully substantiated (Hambleton, 2005).

Cross-lingual assessments are instruments created in one language and adapted into another for use in a different linguistic population (Dorans & Holland, 1993). Practitioners and researchers use cross-lingual assessments for various descriptive, analytical and selection purposes both in comparative studies across nations and within countries marked by linguistic diversity (Hambleton, 2005). The main challenge to valid inferences based on cross-lingual assessments is that in addition to obvious language differences, there may be less visible cultural, contextual, and psychological differences between the tested groups that impact test item response patterns.

Funds provided by the *Research Scholars Program* supported the examination of two key aspects of cross-lingual test adaptation in Kyrgyzstan at the test item level. The first is the extent to which bi-lingual item evaluators are able to predict differential

[1] Kyrgyzstan (2002), Kazakhstan (2004), Georgia (2005), Russia (2008), and Ukraine (2008). Kazakhstan, Russia and Ukraine actually initiated pilot testing of their new admissions tests earlier – Kazakhstan (1999), Russia (2001), Ukraine (2004). These later dates refer to the time when admissions testing became mandatory for all those seeking higher education.

performance (by language group) on a series of adapted tests items. The results of this analysis have implications for the extent to which cross-lingual assessments can provide a sound basis for valid selection inferences. The second is an exploration of the specific linguistic, cultural or other challenges to cross-lingual test item adaptation in the Kyrgyz and Russian languages.

Research Questions

With support from the American Councils for International Education Research Scholars Program, in the summer of 2010 (June – August) considerable fieldwork was conducted to generate data in order to answer the following questions. The primary research question in this study is:

To what extent can item evaluators (bi-lingual) predict how differences (if any) between Kyrgyz and Russian test items on a cross-lingual test will impact item results across these two language groups? Differences here are defined to be linguistic, psychological, cultural, content, format or other differences on each item.

To answer this question, the investigator organized and conducted a substantive review of forty verbal skills test items from the university admissions test in Kyrgyzstan.[2] Ten bi-lingual evaluators were selected and trained to complete this process. The items evaluated consisted of twenty analogy items, ten sentence completion items, and ten reading comprehension items adapted from Russian into Kyrgyz. In brief, item evaluators: (1) estimated levels of difference(s), (2) characterized the nature of the differences, (3) described the differences, (4) estimated which group was favored, (5) suggested improvements to make the items more equivalent, and (6) participated in a focus group discussion about each item.

[2] The investigator speaks both Russian and Kyrgyz and has secured permission to utilize data from the Center for Educational Assessment and Teaching Methods (CEATM), the organization that conducts the university admissions test (see attached). The full name of the university admissions test is The National Scholarship Test, or NST.

Then, a statistical DIF analysis using logistic regression was carried out and the relationship between these two estimation approaches assessed. A secondary question was whether or not the use of statistical methods adds value to the process of adapting cross-lingual test items in Kyrgyzstan. A questionnaire was administered to the representatives of the testing center where the tests were constructed after all the data was analyzed.

The second research question of this study is:

What are some of the specific linguistic, cultural or other challenges to creating equivalent Kyrgyz and Russian test items in Kyrgyzstan?

There are relatively few studies that seek to identify the causes or sources of DIF on cross-lingual assessments (Allalouf, Hambleton & Sireci, 1999; Ercikan et al., 2004). As no DIF studies have been carried out between Altaic (Turkic) and Russian language groups, a primary goal of this study was to contribute to our understanding of the particular challenges to test adaptation between these two languages. Characterizing and categorizing the sources of DIF will inform planning and design of future assessments as well as promote theoretical understanding of DIF involving these two language groups (Allalouf et al. 2005). Data for interpretation will come from the item evaluators' description of the items on the evaluation rubrics as well as group discussion.

Work Accomplished in Summer, 2010

Developing the Item Analysis Rubrics

In order to assess how well evaluators could predict group performance on cross-lingual test items, special item analysis rubrics were created. Rubric one was a test booklet specifically designed to present both versions of a single verbal skills test item, one item per page. As the construction of this booklet required access to the test items,

this booklet was put together in Bishkek after the investigator arrived in country. The investigator engaged with the director of the Center for Educational Assessment and Teaching Methods to select from which testing year data would come from; it was agreed to use data from year 2010.

Other rubrics were developed in the spring of 2010 in East Lansing, MI. They were designed based on the experience gleaned from similar studies and the knowledge gained from the literature review. Once the rubrics were created, a glossary of technical terms also had to be produced which would carefully define all key terms and provide instructions for the participants. Once completed, the rubrics and glossary were translated into Russian and later approved by the Institutional Review Board at Michigan State University. The English version of all these materials can be found in appendices.

Selecting the Evaluators

Ten bi-lingual educators were selected to participate as item evaluators. All participants signed consent forms and were compensated for their work. Selection of competent bi-lingual evaluators was essential. It was necessary to include not only linguists and translators in the review process but also teachers. This is because the item review process requires not only the identification of differences in the two language versions of each item, but also a judgment as to whether these differences might lead to performance differences (Mazor, 1993; Ercikan et al., 2004).

Perhaps the biggest challenge in selecting the evaluators was to ensure that all participants were as close to being as purely bi-lingual as possible. While finding bi-linguals was not difficult in Kyrgystan, pure bi-lingualism is rare and bi-linguals are usually stronger in one language than the other. In Kyrgyzstan, it is primarily ethnic Kyrgyz who are bi-lingual as Russians tend not to speak other languages. However, there

is a wide spectrum of skills and knowledge amongst those that claim to be bi-lingual. The graph below demonstrates the range of knowledge from the general population.

A pool of potential item evaluators was identified with the help of test center employees and provided with information about the study and a consent form. If they agreed to participate, they completed a questionnaire which elicited detailed information about their language skills and educational backgrounds. In order to encourage only pure bi-linguals to apply, the investigator informed the participants would be required to use both Russian and Kyrgyz not only on individual analysis but in oral presentations with their peers – many of whom would be translators, linguists and other knowledgable specialists. As part of this investigation, evaluators would be required to state and perhaps defend their views and opinions on the test items under study. Several potential candidates initially applied but declined after they learned that they would need to discuss item equivalence with a large group of their peers. Each candidate provided information about their professional backgrounds and language capabilities. The investigator selected ten evaluators that provided a balance in terms of competency levels in both languages. In other words, because few specialists are equally strong in two languages the committee

was balanced to ensure that if half the group was stronger in Russian the other half was stronger in Kyrgyz.

The chart below presents the characteristics of those eventually selected to serve as evaluators.

Profession(s):	Teacher (secondary or tertiary) (5), Test item writer (3), Philologist/language specialist (6), Methodologist (1), Translator (5), Linguist/editor (2), Lawyer (1)		
Background Characteristics of Evaluators	**Kyrgyz**	**Russian**	**Both**
Language Medium of Secondary Education	☑☑☑☑☑	☑☑☑☑☑	
Language Medium of Higher Education	☑☑	☑☑☑☑☑	☑☑☑
Main Language at Work	☑	☑☑	☑☑☑☑☑☑☑
Main Language at Home	☑☑☑☑		☑☑☑☑☑☑
Language in which you "think"	☑☑	☑☑☑☑	☑☑☑☑

Slightly more literate in Russian	**Slightly More literate in Kyrgyz**	**Equally Literate in both Languages**
☑☑☑☑	☑☑☑	☑☑☑

All evaluators except one were women and all had completed higher education. The fact that the majority were women is due to the fact that they are over-represented in the education sector in teaching, translation, and linguistics in general in the republic. The majority of participants selected more than one profession. This is because in Kyrgyzstan bi-lingual educators often serve in many capacities: as translators, test item writers, or work in other capacities on a short term basis in addition to their primary place of work. Teachers often work in multiple institutions in varying roles and capacities (De Young, 2004).

In an important sense, the investigator considered the broad spectrum of professional experience a plus. Translators who know the school program or educators who have experience creating test items can approach the task from a multitude of perspectives and have knowledge not only about one area (linguistics for example) but

also practical experience in another, relevant discipline. Three of the ten had experience working as test item writers and one of the ten had actually participated in the adaptation this instrument in 2010. None of them had ever participated in a formal DIF evaluation procedure before.

All participants were ethnic Kyrgyz. In terms of schooling, half the evaluators completed their secondary education in the Russian language medium and half in the Kyrgyz language. Three evaluators received higher education in both languages while only two completed their higher educations in the Kyrgyz language medium. This is not surprising as noted in the previous chapters Russian language medium of instruction was the language of cultural and social capital throughout the Soviet period and even today a majority of those receiving higher education are doing so in the Russian language.

However, seven evaluators use both languages at work and six of them use both languages in the home. None of the evaluators reported that Russian was their main home language. Interestingly however, four evaluators reported that they "think" primarily in the Russian language. If there is error in evaluator reporting, it could very well be in under reporting the amount of Russian spoken in the home. I.e. it is possible that ethnic Kyrgyz respondents tended to note "both languages" were spoken at home, even if Russian language was in fact the primary language. Four marked that they were slightly more literate in Russian than Kyrgyz, three marked that they were slightly more literate in Kyrgyz than Russian and four marked that they were equally literate in both languages.

Testing the Rubrics

I conducted a practice analysis (pre-test) with one evaluator in order to determine if adjustments were needed to the rubric or glossary. A separate time was set up and the

one evaluator completed the rubrics several days before the other nine evaluators met to work. The pre-test yielded important results. In addition to the discovery of some minor formatting mistakes, the pre-test evaluator reported in a debriefing that the most challenging aspect of the rubric was interpreting the coding categories in section two of rubric 2.

Although definitions of “adaptation, translation, format and cultural issues” were provided in the pre-packaged materials given to the evaluators, due to the limitations of time, and perhaps evaluator experience, the pre-test evaluator noted that these categories were easily confused and open to various interpretations. She noted, for example, that she spent an inordinate amount of time attempting to classify whether a problem with an item was a “cultural” or “linguistic” problem. She questioned the utility of coding the nature of the problem and was in favor of more focus on description of the problem (section three).

After considerable contemplation, the researcher agreed that the main purpose of the rubrics was to gather good descriptive data about the item, not assess how consistently the evaluators coded the nature of the problem: This is the task of the researcher, to characterize and interpret what kinds of problems were being discovered, after collecting the data from all ten evaluators. However, as the full rubrics had already been printed, this section was left on the rubrics. It was decided that before evaluators started to fill out the rubrics they would be instructed to focus on sections one, three, four and five of the rubrics. Emphasis was placed on section three (description) of the issues they discovered with each item.

Several days before convening the evaluators for item analysis, each evaluator was provided with the glossary of key terms for home study. They were instructed to read through the glossary carefully before the committee was to convene. The evaluator panel was convened at 98 Tynustanova Street at the Center for Educational Assessment and Teaching Methods at 9:00 am. All ten evaluators came on time, the researcher conducted a forty-five minute training, and evaluators were seated in individual work stations.

Administering the Rubrics

There were three main steps to the evaluation process. On the morning of day one all evaluators answered the thirty eight[3] test items test items (rubric 1 a) and provided an initial mark on difference levels (rubric 1b). This process took approximately three and one half hours. The next day evaluators completed rubric 2 for the items they had marked as not identical. This process took approximately four hours. On the evening of day three, all evaluators returned to the test center to discuss each item. This process took approximately three hours. The English version of these evaluation rubrics can be found in appendix three.

Day One

Rubrics 1a and 1b were designed based on a model by utilized by Allalouf et al. (1999). The evaluators attempted to correctly answer all the items in both the Kyrgyz and Russian versions in a single test booklet. Evaluators were split into two groups and asked to start with different items. One group started with item 1 while the other started with item 20. This ensured that all items received at least a minimum amount of coverage. Items were organized one pair to a page with room for comments about equivalence and differences.

[3]

This was a blind review in the sense that the evaluators did not know which items had been identified as DIF (Ercikan et al. 2004). Evaluators read and answered both language versions and responded to the items as if they were test takers; they took notes only on the most important problems that arose. Item pairs coded as "identical" on rubric 1.b were set outside as they would not be needed for the completion of rubric 2. Evaluators were tasked with filling out rubric 2 in detail only for those pairs for which differences were evident. A fifteen minute coffee break was organized after the second hour. All test booklets were collected at the end of the day and stored in a secure location until the continuation of work the next day.

Day Two

Rubric 2 was filled in the next day. This step required the evaluators to take their notes from step one and code their comments on specially developed rubrics for those items not selected as identical. Per the changes following the pre-test, rubric 2 now contained the following sections: (1) estimation of the level of differences (if any), (3) description of differences in detail, (4) estimation of which group (if any) might benefit from the noted differences, (5) suggestions for improvements.

Section 1 (level of difference) of the rubric required evaluators to classify each pair of items as identical, somewhat similar, somewhat different, or different. This coding scheme, adapted from both Ercikan et al.'s (2004) and Reckase and Kunce's (2002) work, defined these terms as follows:

0- Identical: no difference in meaning between two versions;
1- Somewhat similar: small differences in meaning between two versions, will not likely lead to differences in performance;
2- Somewhat different: clear differences in meaning between the two versions may or may not lead to differences in performance between two groups;

3- Different: differences in meaning between two versions that are expected lead to differences in performance between two groups.

Depending on the nature of the differences, evaluators filled out color-coded rubrics based on whether their comments are related to content (violet form), format (green form), or cultural/linguistic issues (pink form). This allowed the researcher to more easily collate the forms later during analysis. This stage of the process took approximately four hours to complete.

At the end of the day, booklets and rubrics were collected by the investigator and analyzed that evening to look for key patterns and issues. This was done because the time to be allowed on day three for discussion was only 2-3 hours and the investigator wanted to make sure that items were prioritized for discussion. This initial review by the investigator primarily considered to what extent evaluators estimated "the level of differences" to be. That is, if certain items elicited much commentary, discussion or varying views, it was essential that the group discussed these issues as a group.

Day Three

The discussion session was held on a Monday evening after the normal workday for most evaluators. The evening session with entire group lasted three hours. The investigator facilitated the discussion in the Russian language. A note taker from the test center recorded the conversation and it was also audio-recorded by the investigator. Areas of agreement and disagreement were noted and recorded. The investigator allowed the conversation to flow but on occasion required evaluators to support their opinions and kept the conversation on task. Data from these discussions (along with data from the item analysis rubrics) was recorded for later analysis. This data would be utilized to examine

the relationship between evaluators' marks and the DIF statistics as well as disentangle the many potential sources of DIF on the test items.

Examples of individually completed rubrics are also presented as examples in the appendices. The complete item summary rubrics with fully coded data can be found in the appendices. There is one summary rubric per item which contains the marks and evaluative comments from each of the ten evaluators for that test item. The statements and marks are unedited, reported as written on each of the individual rubrics that they completed in step one and two of the analysis. For example, for item two, the researcher received ten different individual rubrics. The researcher translated the data from Russian and Kyrgyz into English from each of the rubrics and collated all comments onto one summative protocol representing the entire spectrum of marks for that item.

Current Status of the Research: Analyzing the Results

At the present time, the researcher is analyzing the rubrics and the results of the statistical analyses. The first step of this analysis is the transcription and translation from Russian and Kyrgyz into English. Recall that these rubrics contain estimations by evaluators of levels of differences, nature of differences, descriptions of differences, estimations of which group is favored, suggestions for improvement as well as transcripts from item discussion.

The data from each of the ten evaluators will be summarized onto one rubric per item. For example, all descriptions about item number 2 will be collated together to present the overall evaluation of the item. The discussion transcript for each item will also be written up onto the same rubric. The result should be a summative collection of

all the data for each item in one location. Depending on the activeness of the evaluators and the length of discussion, there might be 3-4 pages of data for each item.

Data about the level of difference will be coded and compared to the statistical data. I will use Spearman's rank order correlation to see how the evaluators' predictions matched statistical measurement on each item. Then conclusions will be drawn as to the extent of overlap of these two methods.

Descriptive data and discussion notes will provide information for research question number two. I will be able to elicit what specific challenges to adapting items in the Russian and Kyrgyz languages are occurring and see what grammar, syntax, or other linguistic or cultural issues arose in the adaptation process. This data will be presented in an item by item analysis drawing on the direct transcripts and coding of the evaluators. The researcher will codify and summarize these main findings.

Important questions that I hope to answer through the analysis of this data are what proportion of the detected DIF can be attributed to resolvable adaptation errors vs. other identifiable or unidentifiable lack of measurement invariance between the two groups? For example, are differences in meaning on item pairs due to mistakes in item translation or due to inherent linguistic differences in the way a particular language expresses or represents certain meanings or constructs? And finally, in regard to this particular test, is the amount of DIF detected pervasive enough to challenge the inferences for which this selection test was designed?

This study presents an opportunity to advance our specific understanding of the linguistic and methodological challenges in adapting cross-lingual verbal items on a high stakes test. In particular, the study will: (1) provide empirical evidence as to the utility of

employing statistical DIF detection methods in a context where substantive item reviews are usually employed. The use of these two methods congruently will also provide an estimate the amount and sources of DIF on this high stakes selection test; (2) generate and disseminate new knowledge about the specific linguistic and cultural challenges to cross-lingual comparative assessments involving Slavic and Altaic language groups;

Results of the study will assist test developers and policy makers in Eurasian countries decide how to best approach the complex task of cross-lingual test adaptation by providing them information about the efficacy of various approaches. Depending on the level and nature of DIF discovered, the results of the study might also assist policy makers in the Eurasian region decide whether or not cross-lingual assessments alone should be utilized as the single selection criteria for high stakes test like university admissions.

Bibliography

Ackerman, T. (1992). "A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective." *Journal of Educational Measurement*, 29, no.1, pp. 67-91.

Agnoff, W.H. & Cook, L.L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Examination Board.

Allalouf, A., Hambleton, R., & Sireci, S. (1999). "Identifying the causes of DIF in translated verbal items." *Journal of Educational Measurement*. Vol. 36, No. 3. pp. 185-198.

Archer, M. (1979). *Social origins of educational systems*. London: SAGE Publications.

Blau, A. (2004). "Central Asia: Buying Ignorance – Kyrgyz, Kazakhs Lead in Education Reform (Part 4)." *Radio Free Europe/Radio Liberty, July 7, 2004*.

Bejar, I., Chaffin, R., & Embertson, S. (1991). *Cognitive and psychometric analyses of analogical problem solving*. New York: Springer-Verlag.

Beller, M. (1995). "Translated Versions of Israel's Inter-University Psychometric Entrance Test (PET)." In T. Oakland & R.K. Hambleton (Eds.), *International perspectives of academic assessment* (pp. 207-217). Boston, MA: Kluwer Academic Publishers.

Beller, M., Gafni, N., Hanani, P. (2005). "Constructing, adapting, and validating admissions tests in multiple languages: the Israeli case," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Hambleton, R., Merenda, P. & Spielberger, C. (Eds), Mahwah, NJ & London: Lawrence Erlbaum Associates.

Bereday, G. (1960). *The Changing Soviet School*. The Riverside Press: Cambridge, MA.

Berk, R.A. (Ed.). (1982). *Handbook of methods for detecting test bias*. Baltimore MD: Johns Hopkins University Press.

Birbaum, M., & Tatsuoka, K.K. (1982). "On the dimensionality of achievement test data." *Journal of Educational Measurement*, 19, pp. 259-266.

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press.

Brunner, J. & Tillet, A. (Eds.). (2007). "Higher education in central asia: The challenges of modernization (case studies from Kazakhstan, Tajikistan, the Kyrgyz Republic and Uzbekistan)." The International Bank for Reconstruction and Development/TheWorld Bank.

Camilli, G. & Shephard, L. (1994). *Methods for Identifying Biased Test Items*. Sage Publications: London, England.

CEATM, Center for Educational Assessment and Teaching Methods. Official Website: http:www.testing.kg

Child, D. (1990). *The Essentials of Factor Analysis.* 2nd Edition: Cassell: London.

Clark, N. (2005). "Education reform in the former soviet union." *World Education News and Reviews.* WES, Dec.2005.http://www.wes.org/ewenr/PF/05dec/pffeature.htm

Clauser, B.E., Mazor, K.M., & Hambleton, R.K. (1991). "The influence of the criterion variable on the identification of differentially functioning items using the Mantel-Haenszel statistic." *Applied Psychological Measurement,* 15 (4), pp. 353-359.

Clauser, B.E., Mazor, K.M., & Hambleton, R.K., (1993). "The effects of purification of the matching criterion on identification of DIF using the Mantel-Haenszel procedure." *Applied Measurement in Education*, 6, 269-279.

Clauser, B.E., Nugester, R.J., & Swaminathan, H. (1996). "Improving the matching for DIF analysis by conditioning on both test score and an educational background variable." *Journal of Educational Measurement*, 33, 453-464.

Clauser, B.E., Nungester, R.J., Mazor, K., & Ripkey, D. (1996). "A Comparison of Alternative Matching Strategies for DIF detection in Tests that are Multidimensional." *Journal of Educational Measurement*, 33, pp. 202-214.

Clauser, B. & Mazor, K. (1998). "Using statistical procedures to identify differentially functioning test items." *Educational Measurement: Issues and Practice*. Spring 1998. pp. 31 – 44.

Cohen, J. (1992). "A power primer." *Psychological Bulletin*, 112, 155-159.

De Young, A., Reeves, M., & Valyeva, G. (2006). *Surviving the Transition? Case Studies and Schooling in the Kyrgyz Republic Since Independence*. Greenwich, Connecticut: Information Age Publishing:

Dienes, L. (1987). Soviet Asia: Economic Development and National Policy Choices. Westview Press: Boulder and London.

Dorans, N.J., & Holland, P.W. (1993). "DIF detection and description: Mantel-Haenszel and standardization." In Holland, P.W. and Wainer, H. (eds), *Differential Item Functioning.* Hillside, NJ: Lawrence Earlbaum, 35-66.

Douglas, J.A., Roussos, L.A., & Stout, W. (1996). "Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning." *Journal of Educational Measurement*, 33, No. 4, pp. 465-484.

Drummond, T. & Titov, C. (2004). "Analiz Obshe Respublikanskova Testirovaniya 2003 goda v Kirgizskoi Respublikii: Perviyee Vzglad," (Tamga Digital, Bishkek), with support from American Councils for International Education and USAID.

Drummond, T. & De Young, A. (2004). "Perspectives and problems in education reform in kyrgyzstan: The case of national scholarship testing 2002," in *Challenges for Education in Central Asia*, S. Heyneman and A. De Young (Eds.), Greenwich, CT: Information Age Publishing, pp. 225-242.

Duncan, A. (June 14, 2009). "States Will Lead the Way Towards Reform," Address by the Secretary of Education at the 2009 Governors Education Symposium. www.ed.gov

Education Development Strategy for the Kyrgyz Republic: 2007-10, (2006). Ministry of Education, Science, and Youth Policy, The Kyrgyz Republic.

Ellis, B.B. (1995). "A partial test of Hulin's psychometric theory of measurement equivalence in translated tests." *European Journal of Psychological Assessment,* 11, 184-193.

Englehard, G. David, M., Hanshe, L.(1999). "Evaluating the Accuracy of Judgments Obtained from Item Review Committees." *Applied Measurement in Education*, 12(2), pp.199-210

Ercikan, K., & McCreith, T. (2002). Effects of adaptations on comparability of test items and test scores. In D. Robitaille & A. Beaton (Eds.), *Secondary analysis of the TIMSS results: A synthesis of current research* (pp. 391-407). Dordrecht, the Netherlands: Kluwer.

Ercikan, K., Gierl, M., McCreith, T., Phan, G., & Koh, K. (2004). "Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of the Canada's national achievement tests." *Applied Measurement in Education*, 17 (3), pp. 301-321.

Ercikan, K. & Koh, K. (2005). "Examining the construct comparability of the English and French versions of TIMSS." *International Journal of Testing.* Vol. 5, No. 1, pp. 23-35.

Furr, M., & Bacharach, Y. (2008). *Psychometrics: An Introduction.* Sage Publications: Los Angeles.

Gafni, N., & Canaan-Yehoshafat, Z. (1993, October). *An examination of differential item functioning for Hebrew and Russian-speaking examinees in Israel*. Paper presented at the Conference of the Israeli Psychological Association, Ramat-Gan.

Gierl, M., Rogers, W.T., & Klinger, D. (1999). "Using statistical and judgmental reviews to identify and interpret translation DIF." *Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME)* at the Symposium entitled, "Translation DIF: Advances and Applications." Montreal, Canada. April 20-22, 1999.

Gierl, M.J. & Khaliq, S.N. (2001). "Identifying sources of differential item and bundle functioning on translated achievement tests." *Journal of Educational Measurement,* 38, pp. 164-187.

Glenn, C. (1995). *Educational Freedom In Eastern Europe*. Cato Institute: Washington, DC.

Grisay, A. de Jong, J.H.L., Gebhardt, E., Berezner, A., & Halleux, B. (2006). *Translation equivalence across PISA countries.* Paper presented at the 5th Conference of the International Test Commission, Brussels, Belgium, 6-8, July 2006.

Grisay, A. & Monseur, C. (2007). "Measuring the equivalence of item difficulty in the various versions of an international test." *Studies in Educational Evaluation*: 33, p. 69-86.

Hambleton, R.K., & Rogers, H.J. (1989). "Detecting potentially biased test items: Comparison of IRT and Mantel-Haenszel Methods." *Applied Measurement in Education*, 2(4), pp/313-334.

Hambleton, R.K., Clauser, B.E., Mazor, K.M., Jones, R.W. (1993). Advances in the detection of differentially functioning test item. *European Journal of Psychological Assessment, 9, 1-18.*

Hambleton, R. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Hambleton, R., Merenda, P. & Spielberger, C. (Eds), Mahwah, NJ & London: Lawrence Erlbaum Associates.

Heyneman, S., Anderson, K. & Nuralieva, N. (2008). "The cost of corruption in higher education." *Comparative Education Review.* 52(1), 1-25.

Holland, P. & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Earlbaum Associates.

International Crisis Group, (2003). *Youth in Central Asia: Losing the New Generation*. Asia Report No. 66. Osh/Brussels.

Jodoin, M.G. & Gierl, M.J. (2001). "Evaluating Type I Error and Power Using an Effect Size Measure with the Logistic Regression Procedure for DIF Detection." *Applied Measurement in Education*, 14(4), pp. 329-349.

Johnson, M. (2004). The legacy of Russian and Soviet Education, in *Challenges for Education in Central Asia*, S. Heyneman and A. De Young (eds), Greenwich, CT: Information Age Publishing. pp. 21-36.

Joldersma, K. (2008). *Comparability of Multi-Lingual Assessments: An Extension of Meta-Analytic Methodology to Instrument Validation*. PhD. Dissertation, Michigan State University.

Kok, F. (1988). "Item bias and test multidimensionality." In R. Langeheine & J. Rost (Eds.), *Latent Trait and Latent Class Models.* (pp. 263-274. New York: Plenum.

Kutueva, A. (2008, May 22). *Po dannym antikorruptsionnogo komiteta za 2007 god, Ministerstvo obrazovaniya i nauki Kyrgyzstana stoit na vtorom meste po urovnyu korruptsii.* [According to data from the anti-corruption committee in 2007, the ministry of education and science is in second place for highest levels of corruption]. Retrieved from the website of Information Agency 24.KG: http://www.24.kg/community/2008/05/22/85241.html

Mazor, K. (1993). *An Investigation of the Effects of Conditioning on Two Ability Estimates in DIF Analyses when the Data are Two-Dimensional*. PhD. Dissertation. University of Massachusetts.

Mazor, K.M., Kanjee, A., Clauser, B.E. (1995). "Using logistics regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning." *Journal of Educational Measurement, 32,* pp. 131-144.

Mazor, K., Hambelton, R.K., & Clauser, B.E. (1998). "The effects of matching on unidimensional subtest scores." *Applied Psychological Measurement,* 22, pp. 357-367.

Mellenbergh, G.J. (1982). "Contingency table models of assessing item bias." *Journal of Educational Statistics*, 7, pp. 105-118.

Messick, S. (1988). "The once and future issues of validity: Assessing the meaning and consequences of measurement." In H. Wainer & H.I. Braun (Eds.), *Test Validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Ministry of Education Statistics for 2001.www.moik.gov.kg

National Governors Association (NGA), Council of Chief State School Officers, & Achieve (2008). *Benchmarking for success: Ensuring U.S. students receive a world-class education.* http://www.achieve.org/BenchmarkingforSuccess

Narayanan, P. & Swaminathan, H. (1994). "Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning." *Applied Psychological Measurement*, 18, pp. 315-338.

Narayanan, P., & Swaminathan, H. (1996). "Identification of items that show non-uniform DIF." *Applied Psychological Measurement*, 20, no.3, pp. 257-274.

Osipian, A. (2007, February*). Corruption in higher education: Conceptual approaches and measurement techniques***.** Paper presented at the meeting of the Comparative and International Education Society (CIES), Baltimore, MD.

Plake, B.S. (1980). "A comparison of statistical and subjective procedures to ascertain validity: one step in the test validation process," *Educational and Psychological Measurement*, 40, pp. 397- 404.

Poortinga, Y.H. (1983). "Psychometric Approaches to Intergroup Comparison: The Problem of Equivalence." In S.H. Irvine and J.W. Berrey (Eds.), Human Assessment and Cross-Cultural Factors? New York: Plenum Press. pp. 237-258.

Poortinga, Y.H. (1989). "Equivalence of Cross-Cultural Data: An Overview of Basic Issues." *International Journal of Psychology,* 24, pp. 737-756.

Presidential Decree No. 91. (2002, April 18). *"O dal'neyshih merah po obespecheniyu kachestva obrazovaniya i sovershenstvovaniyu upravleniya obrazovatel'nymi protsessami v Kyrgyzstkoy Respublike."*. [About further measures for ensuring quality education and improving the administration of educational processes in the Kyrgyz Republic].

Reckase, M.D. (1985). "The difficulty of items that measure more than one ability." *Applied Psychological Measurement.* 9: 401.

Reckase, M.D., & Kunce, C. (2002). "Translation accuracy of a technical credentialing examination," in *International Journal of Continuing Engineering Education and Lifelong Learning*, Vol. 12, Nos. 1-4, pp. 167-180.

RIA News, Moscow. (2007, February 5). *Vvedenie v Rossii Edinogo gosudarstvennogo ekzamena (EGE) yavlyaetsya oshibkoy, ubezhden spiker Soveta Federatsii Sergey Mironov.* [Speaker of federal Soviet thinks the USE is a systematic mistake]. Retrieved from http://www.spravedlivo.ru/news/section_385/738.smx

Robin, F., Sireci, S., & Hambleton, R. (2003). "Evaluating the equivalence of different language versions of a credentialing exam," *International Journal of Testing*, 3 (1), (Lawrence Erlbaum Associates, Inc.), pp. 1-20.

Rogers, H.J. (1989). *A Logistic Regression Procedure for Detecting Item Bias.* Unpublished Doctoral Dissertation, University of Massachusetts, Amherst.

Robin, F., Sireci, S., & Hambleton, R. (2003). "Evaluating the equivalence of different language versions of a credentialing exam," *International Journal of Testing*, 3 (1), Lawrence Erlbaum Associates, Inc. pp. 1-20.

Roccas, S., & Moshinsky, A. (1997). *Factors affecting the difficulty of verbal analogies (NITE Report No. 239).* Jerusalem: National Institute for Testing and Evaluation.

Rogers, J. & Swaminathan, H. (1993). "A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning." *Applied Psychological Measurement*. Vol. 17, no.2, pp. 105-116.

Roussos, L., & Stout, W. (1993). "A multidimensionality-based DIF analysis paradigm." *Applied Psychological Measurement*, 20, pp. 355-370.

Sait Halma, T. (1981). *201 Turkish Verbs Fully Conjugated in All the Tenses*. Barron's Educational Series: NY, New York.

Scheneman, J.D. (1982). "A posteriori analyses of biased items." In R.A. Berk (Ed.) *Handbook of Methods for detecting test bias* (pp. 180-198). Baltimore, MD: Johns Hopkins Press.

Schumacker, R. (2005). *Test Bias and Differential Item Functioning*. Applied Measurement Associates.

Schmitt, A.P., & Bleistein, C.A. (1987). *Factors affecting differential item functioning of black examinees on Scholastic Aptitude Test and analogy items* (Research Report 87-23). Princeton, NJ: Educational Testing Service.

Schmitt, A.P. (1988). "Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the scholastic aptitude test." *Journal of Educational Measurement*, Vol. 25, No. 1 (Spring) pp. 1-13

Shealy, R., & Stout, W.F. (1993). "A model-based standardization approach that separates true DIF/Bias from group differences and detects test bias/DTF as well as item bias/DIF." *Psychometrika*. 58, 159-194.

Sireci, S.G., & Allalouf, A. (2003). "Appraising item equivalence across multiple language and cultures." *Language Testing*, 20(2), pp. 148-166.

Sireci, G., Patsula, L., & Hambleton, R. (2005). "Statistical methods for identifying flaws in the test adaptation process," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Hambleton, R., Merenda, P. & Spielberger, C. (Eds), Mahwah, NJ & London: Lawrence Erlbaum Associates.

Subkoviak, M.J., Mack, J.S., Ironson, G.H., & Craig, R.D. (1984). "Empirical comparison of selected item bias procedures with bias manipulation," *Journal of Educational Measurement*, 25, pp. 301-319.

Swaminathan, H. & Rogers, J. (1990). "Detecting differential item functioning using logistic regression procedures." *Journal of Educational Measurement*, vol. 27, No. 4, pp. 361-370.

Tittle. C.K. (1982). "Use of judgmental methods in item bias studies." In R.A. Berk (Ed.) *Handbook of Methods for detecting test bias* (pp. 31-63). Baltimore, MD: Johns Hopkins Press.

Van de Vijver, F. & Tanzer, N.K. , (1999). "Bias and equivalence in cross-cultural assessment: An overview," *European Review of Applied Psychology*, 47, 263-279.

Valkova, I. (2001). My Symphony: Interview with the Minister of Education of Kyrgyz Republic, Camilla Sharshkeeva. In *Thinking Classroom*, October (6). Vilnius, Lithuania: International Reading Association.

Valkova, Inna. 2004. *Getting Ready for the National Scholarship Test: Study Guide for Arbiturents.* Bishkek: CEATM.

Valyaeva, G. (2006, September). *Standardized testing for university admissions in Kazakhstan: A step in the right direction?* Paper presented at the Central Eurasians Studies Conference, University of Michigan, Ann Arbor, MI.

Van de Vijver, F. & Poortinga, Y., (2005). "Conceptual and methodological issues in adapting tests," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, Hambleton, R., Merenda, P. & Spielberger, C. (Eds), Mahwah, NJ & London: Lawrence Erlbaum Associates.

Yeh, S.S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives,* 13(43), 1-23.

Zheng, Y., Gierl, M., & Cui, Ying, C., (2005). "Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and Logistic Regression Procedures." Centre for Research in Applied Measurement and Evaluation, University of Alberta.

Zumbo, B.D., & Thomas, D.R. (1997). *A Measure of Effect Size for a Model-Based Approach for Studying DIF*. Working paper of the Edgeworth Laboratory for Quantitative Behavorial Science, University of Northern British Columbia: Prince, George, B.C.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF): Logistic Regression Modeling as a Unitary Framework for Binary and Likert-Type (Ordinal) Item Scores*. Ottawa, ON: Directorate of Human Resources and Evaluation, Department of National Defense.

Zumbo, B.D., (2003). "Does item level DIF manifest itself in scale level analysis? Implications for translating language tests." *Language Testing*, 20(2) 136-147.

.

Glossary of Rubric Terms

(Provided to evaluators)

It is generally understood that *different* means *not the same*. Differences of wording, content, format, and structure of cross-lingual test items can lead to differences in overall item meaning and difficulty. If we intend to use items to measure the same constructs and knowledge, differences on two versions of the same item are problematic because accurate measurement (regardless of language group) requires equivalent items: Thus, differences between the two versions of the item can potentially invalidate the inferences based on test results.

In the context of this study, there are four aspects of difference that merit attention – differences in the meaning of individual words, differences in overall meaning, differences in relative difficulty, and differences in cultural interpretation of the two versions of the item. In this study, causes of such differences will also be examined. *Equivalent* means the opposite of different in this context: Equivalence is achieved when the item in both language versions has the same meaning, same relative difficulty level, and can be interpreted similarly in the different language groups (cultural interpretation).

Meaning of Individual Words and Overall Meaning

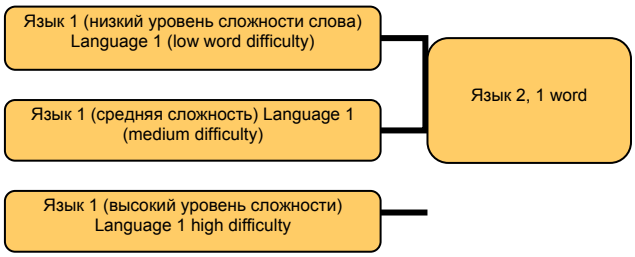
The incorrect translation of individual words, the addition or omission of a word can cause differences in item meaning. This problem can sometimes be resolved relatively easily by improvements in translation. The word ***Translation*** (1.a) will be used in this study in a narrow sense to refer to *direct, one to one correspondence* of words and sentences. In many instances, direct correspondence is needed to make words and ideas expressed by test items equivalent. If a single word is mistranslated, overall meaning can change or the item can make no sense at all.

In many cases, however, two items translated correctly (*word for word*) can result in different overall meaning. For example if literal translation was used when the actual properties of the two languages require a more nuanced adaptation to retain similar meaning. So, the lack of direct correspondence of words is not necessarily always problematic. In recognition of the above, test developers often prefer to speak of test or item ***adaptation*** rather than translation (Hambleton, 2005). Adaptation (1.a) is preferred term because it acknowledges that direct, literal, translation is often not possible (or desired) across disparate languages if we seek to maintain the overall similarity in meaning of two test items. A sentence or text can have little direct, literal correspondence to the same material in another language, yet maintain the same overall meaning. Sometimes this is because there are differences in linguistic structures (lack of a certain verb tense in one language for example) or other inherent differences (see more below). For this rubric, the term adaptation is utilized to denote this broader idea of similar overall meaning, regardless of how individual words may or may not correspond.

Relative Difficulty

Individual words as well as phrases, concepts and ideas can have equivalent overall meaning but still be problematic for test developers. That is because while some words, phrases and ideas can be similar in overall meaning in two groups, they can be different in terms of their conceptual difficulty in the two groups. An obvious example of this is when one language has five synonyms for the same word or idea while the other language has two. In the language that has five words, two of them might be rarely utilized, for example in literary or other scholarly circles. Thus, the *commonality* of their use may be as important as their actual meaning in terms of how differences in item difficulty manifest themselves in different linguistic or cultural groups. While the use of such pairs of words may technically be correct, their usage might pose the problem of relative difficulty for one language group.

Another example of when linguistic adaptation appears correct but remains problematic is the issue of *explicitness* of words or ideas. For example, ideas that are conceptually challenging in one language might get adapted to a more literal or explicit meaning in the second language, making them easier to interpret. Complex metaphors are sometimes adapted to a more literal meaning in the target language which can lead to the target language having greater success on an item.



Cultural/Linguistic Interpretation

Literal translations and even adaptations can appear accurate, but test items may still have different meanings for two language groups. Cultural, contextual, and conceptual understandings may differ between the groups enough to make the intended meaning of some items unclear, irrelevant, or have a disparate difficulty levels for one group. Due to demographic, socio-economic, or linguistic reasons, some words, concepts, or ideas might be more familiar to one group than the other. The way two languages express or articulate ideas and concepts could make meaning more "difficult to locate" in some languages than others.

Terms from the Rubric 1.b.

Type of Difference

No Difference
The two versions of the item are assessing the same thing in the same way, using equivalent words, ideas, and content, as well as a similar format. Similar cultural meaning and equivalent language is attained. You expect no differences in item performance by the two groups on this item.

Content Differences
Refers to the basic ideas, concepts, knowledge, skills, language, and words assessed on each item (see prompts on Rubric 2.I.).

Format Differences
Refers to the way content is formatted, spaced, edited, and presented visually. Size of text, length of material, punctuation, capitalization, etc. (see prompts on Rubric 2.II.)

Cultural/Linguistic Differences
Meaning and of items to both Russian and Kyrgyz examinees, relevance to different schooling contexts and cultures, similarity of dispositions of two groups, similarity of norms, psychological construct present in both groups, equivalence of linguistic expression, similarity of linguistic structure and grammar, symbolism, metaphor meaningful in both groups, level of explicitness similar, etc. (see prompts on Rubric 2.III)

Terms from the Rubric 2.

Level of Difference

Somewhat Similar
You note small differences between the two versions of the item but they are not very significant. The kind of "daily" differences you see are those that an examinee might also be quite familiar with and be able to negotiate with little or no difficulty.

Somewhat Different
These items appear to be different in more obvious and unambiguous ways. However, you are not sure that these differences will impact item response patterns.

Different
These items clearly indicate differences in meaning, relative difficulty or cultural interpretation. You are confident that these differences will impact the way students answer these questions. In other words, you are confident that these differences will impact item response patterns.

Terms from Rubric 2.III. Cultural Meaning

Meaning Differences

Under the meaning differences for the cultural category, we refer not to meaning differences caused by translation mistakes, but meaning differences that might occur even when the translation is accurate. In regard to comparison of Russian and Kyrgyz examinees, consider the word "family." The definition of family is culturally informed and can very different meanings in different cultural groups (wider understanding or more narrow understanding). Other words/concepts such as "independence, freedom, love, values, respect, etc." are all strongly influenced by cultural norms and values. Test items that use such concepts should do so with regard to possible differences between the two groups.

Contextual Differences

Contextual differences can occur when examinees from different groups have different levels of exposure to ideas, knowledge or situations due to demographic, social, or cultural differences. In Kyrgyzstan, Russian speaking examinees are (on average) concentrated in urban areas while Kyrgyz speakers (on average) are concentrated in rural areas. For example, urban Russian examinees might have less knowledge about horsemanship and the vocabulary, knowledge and norms that surround the keeping and use of horses than say Kyrgyz youth who have been raised around horses. Geographic concepts and terminology about mountains might also advantage those who live in such regions. Or, the opposite, many urbanites might be more familiar with questions connected to the ways of life of urban dwellers. Focus on cultural heroes, myths, legends might also be problematic. Like meaning differences, contextual differences might not be apparent in the quality of translation/adaptation (which may be accurate) but must be considered nonetheless. Finally, success on an item should not depend on exposure to similar curriculum and schooling practices.

Linguistic Differences

Language falls under several rubrics. The most obvious form of "linguistic difference" becomes evident when items are poorly translated or adapted. However, there are also inherent differences in the way languages forms express and convey meaning. For example, an adaptation might be accurate but it might take many more words to express a concept in one language than another. How (if at all) does this impact the difficulty of an item? Some languages might have more nuances of meaning due to having more verb tenses which create meaning not easily captured in another language. Some languages might more "efficiently" convey meaning than others in some situations. As bi-lingual speakers, consider the times that you consciously or subconsciously prefer to use one of the languages you know more often than the other because the language allows a more precise or efficient expression of your intended meaning. Some languages might have many more words for richer variation of nuance of certain concepts. Word order can also be important. Consider the example of the item instructions in the Russian and Kyrgyz items below. Are there differences in meaning and/or difficulty of these items inherent in the way the directions are expressed? Is the issue easily resolved?

"каждое задание состоит из пяти пар слов. Выделенная жирным шрифтом пара показывает образец отношении и между двумя словами. Определите,

какие отношения сушествуют между словами в этой паре, а затем выберите в вариантах ответа пара слов с такими же отношениями. Порядок слов в выбранном вами ответе должен быть таким же как и в образце.”

“Ар бир тапшырма беш жуп создон турат. кара тамгалар менен белгиленген жуп соз эки создун ортосундагы мамиленин улгусун корсотуп турат. адегенде бул жуптагы создордун ортосундагы мамелени анектаныз да, андан сонг жооптун вариантгарынын ичинен ушундай мамеледе турган жуп созду тандап алыныз.”

Item Rubric 1.a. How equivalent are these test items?

Directions: Please read and answer both the Kyrgyz and Russian versions of the test item below. In the comments boxes to the right, make a brief note about how well (in your estimation) these two items are assessing the same thing in the same way. You may write notes directly on the items. Consider content, format and cultural issues, as well as linguistic comparability. In the lower box, please comment on the quality of the translation/adaptation. Make only brief notes as you will return for a more in depth analysis of these items later.

Item 1

Kyrgyz Version

Item here…	Notes: Equivalent/ Different

Russian Version

Item here…	Notes: Translation/Adaptation

Item Rubric Summary 1.b

Directions:

Please review the notes you took while taking the test items in 1.a. and circle the descriptor that best characterizes each pair of items. Please circle differences if *any* level of difference is apparent (small, medium, or large):

Item 1:	**0. No differences**	**I. Content Differences**	**II. Format Differences**	**III. Cultural/Linguistic Differences**
Item 2:	**0. No differences**	**I. Content Differences**	**II. Format Differences**	**III. Cultural/Linguistic Differences**
Etc.:	**etc…**			
Item 40:	**0. No differences**	**I. Content Differences**	**II. Format Differences**	**III. Cultural/Linguistic Differences**

Item Rubric 2

Item Rubric 2 is to be filled out *only* for those items you identified as different (somewhat similar, somewhat different or different). The purpose of Item Rubric 2 is to collect data that will facilitate an understanding of the cause (source) of problems of equivalence.

(Note that the term DIF *is not used when non-statistical procedures identify differences in item functioning. When evaluators identify differences between items, we flag them as problematic because they are not equivalent. When speaking about DIF, we are speaking about empirical data patterns observed through statistical means. When we speak about "problems of equivalence" identified by evaluators, we are speaking about their "best estimates" of how items may or may not be different and how students may or may not perform).*

The rubric is broken into three primary categories to characterize the nature of the differences. The main categories are: I. Content differences, II. Format differences, and III. Cultural differences. Note that these categories are not always clearly mutually exclusive.

However, these three categories provide a strong foundation from which to classify core item issues. Evaluators also have an option to note “other reason for difference.”

At the top of the rubric for each category, evaluators are provided a series of prompts – or possible explanations for differences. These prompts are not meant to be exhaustive but are there to help evaluators classify the cause of the problem. In question 1 of each category, evaluators are asked to score the item as “somewhat similar” “somewhat different” or “different.” (These nominal categories are also coded with an ordinal indicator which will later be utilized to calculate inter-rater reliability). Then, evaluators are asked to speculate on the cause/source of the differences. In question 3 evaluators are asked to describe in as much detail as possible the problem of equivalence. Then, they are asked to estimate which group, if any, the item favors (item 4). Finally, they are asked to provide an improved item if they can, or a solution to the hypothesized problem with the item (item 5).

Item Analysis Rubric 2

Item Number: _____

Directions:

Using data and notes from Item Rubrics 1.a. & 1.b., for the item pairs *you identified as different* or exhibiting problematic characteristics that makes them incomparable, fill in the rubric below. Please circle the category that best characterizes the nature of the differences you identified:

I. Content Differences **II. Format Differences** **III. Cultural/Linguistic Differences** **IV. Other**

Go to the section circled above and fill in the information required. If you found it difficult to classify the problem or see problems in more than one area, please fill in all appropriate sections below. If you circled "IV. Other", please provide an explanation at the end of the protocol in the appropriate section.

Please describe the issue or problem you see with the item in as much detail as possible. You need not comment on each prompt but please do your best to characterize the items in a complete and descriptive way. We will review these items together during our next session.

I. Content	**Prompts: Consider the Equivalence of:** *Skills or knowledge demanded; vocabulary, ideas, situations, topics; word, expression, sentence or phrasal difficulty; word omission or word addition; grammar; the frequency of words, level of nuance, level of explicitness, literal vs. figurative meaning, the use of metaphor, idiom, etc.*		
1. The content of these items is (*circle one*):	**a) Somewhat Similar (1)**	**b) Somewhat Different (2)**	**c) Different (3)**

2. The problem is best characterized as related to (*circle one*):	**a) Translation** (individual word issues)	**b) Adaptation** (general meaning)	**c) Other**
3. Describe the difference(s) in detail:			
4. Advantage:	**If the item content is different, do you think that it favors one of the groups? Which one? *(circle one*): <u>Russian</u> or <u>Kyrgyz</u>**		
5. Improving Equivalence:	**Can the equivalence problem(s) with this item be resolved? How?**		
<u>II. Format</u>	**Prompts: Consider the Equivalence of:** Overall item presentation, item length, clarity of directions, order of words and ideas, number of words, punctuation, capitalization, typeface, editing and general formatting, etc.		

1. The format of these items is (*circle one*):	**a) Somewhat Similar (1)**	**b) Somewhat Different (2)**	**c) Different (3)**
2. The problem is best characterized as related to: *(circle one)*	**a) Adaptation**	**b) Presentation**	**c) Other**
3. Describe the difference(s) in detail:			
4. Advantage:	**If the item format is different, do you think that it favors one of the groups? Which one? *(circle one*) Russian or Kyrgyz**		
5. Improving Equivalence	**Can the problem(s) with this item be resolved? How?**		

III. Cultural Meaning	**Prompts: Consider the equivalence of:** Russian and Kyrgyz schooling contexts in relation to items, importance or relevance to both cultures, similarity of dispositions, similarity of norms, psychological construct present in both groups, equivalence of linguistic expression, similarity of linguistic structure and grammar, symbolism, metaphor meaningful in both groups, level of explicitness similar, etc.			
1. The cultural equivalence between the two items is (circle one):	**a) Somewhat Similar (1)**	**b) Somewhat Different (2)**	**c) Different (3)**	
2. The problem is best characterized as related to: *(circle one)*	**a) Meaning differences**	**b) Contextual differences**	**c) Linguistic differences**	**d) Other**
3. Describe the Difference(s) in detail:				
4. Advantage:	**If the items are not equivalent for cultural reasons, do you think that it favors one of the groups? Which one? (*circle one*) Russian or Kyrgyz**			

5. Improving Equivalence	**Can the problem(s) with this item be resolved? How?**
IV. Other	**Please explain in detail:**