Russian Language Journal

Volume 64 | Issue 1 Article 8

2014

Accuracy in Predicting Cross-lingual Differential Item Functioning (DIF): A Study of Russian to Kyrgyz Language Test Item Adaptation in the Kyrgyz Republic

Todd Drummond

Accuracy in Predicting Cross-lingual *Differential Item Functioning* (DIF): A Study of Russian to Kyrgyz Language Test Item Adaptation in the Kyrgyz Republic

Todd Drummond

Introduction

Russian-speaking teachers, assessment specialists, and other educators in Eurasia are frequently tasked with effectively translating and adapting sophisticated educational materials from Russian into non-Slavic languages. While standards, textbooks, and other teaching materials have been adapted from Russian to other Eurasian languages for over a century, a contemporary challenge is the adaptation of highly complex, standardized tests and assessments produced in the Russian language (Drummond and Gabrscek 2012). Because the results of educational assessments are often employed in high stakes decision-making, the room for error in the adaptation of cross-lingual tests is small: Capturing exact meaning in all language versions, accounting for cultural nuance, and ensuring corresponding difficulty of the multiple versions in standardized testing are essential.[1]

Due to the lexical, syntax, and other differences between languages, as well as cultural differences between language groups, the test adaptation process is fraught with challenges (Hambleton, 2005). Test developers can only assure the validity of inferences based on assessment results if the content, meaning, and difficulty of test items are similar across different language versions (Camilli and Shephard

[1] Assessment practitioners and researchers employ cross-lingual assessments and tests for various descriptive, analytical and selection purposes both in comparative studies across nations and within countries marked by linguistic diversity. A *cross-lingual assessmen*t is a single assessment (with the same tasks and test questions) that is administered in more than one language (Hambleton, 2005). Employing cross-lingual educational assessments in the Eurasian countries of the former Soviet Union is common practice due to linguistic diversity and the provision of secondary and tertiary education in more than one language (Shamatov, 2012).

1994; Ercikan 2002). It is essential that newly established assessment centers in the Russian-speaking world develop and maintain the institutional capacity to adapt Russian-medium materials to other languages at a high level of quality that can be empirically verified. The appearance of new assessment regimes throughout the region has provided considerable data and the opportunity to conduct studies of that capacity.

The *National Scholarship Test (NST)*, conducted in the Kyrgyz Republic since 2002, is a university admissions test conducted in the Russian, Kyrgyz, and Uzbek languages. Since implementation, this high stakes test has increased the transparency and quality of university selection by providing a reliable, quality assessment from which valid inferences about student performance can be drawn (Shamatov 2012). Early research on the NST has demonstrated that early iterations of the test had reasonably high levels of predictive validity (Davidson 2003). Initially supported by the United States Agency for International Development (USAID) with technical assistance from the *American Councils for International Education*, today the NST is conducted by the *Center for Educational Assessment and Teaching Methods (CEATM)*, the first non-governmental assessment organization in the republic.

With student response data made available by CEATM and funding provided through *American Councils*' Research Scholars Program, a dissertation study was conducted in 2010 to examine two key questions: (1) How accurate were bi-lingual (Russian/Kyrgyz language) test reviewers in predicting how differences in content, meaning, and difficulty across Russian and Kyrgyz test items would impact student response patterns (correct/incorrect answers)? In order to determine accuracy, reviewers' predictions about the scale of difference between the Russian and Kyrgyz items were compared to the results of a statistical test (null hypothesis of "no difference"), using actual student data. (2) What were the primary sources (origins) of difference, if any? Could differences in student outcomes be attributed to variation in cultural interpretation of items, properties of language that made meaning incomparable, technical expertise in translation and adaptation, or other factors? In this paper, the results from the first question are presented and implications discussed.

Key Terms

In the assessment literature, the term test or item *adaptation* from a source language (language items were written in) to a target language (language the original is translated into) is employed to imply a process that produces a variation of the original item that may or may not be a literal translation of the original: Necessary, because literal translation often results in an inadequate correspondence in meaning between two items due to cultural, contextual or linguistic challenges (Camilli and Shephard 1994). In the Eurasian States, test item adapters (called *translators* in most Eurasian contexts) have traditionally relied on *substantive review* of items by bi-lingual educators (reviewers) to ensure measurement invariance across groups. Substantive review in cross-lingual testing relies on bi-lingual experts' *best estimates* of item differences across groups and sometimes a prediction of how language groups may be impacted by those differences (Allalouf, Hambleton, and Sireci 1999). Substantive review relies heavily on the subjective experience, professional judgment, and knowledge of experts, working either in isolation or in review committees, often without statistical analyses of actual student performance (Drummond 2011).

However, when performance data is available and specialists have expertise in statistical methods, these *best estimates* of reviewers can be compared to a statistical test for *differential item functioning* (DIF) of two or more groups (by gender, language, race, or other category).[2] DIF is present when examinees from two or more distinct groups (language in this case) do not have the same approximate probability of responding correctly to a given test item, after controlling for examinee ability (Camilli and Shephard 1994). A large number of DIF items on a given assessment can result in invalid categorization, selection, or policy decisions and consequently have important political and social implications (Ercikan and Koh, 2005; Grisay and Monseur 2007).

[2] It is important to emphasize that the term DIF refers to a *statistically measured difference* after controlling for ability. Item reviewers do not determine DIF levels, but rather predict whether observed differences between group differences are great enough (in their subjective estimation) to lead to DIF between the language versions. The investigator in this study conducted the DIF analyses for each test item. In later sections, both the scoring rubric which quantifies reviewers' estimations as well as the statistical method (logistic regression) used for determining DIF are explicated.

Research in North America and other contexts has shown that substantive reviews are not consistently effective at accurately predicting or interpreting statistical DIF (Plake 1980; Engelhard, Hansche and Rutledge 1990). In cross-lingual assessment situations, if bi-lingual item reviewers can not detect differences or predict performance patterns across language groups with at least a modicum of accuracy, this calls into question the feasibility of accurate test adaptation and hence the feasibility of cross-lingual assessments: Thus, the need to examine the ability of the bi-lingual reviewers to accurately predict statistical DIF. By using statistical approaches to DIF detection, an empirically-based case can be made as to whether or not adapted tests can be considered equivalent in meaning and difficulty and thus fair to all groups assessed.

To my knowledge, no cross-lingual DIF studies involving Turkic and Slavic languages in the Eurasian region have been carried out. This study contributes to an understanding of the unique challenges of test adaptation between these two language groups and the Russian and Kyrgyz languages in particular. Understanding the challenges in predicting and explaining DIF will inform the planning and design of future cross-lingual assessments in the Kyrgyz Republic and elsewhere in Eurasia where Slavic and Turkic languages are routinely employed in educational assessment (Gierl and Khaliq 2001; Jodoin and Gierl 2001).[3]

The National Scholarship Test (NST)

The objective of the NST is to assess the mathematical and verbal reasoning skills of university scholarship applicants. High scorers on the NST are awarded full scholarships to state institutions of higher education (Shamatov, 2012). In 2010, 30,264 examinees sat for the NST; approximately 18,720 in the Kyrgyz medium, 10,994 in the Russian

[3] Recent research has shown that the more disparate the language families involved in a cross-lingual assessment, the more challenging it can be to ensure the equivalence of test forms or unambiguously interpret assessment results (Sireci, Pastula and Hambleton 2005; Ercikan and Koh 2005; Grisay, de Jong, Gebhardt, Berezner, and Halleux 2006; Grisay and Monseur 2007). While there may be some common challenges to cross-lingual test adaptation (regardless of specific languages involved), it is increasingly clear that the feasibility of employing equivalent cross-lingual tests is also a function of the particular languages in question.

medium, and 1,000 in the Uzbek language medium (CEATM[a], 2010). In 2010, the NST lasted 3 hours and 35 minutes and had 150 test items (CEATM[a], 2010). The items analyzed in this study were taken from the NST *verbal reasoning* (словесно-логический) domain. The verbal reasoning format of the NST contrasts with what was historically assessed for university entry in the republic, native language and literature which focused on knowledge of grammar and literary works (Drummond & De Young, 2004).

The verbal reasoning domain on the NST consists of four sections: Reading comprehension (24 items, 3 texts), analogies (20 items), sentence completion (10 items), and grammar use (20 items) (Valkova 2004). All the test items were multiple-choice and each item had three distractors (incorrect answers) and one answer key (correct answer). The 38 items reviewed consisted of eighteen analogy items, ten sentence completion items, and ten reading comprehension items; all item types similar in purpose and format to the items found on the verbal reasoning skills section of the North American SAT.

According to the test developers, the purpose of the analogies and sentence completion sections were to check verbal reasoning skills at the word, sentence and text level. More specifically (text translated from Russian into English):

> "Analogies check (a) lexical richness, (b) ability to analyze logical relations between concepts, (c) ability to find relations (dependencies) between words in pairs (d) ability to determine similarities or differences by one or several indicators, (e) ability to analyze, synthesize, compare, generalize, and classify" (CEATM 2007, 14-16).

The sentence completion items:

> "… check (a) the ability to understand logical connections between different parts of verbal expression, (b) vocabulary richness" (Ibid., 14-16).

In regard to the reading comprehension items:

> "The questions from this section evaluate the ability to carefully read different texts of 400 to 850 words, understand and analyze what has been read. Fragments of texts can be taken from

different domains of knowledge: humanities, social science, and physical science. Popular literature is also utilized. This section has two independent texts and two related text fragments for comparison with each other. Each text or pair of texts is accompanied by questions that check: (a) understanding of the content of the text, its basic concept; (b) ability to interpret portions, connections between such portions in the text; (c) connections between the text and the real world; (d) ability to understand hidden meaning; (e) ability to determine the style of the author and his/her disposition, as articulated in the text, and; (f) understanding of the structure of the text and its connection to content…" (Ibid., 14-16).

Below are two translated versions (from Russian into English) of the type of NST items analyzed in this study. These are example items from a previous NST year as items from the 2010 test have not yet been released to the public. Due to the length of the reading comprehension texts, translations from that section are not provided here. However, the reading comprehension section is similar to the reading comprehension section found on tests such as the American SAT or Graduate Record Examination. For more examples of NST items in the Russian or Kyrgyz languages, including reading comprehension sample items, see Valkova (2004), CEATM (2007), or the center's website at: www.testing.kg.

Example Analogy and Sentence Completion Items

Analogies

Instructions: Every task has five pairs of words. The highlighted pair of words presents a relationship between two words. Determine the relationship between those two words and then select another pair below with the same relationship. The order of the words should be the same as in the example.

7. music: composer
(A) poem : poet
(B) aerodrome : pilot
(C) fuel : engineer
(D) doctor : patient

Sentence Completion

Instructions: Each sentence below contains two to four blanks. There are four groups of possible answers to complete the sentence. Select the best answer to make the sentence logical.

3. _____ to believe this theory, _____ nobody has _____ yet.

(A) It is easy / because / formulated it
(B) It is not possible / for / refuted it
(C) It is easy / although / proven it
(D) It is common / although / cancelled it

(Valkova 2004)

The investigator did not have access to the schools or names of the individual examinees who sat for the 2010 NST. Reliability estimates calculated by the test center for the items analyzed in this study were .907 for the Russian language verbal items and .702 for the Kyrgyz language verbal items.

Methods

Selecting and Preparing Item Reviewers

The first step in the study was to select bi-lingual test item reviewers.[4] It was important that the pool of selected reviewers be skilled bi-linguals, preferably with experience in educational assessment, test item writing and translation (adaptation). After conversations with CEATM, it was determined that eligible candidates could be those with experience writing or adapting NST test items in previous years, those who worked on 2010 NST sections not under study, materials translators with extensive experience, and content specialists knowledgeable about assessment issues. Reviewers selected served as proxies for "as qualified as any other feasible sample" of potential reviewers in the republic, but did not have a conflict of interest due to experience working directly with the items under study.

[4] In the 2011 dissertation study, the term item *evaluator* is used instead of item *reviewer*.

The candidate pool included linguists, translators, philologists, and teachers because the task required not only the identification of linguistic differences in the two language versions but an understanding of student cognition and problem solving skills which would enable accurate predictions as to whether item differences would lead to performance differences by group (Mazor 1993; Ercikan et al. 2004). Potential reviewers were identified with the assistance of CEATM employees. Each prospective candidate was provided with full information about the study. If they agreed to participate, they first completed a questionnaire which elicited detailed information about their language knowledge and skills, as well as educational backgrounds. In order to encourage only true bi-linguals to participate, potential participants were informed in an interview that they would be required to speak and write in Russian and Kyrgyz equally, not only on an individual written analysis but in discussion with their peers – many of whom would be translators, linguists and other knowledgeable specialists. As part of this investigation, reviewers would be required to state and perhaps defend their views on the test items under study using both languages. Several of the candidates who initially applied declined to participate in the study after learning about these high expectations.

Half of the reviewers selected had completed their secondary education in the Russian language medium of instruction and half in the Kyrgyz language medium of instruction. Three reviewers had received higher education in both languages while only two had completed their higher educations in the Kyrgyz language medium. Seven reviewers reported using both languages at work and six of them reported using both languages in the home. None of the reviewers reported that Russian was their primary home language. Interestingly, however, four reviewers reported that they "think" primarily in the Russian 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 three marked that they were equally literate in both languages.

All the reviewers had completed higher education and nine of the ten were women. The majority were women because women are over-represented in teaching and in areas related to pedagogy, translation, philology and linguistics in the republic (De Young, Reeves,

and Valyaeva 2006). As it is primarily ethnic Kyrgyz who are bi-lingual (Russian speakers from other nationality groups tend not to be proficient in Kyrgyz), all reviewers were ethnic Kyrgyz (Korth 2005). The majority of participants indicated that they had more than one workplace because in Kyrgyzstan educators often work in many capacities or teach at more than one institution (Ibid. 2005). None of the selected reviewers had ever participated in a formal DIF analysis. Table 1 presents the background characteristics of those selected to serve as reviewers. All participants signed consent forms and were compensated with an honorarium for their work.

Prior to convening the group of item reviewers, a glossary of technical terms that defined all key concepts was distributed. A pre-test of the item scoring rubrics was conducted with one reviewer in order to determine if adjustments were needed to the glossary or rubrics. The pre-test yielded important results: In addition to the discovery of some minor formatting and typographical mistakes, in the debriefing the pre-test reviewer reported that the most challenging aspect of the item scoring rubric was interpreting the coding categories in section 2.2. Changes were made to the rubric based on this feedback (more below).

Item Scoring Rubrics

In order to answer the research questions, item rubrics were developed to capture not only the evaluators' estimations of content, meaning and difficulty differences between item pairs, but also to elicit hypotheses about the cause or source of those differences. They needed to be short enough to allow efficient rubric administration but thorough enough to ensure that essential data was captured that would enable clear interpretation. The test items selected for analyses were collated in test booklets which consisted of each of the 38 item pairs (1 version in Russian, 1 version in Kyrgyz), one test item pair per page. Rubric 1 was a graphic organizer which required evaluators to provide an initial categorization of the type of differences between versions (if any).

Rubric 2, also called "the scoring rubric," was developed and translated before the investigator arrived in country.[5] At the top of Rubric 2 were the item number and a series of prompts offering possible

[5] Table 2 presents a Russian version of Rubric 2 for Cultural/linguistic differences.

explanations for differences between the language versions. Rubric 2 had the following sections: (2.1) a section to estimate the level of difference(s) in content, meaning, or difficulty (if any) between the two items in the pair (Степень различия); (2.2) a section to identify the specific nature of the difference(s) (Причина различий); (2.3) a section to describe the difference(s) in detail (Подробно опишите различия); (2.4) estimation of which group might be advantaged (favored) by differences (Преимущества); (2.5) suggestions for improving equivalency of the item pairs (Улучшение эквивалентности). Rubric 2 was printed in three colors for each category of difference: Content (violet form), format (green form), or cultural/linguistic (pink form). This color scheme allowed the researcher to easily collate the forms by nature of difference during later analysis. The directions (English translation) for the reviewers for completing rubric 2 follow:

Directions
Fill in item rubric 2 for each item not identified as "identical." The purpose of item rubric 2 is to collect data that will facilitate an understanding of the level and nature of difference as well as the cause (source) of difference for each item. 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 group discussion.

The rubric is broken into three color coded categories. The main categories are: Content differences (purple), Format differences (green), and Cultural/Linguistic differences (pink). Match the color of the rubric that best fits the nature of the difference you identified in 1.b and fill it in. Note that these categories are not always mutually exclusive. However, these three categories provide a strong foundation from which to classify core item issues. You can also note other reasons for difference if necessary on any of these rubrics.

At the top of each rubric, you are provided a series of prompts – or possible explanations for differences. These prompts are not meant to be exhaustive but are examples of issues that can help you classify the

nature of the differences. In section 2.1, please score the item as "somewhat similar", "somewhat different" or "different" per the guidance in the glossary of key terms. Then, in 2.2, circle the most likely cause/source of the differences. In section 2.3, describe in as much detail as possible the problem of equivalence. Next, in section 2.4, estimate which group, if any, the item favors. Finally, in section 2.5, provide an improved item if you can, or a solution to the hypothesized problem with the item.

If you find it difficult to classify the problem or see problems in more than one area, please describe the nature of the problems on one of the rubrics under section 2.3.

This use of item scoring rubrics was adapted from item studies conducted by other researchers (Allalouf et al. 1999; Ercikan et al. 2004).[6] In terms of the estimation of the *level of difference*, a marking scheme was adopted from Ercikan (2002) and Reckase and Kunce (2002), which defined these terms as follows:

0- Identical: no difference in meaning, content, or difficulty between two versions;

1- Somewhat similar: small differences in meaning, content, or difficulty between two versions, will not likely lead to differences in performance;

2- Somewhat different: clear differences in meaning, content, or difficulty between the two versions, may or may not lead to differences in performance between two groups;

3- Different: differences in meaning, content, or difficulty between the two versions that are expected to lead to differences in performance between the two groups.[7]

[6] The term *rubric* is used because reviewers were required to provide numerical estimate for difference levels (above). A full explication of the item types, item scoring rubrics, Russian and English language versions of data collection protocols, etc. can be found in the full study.

[7] In Russian, the descriptors read: "somewhat similar" (небольшие различия), "somewhat different" (средние различия), and "different" (значительные различия).

The use of the above scoring scheme provided a way to score the extent to which reviewers believed that differences in the item versions would lead to DIF. How these scores were tabulated is presented below under the section *Scoring the Reviewers' Predictions*.

The Item Review Process

The review of the thirty-eight item pairs took place over a three day period. The review was a "blind review" which meant that the reviewers did not have access to the DIF statistics (i.e. had no idea of actual examinee performance by group) when they conducted their review. On day one, all reviewers participated in a forty-five minute overview of the item review process, asked questions and clarified expectations. Next, test booklets were provided to each reviewer with each pair of items (Russian/Kyrgyz) set on the same single page. Reviewers were asked to try and solve the items and to take notes only on the most important problems that arose. After going through all the items, item pairs coded as "identical" on rubric 1.b were set aside as they were not needed for the completion of rubric 2. On the first day, all reviewers were seated in individual work stations and asked not to communicate with each other about their initial perceptions.

On day two the reviewers completed the scoring rubric (2) for each item they had marked with any rating other than "identical" on day one. This stage of the process took approximately four hours to complete. At the end of this session, the test item booklets and rubrics were collected and in the evening the investigator reviewed the rubrics to make sure that any items needing special attention would be prioritized for discussion on day three. The purpose of the discussion on day three was to provide reviewers an opportunity to reflect on each item pair by discussing their views with their peers, to think more deeply about the items, and to change their predictions (scores) if necessary. The investigator facilitated the discussion in the Russian and Kyrgyz languages and audio recorded the conversation. A note taker from the test center also recorded the conversations in writing. Areas of agreement and disagreement were noted and recorded.

Collating and Analyzing the Rubrics

Over 150 individual item rubrics were filled out by reviewers. Reviewer scores and data from reviewer discussions were next recorded on a summary table for each item which collated all the individual scores from each of the eight reviewers and group discussion comments for each item in one summary table (Tables 3 & 4, presented below). The summary tables for all 38 items can be found in the full study. All comments in Russian or Kyrgyz from the individual rubrics were translated verbatim into English without editing or synthesis on the summary tables. Reviewers' scores and descriptive data from the individual rubrics and discussion notes provided considerable data about reviewers' DIF predictions as well as hypotheses about the causes of plausible DIF.[8]

Under section 2.3 on the summary rubric, each bullet point and comment represents a statement from a different evaluator. This allows the reader to see both the nature of the issues described in detail as well as the "strength of agreement" in the commentary. For example, if six or seven individuals all seem to be saying the same thing, this is visible. Or, the opposite, if only one or two people are noting certain issues or tendencies, this is also on display. The two example summary tables presented here in Table 3 (item 3, statistical DIF item) and Table 4 (item 2, non-statistical DIF item) differ from the individual rubrics (Table 2) completed by each evaluator in a few important ways. On the summary table, section 2.2, the "nature of difference" data was not recoded from each of the item individual summary tables. Based on feedback from the pre-test about conceptual clarity, reviewers were instructed to focus on item description in section 2.3 and not to worry about categorizing their coding in section 2.2. The *a priori* coding categories under section 2.2 were thus used to guide reviewers' thinking in how best to characterize the differences between the item versions but were not analyzed rigorously.

The "level of difference" on section 2.1 of the summary table was coded under the color-coded categories (content, format, cultural/language, or other) as submitted by each reviewer. Notice in the summary tables that a difference that was defined by one as

[8] A full explication on sources (origins) of predicted and actual DIF can be found in the full study.

"cultural" for example, might have been characterized by another as a "content" issue. Again, these categories were employed as a way to collate the data but the investigator did not focus on the consistency of the reviewers in marking these categories. The important data for analysis was the totality of the description, not how the issues were coded according to each individual reviewer. Otherwise, the summary table in Tables 3 and 4 in this paper reflect the same organizing principles and data as collected from each of the reviewers on each of the individual rubrics.

Scoring the Reviewers' Predictions

Before describing how the statistical DIF levels were calculated for each item, it is necessary to explain how the reviewers' predictions (scores) were tabulated. The critical part of the scoring rubric required reviewers to estimate the *level of difference* between item versions. Recall that the possible values were 0, 1, 2, or 3; the higher the value, the stronger the belief that differences between the item versions would lead to statistical DIF. Recall from above that a score of "3" (different) meant that the reviewer believed that the difference "would likely lead to differences in performance outcomes between the two groups," or, DIF.

The scores for each item were totaled across all reviewers to produce a combined total score for each item.[9] Recall that the accuracy of these predictions would later be tested by doing an actual statistical DIF analysis. The total scores for each item could thus range from 0 to 24 total points per item. For example, 8 reviewers @ 3 possible points (maximum) is equal to a maximum score of 24 total points per item. The quantification of the reviewers' predictions enabled a rank order of correlation estimation between the statistical DIF results and the reviewers' predictions (more below).

In order to facilitate a coherent discussion about each item and come to a common agreement about when the group believed an item was likely to exhibit DIF, it was necessary to set a kind of cut score beyond which "the group" predicted an item to be DIF (as distinct from individual predictions). As there was inevitable variability in scoring by the reviewers, this was not a straightforward task. Before the process of

[9] While ten item reviewers were initially selected, the data from two reviewers were not tallied.

scoring items began on day two, reviewers were asked to consider what total score per item would serve to indicate that *the reviewer group* predicted DIF. Ultimately, it was determined that it was not the item total score (a sum of the scores of 0, 1, 2, or 3) that mattered most, but rather the number of reviewers who predicted difference. The reviewers proposed that four total marks in any combination from the two categories of "somewhat different" (2 points) or "different" (3 points) would be considered a vote for DIF.

Four total marks from these two categories thus served as a "cut score" (resulting in 8 points minimum if considering the sum of scores) for DIF prediction; less than four total marks for any item pair meant reviewers (as a group) believed that differences were not likely to impact group performance (statistical DIF). The rationale for using a certain number of "high marks" as the criterion rather than a total numerical score was the concern that a small number of "outlier opinions" could result in over prediction of DIF. For example, the argument made was that an item could also receive a score or 8 when only three evaluators marked it as DIF – scores of 3, 3, and 2, or 3, 3, and 3, for a total of 9 points. In other words, common agreement by at least half the group was perceived to be of more value than the possibility of 2 or 3 very high scores from just a few group members.

Estimating Statistical DIF

Up to this point, reviewers' predictions of difference and their beliefs about the likelihood that a perceived difference in content, meaning or difficulty would impact performance of a group have been calculated to provide a means of scoring "likelihood of difference" from a subjective perspective. This is, of course, not the same thing as calculating DIF with actual examinee response patterns for each item. As highlighted above, DIF is a function of statistical analysis and needed to be calculated accordingly. The test center did not have the capacity to conduct DIF analyses independently so the investigator calculated statistical DIF estimations for each of the items as a key part of the study.

Logistic regression (LR) analysis was employed to detect DIF levels for each of the 38 items. The LR method is a non-parametric probabilistic approach to DIF detection that utilizes observed scores to test for the likelihood of difference in group performance on an

individual item, after conditioning on ability. In most non-parametric DIF studies, the total test or sub-score (verbal reasoning score here) on the test examined can be used as a proxy for examinee ability (Sireci, Patsula and Hambleton 2005).[10] It is important to recall that aggregated "average score differences" between groups on a given test or item is not an indication of DIF: There was in fact a large achievement gap between Russian and Kyrgyz language groups on the NST with the Russian language groups scoring significantly higher (almost 1 standard deviation) throughout the republic. The comparison that the LR method employs compares only "like students to like students" using ability (test score as a proxy for ability) as a control.

The LR approach to DIF analysis relies on a chi-squared test of statistical significance and has an established measure of effect size.[11] The LR model is easy to implement and has power comparable to other DIF detection methods (Swaminathan and Rogers 1990; Zumbo 1999; Gierl, Rogers, and Klinger 1999; Jodoin and Gierl 2001). The logistic regression model for predicting the probability of a correct response to an item can be formulated as (Swaminathan & Rogers, 1990):

$$P(u = 1) = \frac{e^z}{[1 + e^z]},$$

where: $z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta G)$

For each item analysis, the dependent variable was dichotomous - "1" for correct item response, "0" for incorrect response. On the right hand of the equation (independent variables), θ was a measure of examinee

[10] It is important to note that most non-parametric DIF studies measure on internal criteria. In essence, DIF detection assumes at least a modicum of overall validity because if all items were biased (systematically) no DIF would be evident (Hambleton, Clauser, Mazor and Jones 1993).

[11] This was not the case with LR originally until Zumbo (1999) and Jodoin and Gierl (2001) introduced a pseudo R-squared measure of effect size for LR in DIF analyses. Effect size measures are employed in statistical tests to ensure that high incidences of statistical significance (common in in hypothesis testing with large samples), does not lead to faulty inferences about the meaning of that significance level, i.e., a test can be statistically significant but not necessarily have practical significance.

ability - verbal reasoning scores in this case. Language group membership was a categorical variable "G" and was coded "1" for Kyrgyz or "0" for Russian. The term θG represented an interaction between the two independent variables and in DIF studies serves as a test for non-uniform DIF. Uniform DIF is evident when differences between groups are found across the continuum of ability and non-uniform DIF is evident when an item shows DIF for lower or higher scorers in a group but does not for the other end of the ability spectrum.

In each item analysis, the null hypothesis is "no difference" in item response patterns for the Russian and Kyrgyz language groups under study (Swaminathan and Rogers 1990). A chi-square test of significance was applied to assess this null hypothesis at the .05 level. At 1 degree of freedom at the .05 level, the test statistic was 3.841. Jodoin and Gierl (2001) propose assessing separately for uniform and non-uniform DIF in order to capitalize on the use of a 1 degree of freedom model. Using the steps they propose, each item was assessed in a two-step process using SPSS software. In order to assess for uniform DIF, two models were identified (Swaminathan & Rogers, 1990). A "compact model" - where $z = \beta_0 + \beta_1\theta$ - was entered first. The presence of uniform DIF was then tested by examining the improvement in chi-square model fit when the group membership (G) term was added, resulting in the "full model" ($z = \beta_0 + \beta_1\theta + \beta_2 G$). The chi-square value of the "compact model" was then subtracted from the chi-square value of the "full model" and this difference was compared to the test statistic (3.81) for statistical significance. Then, the presence of non-uniform DIF was tested in similar fashion by examining the improvement in chi-square model fit associated with the "full model" (above) and the addition of the interaction term (θG), or ($z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3(\theta G)$).

In LR, which group is favored by DIF (in terms of who was disproportionately getting the correct answer) is determined by the sign of the β_2 value (Jodoin & Gierl 2001). An early criticism of the LR approach was that it did not have a measure of effect size (Kirk 1996), considered important to reduce type 1 error. An effect size measure ($R^2\Delta$) proposed by Jodoin & Gierl (2001) for DIF detection was employed to address that concern. As with the two step chi-squared comparisons, the effect size from the compact model was subtracted from the effect

size from full model to determine effect size value. The resulting effect size values utilized to classify the practical significance of DIF were:

- Negligible DIF: $R^2\Delta < .035$
- Moderate DIF: $.035 \leq R^2\Delta < .070$, and the null hypothesis is rejected
- Large DIF: $R^2\Delta \geq .07$, and the null hypothesis is rejected

Before conducting the statistical DIF analyses the sample of student item responses was selected. Large sample sizes enable more accurate DIF detection across different combinations of item types, ability distributions and other experimental conditions (Mazor, Clauser, and Hambleton 1992; Rogers and Swaminathan 1993; Hambleton 1993). However, it is also known that in chi-squared tests, large sample sizes can result in high levels of type 1 error (statistical significance of a finding that is a result of a statistical artifact rather than a finding of practical significance). Thus, the need for sample sizes of between 200-1,000 responses. The test version of the NST from which the 38 items were drawn had been administered in all regions of the country and had a total 4,407 examinees.[12] This selection included a total of 1,550 Kyrgyz language and 2,850 Russian language examinees. From this test version, using SPSS software, I randomly selected a sample of 1,000 examinees per language group to be analyzed.[13]

Study Results

Reliability of Reviewers

An important question on the use of reviewers or item raters in any DIF study is the extent to which their estimations can be considered reliable. As bi-lingual reviewers represent a sample of a larger possible domain of reviewers, it was necessary to see how much measurement error existed. After collecting the initial data, the first step of the analysis was to determine the inter-rater reliability of the reviewers' scores and how much variation there was in their estimations. In order to do this an inter-class correlation coefficient was estimated with SPSS statistical software. Inter-class correlations are ratios of rating variance to total variance and can be used as reliability coefficients for assessments of

[13]The investigator did not have access to the schools or names of the individual examinees who sat for the 2010 NST.

raters that are deemed to be in the same category or class (McGraw and Wong 1996).

In order to estimate this coefficient a scoring system was developed that would allow the coding of the reviewers' marks for each item. A matrix of their scores (0, 1, 2, 3) for each item was produced in an Excel file. Each column represented a reviewer and the thirty-eight rows represented each of the items analyzed. Before conducting the reliability analysis two reviewers' scores were dropped from the data set. The one reviewer who had worked as a translator on the NST 2010 filled out only six total scoring rubrics and these six contained a considerable number of missing values. A second reviewer filled out the rubrics incorrectly using the same single rubric to record scores for many different items.[14]

The scores from the eight remaining reviewers were then examined for missing data. There were 13 missing entries from a total of 304 possible entries (38 items x 8 scores). Data was imputed for these missing scores by entering the average scores from the other seven reviewers into each cell where data was missing. Then, Pearson's reliability was calculated in SPSS. Two-way random effects models are used where people effects and measures effects are random. The inter-rater reliability coefficient when "consistency" was selected was .66 with a 95% confidence interval of .473 to .804. This positive correlation is indicative of a fair amount of agreement between reviewers.

Comparing Reviewer Predictions and Statistical DIF Estimations

In all, a total of only six items had no significant statistical DIF as estimated by the LR analysis. These included four analogy items and two sentence completion items (items 9, 2, 24, 7, 17, and 29). Twenty-eight items had statistically significant but negligible DIF (no practical significance), determined based on their effect size values below .035. According to the statistical calculations, only four items were found to have effect sizes larger than .035 and thus were characterized as having

[14] This led to confusion as it was not clear which marks were meant for which items. Approximately one third of her rubrics were filled in this way. Using these rubrics would have demanded considerable guess work in trying to interpret the intent of this reviewer. Nonetheless, after dropping two reviewers, a group of eight remained which provided an ample number of scores for each of the items.

moderate or large DIF.[15] Three items had moderate DIF (13, 19, and 32), and one item had large DIF (item 3). These four DIF items are referred to going forward collectively as "practical DIF" items when distinguishing them from the statistically significant but "negligible DIF" items. All four of the practical DIF items were uniform DIF. The analogy items (numbers 1-20) were spread throughout all classification categories (not significant, negligible, moderate or large DIF) relatively evenly. The sentence completion items (numbers 21-30) were concentrated more heavily in particular categories, typically closer to the DIF cut-off.

The total of only 4 practical DIF items from 38 would seem to indicate that there is considerable capacity for cross-lingual test adaptation in the republic. This represents only about 10% of the items studied. By contrast, Gierl, Rogers and Klinger (1999) found that 52% of English–French item pairs on a Canadian elementary social studies test exhibited DIF. Ercikan & McCreith (2002) discovered DIF rates of 41% on science items from the *Third International Mathematics and Science Study* (TIMSS) test. Robin, Sireci and Hambleton (2003) reported 21% of items on a credentialing exam exhibited DIF when the two languages studied were both European languages: When looking at a European and Altaic language on the same exam, DIF rates were 46%. However, as discussed below and in the larger study, there are also a considerable number of items from this study (10-12) that were both close to the "large DIF" cutoff of .035 and marked as "probable DIF" by many item reviewers. Thus, the practical implications mean that those responsible for analyzing items flagged as DIF might need to consider reviewing items that show some "negligible DIF" as well.

In order to better understand how to interpret the reviewers' predictions and the statistical results, Table 5 presents them in a clear, readable format with an easy to follow logic. The items can be conceived of as arranged in ascending order from most equivalent (top of table) to least equivalent (bottom of table), or from items that are most similar to those that are clearly DIF (last four items in Table 5). Recall that the test statistic used for interpreting the result in terms of the null hypothesis of no difference was 3.81. This means that the first six items with test

[15] For a discussion on how some DIF results may be a result of statistical artifacts, see the full study.

statistic values lower than 3.81 are items for which the null hypothesis of no difference is retained. Note also their p-values above the .05 level. Finally note also that in the LR model, Exp (β) (last column) indicates an approximate ratio of likelihood by both groups to answer an item correctly. Note that the coefficients for these first six items are all at or near "1" which indicates an approximate 1 to 1 correspondence in likely response patterns between groups for those items. The four items at the bottom of the table have practical DIF and have the highest chi-square values, effect sizes over .035, and Exp (β) values far from 1. Of the four items identified as practical DIF, three favored the Russian group while one favored the Kyrgyz group, determined by the direction (+/-) of β_2.

Item reviewers' scores for the items were tallied and a total of eight items were expected to be DIF according to the criteria presented above (four or more scores of 2 or 3). In order to compare their results clearly with the statistical estimations, note that column 2 in Table 5 presents the results for the reviewers' predictions for easy comparison with the actual statistical DIF estimations. Again, the results of the DIF item analyses are presented in rank order by χ^2 difference (chi-squared values from the full equation subtracted from the value from the compact equation, per methods section); and, as the χ^2 values increase (ascending order in the table), the items move closer to medium and large DIF levels.

In column two in Table 5, each X represents a score from the reviewers of either "somewhat different (2 points)" or "different (3 points)." It is apparent by glancing at the data that a modest correlation in reviewers' scores and statistical DIF values does exist. Note that many of the items that received four or more DIF marks from reviewers are clustered closer to the DIF cut off level at .035, in the bottom half of the table. In order to determine the actual relationship (correlation) between the reviewers' predictions and the statistical DIF estimations, a rank order correlation analysis of their estimated numerical scores and the statistical chi-squared values using Spearman's rho was conducted.

In order to calculate the rank order calculation, the numerical score of the items was calculated. The result of the correlation was a significant, positive relationship of .45, .004 significance at the .01 level (Output in SPSS in Table 6). This modest, positive correlation indicates that as reviewers' total scores for the items increased (greater likelihood

of DIF), so did the chi-square difference values (statistical values). This finding would seem to indicate that reviewers' predictions were far from random and that indeed they were "on to something" in their predictions of difference. However, determining the accuracy of the reviewers is not that straightforward, as the discussion below will reveal.

By the standard above, reviewers predicted eight total items to be DIF (Table 7). Seven of the eight items predicted to be DIF by reviewers were statistically significant. The only item predicted to be DIF by reviewers that turned out not to be statistically significant was item 7 (four marks). However, most of the eight predicted items were indeed located in the lower part of the table with relatively high chi-squared values and effect size values close to the cut-score delineating DIF from non-DIF items at .035. Five of their eight predictions had effect size values above the effect size median of .009, i.e. closer to the DIF cut off. For example, item 21 had a .024 effect size and received six marks for DIF from reviewers. Item 11 received five marks for DIF and had a .028 effect size. Item 33 received five marks and had a .027 effect size. It seems that reviewers' moderately accurate estimations in the middle to lower part of the order best explain the positive rank order correlation of .45.

If there is gray area in terms of interpreting the relationship between reviewer and statistical results, it stems mostly from the challenge posed by the large number of items that were perceived as DIF by reviewers that are *close in proximity to the DIF cut-off* as determined by the effect size measure at .035. The range of the effect size measure for negligible DIF items starts at .003 and goes until .029. The median effect size measure is .009. From a practitioner's perspective, it should be noted that some of the items predicted as DIF did in fact have relatively high effect size values but not quite at .035. As statistical estimations are not always flawless, it is possible that some of these "high effect size items" (.024, .027, .028, .029, .031) items might in fact have moderate DIF but the effect size measures were influenced by statistical artifacts in the estimation process. For example, the relatively lower reliability levels of the Kyrgyz items (.702 compared to the Russian .907) might have impacted accuracy in statistical estimations. From a practical standpoint, this might mean that reviewers and assessment specialists select another

10-12 items with high effect size levels for careful review, perhaps reviewing those items with values over .015.

There were of course several outliers in terms of correspondence between reviewer predictions and statistical DIF estimations which kept the overall correlation from being higher. For example, item 15 received five marks from reviewers but had negligible DIF and a fairly low effect size measure of .008. Items 7 and 18 also demonstrated little correspondence between reviewers' marks and the DIF statistics (many reviewer marks but non-significance or negligible DIF with low effect size values). Item 28 was very close to the DIF cut off at .035 but received no marks as DIF from reviewers. Interestingly, looking at the very bottom of Table 5, it is apparent that three of the four practical DIF items did not in fact receive four or more marks from the reviewers. Only item 3 exhibited a high statistical DIF level *and* received many marks (six) from reviewers as probable DIF. Items 13 and 32 had practical DIF but only 1 and 2 marks for DIF from reviewers, respectively. Thus, the positive rank order correlation cannot be attributed to the close correspondence between estimations of the four practical DIF items and reviewers' predictions for these particular items.

Direction of DIF

A key finding of this study was that for the eight items predicted as DIF, reviewers correctly predicted the direction of DIF only 29% of the time (2 of 7 statistically significant items). Note in Table 8 the difference between their predictions and actual DIF direction (which group is favored by item differences) in columns five and six. Five of the seven items favored the Kyrgyz group which means that differences in the items had a discriminatory impact on Russian language examinees. The reviewers were only correct in their predictions with the one practical DIF item (item 3) and with item 21 because, from the eight items they predicted as DIF, with the exception of one lone vote, the reviewers predicted DIF to favor the Russian group for each item.

In general, from the total pool of 38 items, reviewers marked a total of 26 items as "favoring Russian" and only two as "favoring Kyrgyz." One item received a mark of "no advantage" and four items received no marks at all. Of the items that received mixed marks however, the most marks any item received as "favoring Kyrgyz" was one (items 16 and 21). In terms of the four practical DIF items, three of

these items advantaged the Russian group and the evaluators got all three of these predictions correct. Practical DIF item 32, which advantaged the Kyrgyz group statistically, was not predicted to be a DIF item but still received two marks as "favoring Russian."

Discussion and Conclusions

Despite a satisfactory level of inter-rater reliability in reviewer predictions (.66) and a modest correlation between the reviewers' predicted differences and actual statistical DIF for the Kyrgyz and Russian item pairs (.45), the inference that reviewers had a clear understanding of what was going on with the item pairs remains tenuous: The data indicate that reviewers could not correctly predict the "direction of DIF" (which group is favored by difference) on more than a random basis. At first glance, this is a curious finding considering that the low overall number of practical DIF items (4 from 38 total) indicates that there is capacity in the republic to develop equivalent cross-lingual test items, even if we allow that some of the non-DIF high effect size items (.024, .027, .028, .029, .031) may also actually be DIF items. Below, some possible interpretations are offered for this result.

These results contrast with Gierl and Khaliq's (2001) study of cross-lingual DIF that found Canadian reviewers to have better than random prediction rates for DIF direction for French and English versions of mathematics and science items. However, for their 2001 study the reviewers set out to predict DIF direction on item pairs they knew had been flagged as DIF - while in the Kyrgyz case, this was a blind review. Plake (1980) and Engelhard et al. (1990), however, had similar findings with DIF prediction involving a U.S. study where reviewers were tasked with predicting DIF for black and white examinees. Plake (1980) found that the reviewers scored twice the amount of DIF than the statistical procedures yielded. In this study, reviewers also over predicted DIF (8 total items) and only predicted one of the four DIF items correctly (item 3). Engelhard et al. (1990) found that item reviewers could not predict DIF for blacks and whites in the U.S.A when reviewers had no statistical data. This would seem to indicate that predicting DIF in any context is a challenging endeavor.

Engelhard et al. (1990) suggested that one reason for the low agreement in the U.S. was the infrequent use of the category "favors

blacks" and concluded that asking some reviewers to represent the interests of their race in a high stakes situation might have caused stress and influenced their predictions. As in the Engelhard et al. (1990) study, the category "favors Kyrgyz" was only selected twice (two individual marks) in the item review. It seems plausible that in many contexts (not just the Kyrgyz Republic), reviewers enter DIF analyses with the assumption that DIF most often penalizes minority or disadvantaged groups. Thus, one plausible explanation for the one-sided outcome is strongly-held reviewer dispositions towards the groups being compared.

Considering the turbulent language and educational politics in the republic since independence in 1991, perhaps the tendency to mark almost all the NST items as "favoring the Russian group" should not be surprising when considered in post-Soviet context (Huskey 1995; Grenoble 2003; De Young et al. 2006). The ten ethnic Kyrgyz evaluators in this study were certainly cognizant of both the large NST score gaps (favoring the Russian-medium educated), and the overall state of Kyrgyz-medium instruction in the republic (Korth 2005; De Young et al. 2006; CEATM 2010a; CEATM 2010b). To some extent, subtle, even subconscious, tendencies to "defend" the Kyrgyz examinees against what might be perceived as a privileged and historically hegemonic force (the Russian language) might have resulted in a tendency to mark the Russian groups as advantaged without deep reflection upon the differences between item versions.

These findings, along with the previous work in the U.S. context, underscore the need to conceptualize the review of cross-lingual items as a context-bound, social and political process, not simply a technical endeavor. Language as the key variable in DIF studies is invested with symbolic social meaning and language politics can be the means through which power relations between groups are mediated. Participants enter into the review process with dispositions, prejudices and strongly held beliefs, all shaped by individual experience and social context. This finding underscores Grisay et al.'s (2006) point that each study involving language comparison is a unique endeavor in its own right. While Grisay was referring to the specific linguistic properties of the language(s) themselves, this study indicates that there are also potentially important social and political dimensions to DIF studies.

Of course the one-sidedness of reviewers' predictions may not be attributable to reviewer dispositions alone. Throughout the item discussions, reviewers' comments focused almost exclusively on the quality of the Kyrgyz items and the challenge of adaptation from Russian into Kyrgyz, especially the sentence completion items (see full study). No hypotheses were generated as to problems that might lead to items favoring the Kyrgyz group even though the majority of the statistically significant items did, in fact, favor Kyrgyz students. Reviewers indicated that one of the main differences between the Russian and Kyrgyz languages is the extent to which they are "standardized." Indeed, the contested nature of what constitutes "correct literary Kyrgyz" kept the focus of most item analyses squarely on the Kyrgyz items. Almost all of the adaption, content, format and cultural issues raised by reviewers about the items (see full study for complete listing) were related to alleged problems with the Kyrgyz language items.

Discussions typically focused not on the differences in how Russian and Kyrgyz examinees would respond to item differences, but rather on the correct style, grammar, syntax, meaning, and regional vocabulary of the Kyrgyz item versions. An issue that arose consistently in the analyses was the gap between everyday usage and various (disputed) versions of "correct language." It is indeed difficult to compare Kyrgyz and Russian versions of an item if there is little consensus as to what "correct Kyrgyz" should be. And, as reviewers often noted, the Russian items tended to be "quite good." Thus, whatever the political dimensions of language equivalence, the inherent attributes of a language can also make item adaptation challenging, and can potentially affect prediction of DIF.

A lack of reviewer experience could have also contributed to inaccuracy in prediction. While the sample of reviewers represented the most qualified pool possible from the republic, in general, the reviewers had no experience with probability or statistical methods, nor did they have experience as participants in any type of DIF study (as they are unknown in the republic to date). They had no information about the actual statistical DIF outcomes when they completed the individual rubrics and participated in discussion. It is plausible that lack of experience contributed to the focus on such overt, Kyrgyz-related issues

and distracted reviewers from a more nuanced, in-depth examination of the psychology of item response. Russian items at times seemed to be viewed primarily as "references" against which reviewers could check their understandings of the Kyrgyz items.

Perhaps many issues that may have made Russian items more challenging simply went unnoticed *in lieu* of "finding the mistakes" in the Kyrgyz versions (see transcripts in full study for such conversations). It is conceivable that to novice reviewers, mistakes and contestation in one language version are associated with DIF that disadvantages that group. In other words, the "high quality" (and uncontested) items could perhaps become associated (erroneously) with "advantage" while "lower quality" (contested, more mistake prone or allegedly mistake prone) items could become associated with "disadvantage" in the minds of reviewers. The fact that the Russian items were characterized as high quality might have contributed to the assumption that Russians were favored in most instances where differences were evident.

Whether the reasons for inaccurate prediction of DIF direction were due to political dispositions, the focus of attention on "correctness" of the Kyrgyz versions, or lack of experience with such studies, there is nonetheless room for optimism that reviewers in the republic can improve their estimations. First, the reviewers' inter-rater reliability estimate of .66 and the .45 rank order correlation between their estimations and chi-squared values indicate that their overall estimations were not completely random. Second, reviewer marks on direction of DIF were often more tentative than the marks on levels of difference (sometimes item reviewers left the "direction of DIF" blank or only a few reviewers checked this category, see Table 8). This indecision perhaps indicates that inexperience played as an important role in their estimations as dispositions. Third, as Ercikan (2002) argues, DIF study outcomes can differ depending on whether both versions of the items are reviewed simultaneously or individually by reviewers.

Exposure to statistical DIF detection methods by embedding them in some form of practice with action research might improve reviewer accuracy. One way to do this in the Kyrgyz Republic would be to conduct several individual item analyses and later compare the reviewers' preliminary predictions with the actual statistical estimations and discuss the results together as a group. This would underscore the

need to think deeply about the differences between item versions before predicting the direction of DIF and thus serve as a pedagogical tool in item adaptation. This kind of fine tuning and skills enhancement through the introduction of statistical methods holds promise for better analyses in the Kyrgyz Republic. Educators, philologists and assessment adaptors in the Eurasian context working with Russian and other languages could enhance the quality of adaptation of new standardized tests and assessments by introducing both substantive review and empirical statistical testing.

The paradox of this study seems to be that the cultural intimacy of the *within country* DIF study in some ways makes cross-lingual testing more feasible than in broader cross-national comparisons due to the availability of bi-lingual and experienced translators with cultural knowledge: Only 4 of 38 items being flagged for moderate or large DIF is encouraging from the perspective of the testing center. However, cultural proximity may mean that there are added dimensions of sensitive language politics (and subjectivity) when the research touches on questions such as "who benefits from item differences?" While this was not an anticipated result, it should perhaps not be too surprising considering the context of the DIF study and the history of Russian and Kyrgyz language politics in the republic. With the opening of the Soviet space and the introduction of new cross-lingual assessment and testing regimes, the Eurasian context offers a new and rich area for exploration of these and other questions related to challenges in cross-lingual test adaptation.

Table 1: Background Characteristics of Selected Item Reviewers

Profession(s):			
Teacher (secondary and tertiary) (5), Test item writer (3), Philologist/language specialist (6), Methodologist (1), Translator (5), Linguist/editor (2), Lawyer (1)			
Language Medium	Kyrgyz	Russian	Both/Equal
Medium of secondary education?	5	5	0
Medium of higher education?	2	5	3
Main medium at work?	1	2	7
Main medium at home?	4	0	6
Medium in which you think?	2	4	4
Slightly more literate in?	3	4	3

Table 2: Example of Rubric 2, Cultural or Linguistic Difference (Russian Version)

<table>
<tr><td colspan="5">Опросник 2. Культурные/лингвистические различия

Номер задания: _____</td></tr>
<tr><td colspan="5">Рассмотрите задания на эквивалентность по следующим вопросам:

Кыргызкая и русская образовательная среда,
важность и релевантность, сходство нравов,
сходство норм, психологическая составная
присутствующая в обеих группах, эквивалентность
языковых выражений, сходство языковых структур
и грамматики, символизм, значение метафор,
степень очевидности и т.д.</td></tr>
<tr><td>2.1. Степень различия по культурному признаку (обведите одно):</td><td colspan="2">небольшие различия (1)</td><td>средние различия (2)</td><td>значительные различия (3)</td></tr>
<tr><td>2.2. Причина различий (обведите одно):</td><td>(a) Различия в значении</td><td>(b) Контекстуальные различия</td><td>(c) Лингвистические различия</td><td>(d) Другое</td></tr>
<tr><td>2.3. Подробно опишите различия:</td><td colspan="4"></td></tr>
<tr><td>2.4. Преимущества</td><td colspan="4">Если задания не эквивалентны по культурным признакам, у какой группы больше шансов на правильный ответ: кыргызской или русской? (обведите одно)</td></tr>
<tr><td>2.5. Улучшение эквивалентности</td><td colspan="4">Можно ли решить проблему эквивалентности? Как?</td></tr>
</table>

Table 3. Example of Item Summary Table (Statistical DIF Item)[16]

Summary Table						
Item 3 **2.1. Difference Levels**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑					
Content				☑	☑	
Format				☑	☑☑	
Cult/Ling.					☑	
Other						
						☑☑☑☑☑☑

2.3. Describe Differences in Detail:

Content:

- City kids do not encounter "күл"k. (ash) in distractor (г); they live in apartments and don't know what "күл"k. (ash) means because they have not encountered this (so, this is a lack of vocabulary, nuance).
- There is a problem in distractor (A). In the Kyrgyz version, "бак" (tree) can mean both "дерево"r. (tree) and "сад"r. (orchard). In the Russian distractor - "сад" (orchard) is utilized.

Format:

- Misprint in Kyrgyz distractor (В), which is the answer key; wrote "Чоно" (*no meaning*) – should be "Чопо" (clay)
- Orthographical mistake in distractor (В) – student can't understand the word "Чоно" (*no meaning*) - and the result is that they can't find the correct answer.
- Instead of "Чопо" (clay), the word "Чоно" (*no meaning*) is written, a misprint which results in a loss of meaning.
- Misprint with one word in (В) – the word "Чоно" (*no meaning*) should be "Чопо" (clay).
- The word "Чоно" (*no meaning*) should be "Чопо" (clay).
- Misprint – instead of the letter "п" they printed the letter "н" in distractor (В).
- Incorrect letter in word. The word "Чоно" (*no meaning*), in the pair where Kyrgyz is "Чоно" and Russian is "глина" (clay) - should be "Чопо" (clay).

Culture/Language:

- In distractor (а) in the Kyrgyz pair "бак: алма" (tree: apple) - "бак" (tree) can mean *both* "дерево" (tree) and "сад" (orchard) in Russian. However, the corresponding Russian pair is "сад: яблоня" (orchard: apple trees). In the Russian language the Kyrgyz "алма" k. (apple) means "яблоко" r. (apple)

[16] the letters "k" and "r" following words in written in Cyrillic indicate whether the word is a Kyrgyz or Russian word when it is not indicated clearly in the explanation.

and "яблоня" (apple trees) is "алма бак" in Kyrgyz.

- The problem is incorrect translation - "бак" k. (tree) is both "дерево"r. (tree) and "сад" r. (orchard). In Kyrgyz, apple trees is "алма бак"k. which is "яблоня" r. (apple trees) in Russian. The Kyrgyz "алма" k. (apple) is "яблоко" (apple) in Russian.
- The word "бак"k. (tree) and "алма" k. (apple) in comparison to Russian "сад" (fruit orchard) and "яблоня" (apple trees) have many meanings.
- The word "алма" k. (apple) is not correctly translated. The correct variant is ""яблоко" (apple)." (difference in meaning)

2.4. Advantage (DIF Direction): Russian: ☑☑☑☑ **Kyrgyz:**

2.5. Can the items be reconciled?

- Yes, with the correct letter added in distractor (B).
- The translation needs to be tested. You can't rely on only one person for translation.
- Improve translation in distractor (A) by using "бакча"k. (garden)

Discussion:

MK: There are many problems with this item, especially with the item distractors. The first problem I see is confusion in distractor (A) because of the translation of "сад: яблоня"r. (orchard, apple trees) into Kyrgyz is incorrect. The given Kyrgyz version – "бак: алма" (tree, apple). **NO**: Yes, but in Kyrgyz "бак" can mean trees or orchard. **MK**: OK, but we must consider that the Russian variant "сад" (orchard) is only fruit garden, not trees - that is the problem. A better analogy might thus be "tree: apple" – not "orchard: apple." In other words, "from what/where" (material) comes.

MD: I agree, "бак"k. (tree) is "сад"r. (orchard) *and* "дерево"r. (tree). The word "бакча" k. is "огород" (vegetable garden). I think a problem arises in analogies when the Kyrgyz words have many different meanings, and these same words in Russian have only one meaning. I do not know how much this affects overall results but this is true. Again, the problem is the use of multiple meaning and uncommon words in the Kyrgyz language when in the Russian language they have only one meaning. This is a problem of item adaptation.

RM: Another problem is distractor (B). There is a typographical error in this distractor that might cause the question not to work. **ZS:** Yes, the problem is the format (it could have been done correctly, but it wasn't). The results might be influenced by the fact that kids cannot determine the meaning of the word "Чоно" because there is no such word in Kyrgyz! **NO**: Yes, item distractor (B) Чоно is the problem– this question will definitely not work because there is no correct answer; and, there is no way to find the correct answer. AA: I agree, further, many kids in Bishkek do not know the meaning of the word "Чопо" (clay) as this word is rarely used and therefore can lead to problems. So, they couldn't have guessed that there was a misprint in this word.

MD: In regard to city- village kids, we can probably divide kids in into three socio-linguistic groups – Kyrgyz who study in Kyrgyz schools in villages (and don't know Russian), Kyrgyz who study in Russian schools (and speak primarily Russian), and Kyrgyz who study in Kyrgyz schools but communicate often in the Russian language (kids from Bishkek). **AA:** That's true in general, there are different cultural groups who took the test, but I don't see how that effects this item because all the kids tested here took the test only in Kyrgyz, which doesn't impact the result. We can't compare how different Kyrgyz groups will react… but it is clear that the incorrect word use is a problem. Thus, I think the problem is the typographical mistake (format).

Table 4. Example of Item Summary Table (Non-DIF Item)

Summary Table						
Item 2 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑					
Content		☑				
Format		☑				
Cult/Ling.		☑☑				
Other						

2.3. Describe Differences in Detail:

Content:

- In the second word of the analogy pair in the item stem, there are some differences in meaning between the two language groups. In the Kyrgyz stem, the word "шорпо" (broth) suggests "first course," that is "something liquid." In the Russian stem, some people might not understand the corresponding "борщ" (borscht) like "шорпо," as it is the name of a soup.
- Poor translation of item stem: soup is not a direct equivalent to "борщ"r. (borscht) - soup is however, equivalent to "шорпо"k. (broth).
- In distractor (г), there is a Kyrgyz word for "деталь"r. (detail) used in the Russian version. The word is "тетик" in Kyrgyz.

Culture/Language:

- In Kyrgyz, "шорпо" (broth) implies "first course" – soup. In Russian, "борщ" (borscht) is the name of a kind of soup. The item stems are thus not perfectly matched.
- A literal translation of "шорпо"k. (broth) will be "soup."
- The translation of "шорпо"k. (broth) will be "soup." (this is a difference in meaning)
- The word Russian word "деталь" (detail) perhaps won't be understood by village kids as this is a Russian loan word. Should have used the Kyrgyz word "тетик" (detail).
- The word "деталь"r. (detail) – city kids (those who know Russian) will know this, but village kids may not, which will create difficulties in understanding.
- The equivalent word for "деталь"r. (detail) in the Kyrgyz language is "тетик."
- The word "деталь"r. (detail) = тетик; (these are linguistic differences)

2.4. Advantage (DIF Direction): **Russian:** ☑☑ **Kyrgyz:**

2.5. Can these items be reconciled?

Discussion:

MD: I think we agree that the words utilized in the analogy stem are not strictly equivalent; however, there is disagreement as to whether or not this lack of equivalence should be considered a serious enough difference to estimate a lack of equivalence in outcomes. **CJ**: the problem here is the incorrect translation (not adaptation) of the item stem from Russian into Kyrgyz. **KK**: Yes, they are different, but I don't think the differences affect the relationship of the words in the analogy pair.

ZS: also, in regard to item stem (г) it is important to utilize commonly used words, as some terms in this item are rarely used or completely unknown. **NO**: Yes, I agree, the use of uncommon words and terms is problematic. So, the problem is translation, the use of uncommon words, sometimes due to the poorness of the language itself. Some kids in rural areas do not know some of these equivalents, like "деталь" (detail); And, there is a Kyrgyz equivalent for it. It is "тетик," and it should be used.

Table 5: Reviewers' Scores and DIF Statistics in Rank Order by χ^2 Difference

Item	Scores	χ^2 Difference	Effect Size	β_2	sig.	Exp (β)
9	0	0.494	0.000	-0.085	0.482	0.919
2	0	0.733	0.000	0.112	0.393	1.118
24	xx	0.752	0.001	-0.097	0.385	0.908
7	xxxx	1.318	0.000	0.122	0.252	1.13
17	0	2.077	0.001	-0.202	0.149	0.817
29	x	2.369	0.001	0.17	0.125	1.185
39	0	4.733	0.003	0.278	0.031	1.32
35	x	4.796	0.003	-0.242	0.028	0.785
36	xx	4.99	0.003	0.298	0.026	1.347
27	x	5.293	0.003	0.268	0.022	1.307
30	x	6.208	0.004	0.307	0.013	1.359
14	0	6.399	0.004	-0.275	0.011	0.759
31	xx	7.638	0.004	-0.308	0.006	0.735
34	x	9.704	0.006	-0.331	0.002	0.718
12	xx	9.779	0.006	0.385	0.002	1.469
40	0	10.304	0.006	-0.351	0.001	0.704
15	xxxxx	14.890	0.008	0.451	0.000	1.57
18	xxxx	15.464	0.008	0.456	0.000	1.578
10	xxx	15.510	0.009	-0.428	0.000	0.652
37	0	15.595	0.009	0.429	0.000	1.536
8	x	18.174	0.010	0.515	0.000	1.673
4	xx	19.501	0.011	0.574	0.000	1.776
38	xxx	20.21	0.011	-0.507	0.000	0.602
20	0	20.749	0.015	0.741	0.000	2.098
5	xxx	22.576	0.015	0.497	0.000	1.644
25	xxxxx	23.006	0.016	0.583	0.000	1.792
22	xx	23.57	0.013	-0.532	0.000	0.587
26	xx	34.093	0.019	-0.634	0.000	0.531
23	xxx	38.703	0.019	-0.694	0.000	0.5

Item	Scores	χ^2 Difference	Effect Size	β_2	sig.	Exp (β)
21	xxxxxx	42.413	0.024	-0.738	0.000	0.478
16	xx	43.413	0.031	0.98	0.000	2.663
33	xxxxx	43.427	0.027	0.76	0.000	2.138
11	xxxxx	49.326	0.028	0.791	0.000	2.205
28	0	50.145	0.029	0.796	0.000	2.127
19	**xxx**	**94.270**	**0.048**	**-1.171**	**0.000**	**0.31**
32	**xx**	**96.334**	**0.057**	**1.101**	**0.000**	**3.007**
3	**xxxxxx**	**111.086**	**0.05**	**-1.247**	**0.000**	**0.287**
13	**x**	**128.334**	**0.072**	**-1.218**	**0.000**	**0.296**

Table 6: Rank Order Correlation Results

Correlations

			eval	chi
Spearman's rho	eval	Correlation Coefficient	1.000	.451**
		Sig. (2-tailed)	.	.004
		N	38	38
	chi	Correlation Coefficient	.451**	1.000
		Sig. (2-tailed)	.004	.
		N	38	38

**Correlation is significant at the 0.01 level (2-tailed).

Table 7: Reviewers' Scores and DIF Statistics for Items Predicted as DIF

Item	Reviewers' Scores	χ^2 Difference	χ^2 Rank Order	Effect Size
7	*xxxx*	*1.318*	*4*	
15	xxxxx	14.890	17	.008
18	xxxx	15.464	18	.008
25	xxxxx	23.006	26	.016
21	xxxxxx	42.413	30	.024
33	xxxxx	43.427	32	.027
11	xxxxx	49.326	33	.028
3	xxxxxx	111.086	37	.050

Table 8: Prediction of DIF Direction for Items Reviewers Predicted as DIF

1	2	3	4	5	6
Item	Evaluators' Marks	χ^2 Difference	Effect Size	Evaluators Predict*	Statistics Favor
7	*xxxx*	*1.318*			
15	xxxxx	14.890	.008	Russian (5)	Kyrgyz
18	xxxx	15.464	.008	Russian (4)	Kyrgyz
25	xxxxx	23.006	.016	Russian (1)	Kyrgyz
21	xxxxxx	42.413	.024	Russian (5)	Russian
33	xxxxx	43.427	.027	Russian (3)	Kyrgyz
11	xxxxx	49.326	.028	Russian (3)	Kyrgyz
3	xxxxxx	111.086	.050	Russian (3)	Russian

* Numbers in parentheses are number of votes for DIF direction

Works Cited

Allalouf, Avi, Ronald Hambleton, and Stephen G. Sireci.1999. "Identifying the Causes of DIF in Translated Verbal Items." *Journal of Educational Measurement* 36 (3): 185-198.

Camilli, Gregory, and Lorrie Shephard. 1994. *Methods for Identifying Biased Test Items*. London: Sage Publications.

CEATM. 2007. *Rezul'tati obsherespublikanskova testirovaniya i zachisleniya na grantovie mesta vuzov v Kirgizskoi Respubliki v 2007 godu*. [Results of National Scholarship Testing and Enrollment in University Grant Places in the Kyrgyz Republic in 2007]. Bishkek: CEATM. www.testing.kg.

CEATM. 2010a. *Rezul'tati obsherespublikanskova testirovaniya i zachisleniya na grantovie mesta vuzov v Kirgizskoi Respubliki v 2010 godu*. [Results of National Scholarship Testing and Enrollment in University Grant Places in the Kyrgyz Republic in 2010]. Bishkek: CEATM. www.testing.kg.

CEATM. 2010b. *Natsional'noye otsenivanie obrazovatel'nix dostizhenii uchashixsya.* [National Assessment of Educational Quality]. Bishkek: CEATM. www.testing.kg, 2010.

Davidson, Dan. 2003. *Prognozirovaniie uspeshnosti studentov pervyx kursov vysshix uchebnyx zavedenii Kyrgyzskoi Respubliki po rezul'tatam obsherespublikanskogo testa 2003 goda: verifikatsia validnosti testa.* [Prognosis of First Year Student Achievement in Higher Education Institutions in the Kyrgyz Republic According to the Results of the National Scholarship Test 2003: Verification of the Validity of the Test]. American Councils for International Education.

De Young, Alan, Madelaine Reeves, and Gallina Valyaeva. 2006. *Surviving the Transition? Case Studies and Schooling in the Kyrgyz Republic Since Independence*. Greenwich, Connecticut: Information Age Publishing.

Drummond, Todd, and Sergij Gabrscek. 2012. "Understanding Higher Education Admissions Reforms in the Eurasian Context," *European Education: Issues and Studies* 44 (7): 7-26.

Drummond, Todd. 2011. "Predicting Differential Item Functioning In Cross-Lingual Testing: The Case of a High Stakes Test in the Kyrgyz Republic." PhD diss., Michigan State University.

Drummond, Todd, and Alan De Young. 2004. "Perspectives and Problems in Education Reform in Kyrgyzstan: The Case of National Scholarship Testing 2002." In *The Challenge of Education in Central Asia,* edited by Stephen Heyneman and Alan De Young, 225-242. Greenwich, CT: Information Age Publishing.

Engelhard, George, Linda Hansche, and Kay Ellen Rutledge. 1990. "Accuracy of Bias Review Judges in Identifying Differential Item Functioning on Teacher Certification Tests." *Applied Measurement in Education* 3: 347-360.

Ercikan, Kadriye. 2002. "Disentangling Sources of Differential Item Functioning in Multi-language Assessments." *International Journal of Testing,* 2 (3&4): 199-215.

Ercikan, Kadriye, and Tanya McCreith. 2002. "Effects of Adaptations on Comparability of Test Items and Test Scores." In *Secondary Analysis of the TIMSS Results: A Synthesis of Current Research,* edited by David Robitaille and Albert Beaton, 391-407. Dordrecht, the Netherlands: Kluwer.

Ercikan, Kadriye, Mark J. Gierl, Tanya McCreith, Gautam Puhan, and Kim Koh. 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): 301-321.

Ercikan, Kadriye, and Kim Koh. 2005. "Examining the Construct Comparability of the English and French Versions of TIMSS." *International Journal of Testing* 5 (1): 23-35.

Gierl, Mark J., and Shameem Nyla Khaliq. 2001. "Identifying Sources of Differential Item and Bundle Functioning on Translated Achievement Tests." *Journal of Educational Measurement* 38: 164-187.

Grisay, Aletta, John H., A. L. de Jong, Eveline Gebhardt, Alla Berezner, and B. Halleux. 2006. "Translation equivalence across PISA countries." Paper presented at the 5th Conference of the International Test Commission, Brussels, Belgium, July.

Grisay, Aletta, and Christian Monseur. 2007. "Measuring the Equivalence of Item Difficulty in the Various Versions of an International Test. *Studies in Educational Evaluation* 33: 69-86.

Grenoble, Lenore A. 2003. *Language Policy in the Soviet Union.* Boston, MA: Kluwer Academic Press.

Hambleton, Ronald K., Brian E. Clauser, Kathleen M. Mazor, and Russell W. Jones. 1993. "Advances in the Detection of Differentially Functioning Test Items." *European Journal of Psychological Assessment* 9: 1-18.

Hambleton, Ronald K. 2005. "Issues, Designs, and Technical Guidelines for Adapting Tests into Multiple Languages and Cultures." In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, edited by Ronald K. Hambleton, Peter F. Merenda, & Charles D. Spielberger, 3-38. London: Lawrence Erlbaum Associates.

Huskey, Eugene. 1995. "The Politics of Language in Kyrgyzstan." *Nationalities Papers* 23 (1): 549- 572.

Jodoin, Michael G., and Mark J. Gierl. 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): 329-349.

Kirk, Roger E. 1996. "Practical Significance: A Concept Whose Time Has Come." *Educational and Psychological Measurement* 56: 746-759.

Korth, Britta. *Language Attitudes Towards Kyrgyz and Russian: Discourse, Education and Policy in Post-Soviet Kyrgyzstan.* Bern: Peter Lang, 2005.

Mazor, Kathleen M. 1993. "An investigation of the effects of conditioning on two ability estimates in DIF analyses when the data are two-dimensional." PhD diss., University of Massachusetts, Amherst.

Mazor, Kathleen M., Brian E. Clauser, and Ronald K. Hambleton. 1992. "The Effect of Sample Size on the Functioning of the Mantel-Haenszel Statistic. *Educational and Psychological Measurement* 52: 443-451.

McGraw, Kenneth O., and S. P. Wong. 1996. "Forming Inferences about some Intraclass Correlation Coefficients." *Psychological Methods* 1 (1): 30-46.

Plake, Barbara S. 1980. "A Comparison of Statistical and Subjective Procedures to Ascertain Validity: One Step in the Test Validation Process." *Educational and Psychological Measurement* 40: 397- 404.

Poortinga, Ype H. 1983. "Psychometric Approaches to Intergroup Comparison: The Problem of Equivalence. In *Human Assessment and Cross-Cultural Factors,* edited by S.H. Irvine and John W. Berrey, 237-258. New York: Plenum Press.

Poortinga, Ype H. 1989. "Equivalence of Cross-cultural Data: An Overview of Basic Issues." *International Journal of Psychology* 24: 737-756.

Reckase, Mark D., and C. Kunce. 2002. "Translation Accuracy of a Technical Credentialing Examination." *International Journal of Continuing Engineering Education and Lifelong Learning* 12 (1-4): 167-180.

Robin, Frederic, Stephen G. Sireci, and Ronald Hambleton. 2003. "Evaluating the Equivalence of Different Language Versions of a Credentialing Exam." *International Journal of Testing* 3 (1): 1-20.

Rogers, Jane, and Hariharan Swaminathan. 1993. "A Comparison of Logistic Regression and Mantel-Haenszel Procedures for Detecting Differential Item Functioning." *Applied Psychological Measurement* 17 (2): 105-116.

Shamatov, Duishon. 2012. "The Impact of Standardized Testing on University Entrance Issues in the Kyrgyz Republic," *European Education: Issues and Studies* 44 (7): 71-92.

Sireci, Stephen G., Liane Patsula, and Ronald Hambleton. 2005. "Statistical Methods for Identifying Flaws in the Test Adaptation Process." In *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, edited by Ronald K. Hambleton, Peter F. Merenda, & Charles D. Spielberger, 93-117. London: Lawrence Erlbaum Associates.

Swaminathan, Hariharan, and Jane Rogers. 1990. "Detecting Differential Item Functioning Using Logistic Regression Procedures." *Journal of Educational Measurement* 27 (4): 361-370.

Valkova, Inna. 2004. *Getting Ready for the National Scholarship Test: Study Guide for Abiturients.* Bishkek: CEATM. www.testing.kg.

Zumbo, Bruno 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 cores*. Ottawa, ON: Directorate of Human Resources and Evaluation, Department of National Defense.