PREDICTING DIFFERENTIAL ITEM FUNCTIONING IN CROSS-LINGUAL TESTING: THE CASE OF A HIGH STAKES TEST IN THE KYRGYZ REPUBLIC

By

Todd W. Drummond

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirement
for the degree of

DOCTOR OF PHILOSOPHY

Educational Policy

2011

ABSTRACT

PREDICTING DIFFERENTIAL ITEM FUNCTIONING IN CROSS-LINGUAL TESTING: THE CASE OF A HIGH STAKES TEST IN THE KYRGYZ REPUBLIC

By

Todd W. Drummond

Cross-lingual tests are assessment instruments created in one language and adapted for use with another language group. Practitioners and researchers use cross-lingual tests for various descriptive, analytical and selection purposes both in comparative studies across nations and within countries marked by linguistic diversity (Hambleton, 2005). Due to cultural, contextual, psychological and linguistic differences between diverse populations, adapting test items for use across groups is a challenging endeavor. The validity of inferences based on cross-lingual tests can only be assured if the content, meaning, and difficulty of test items are similar in the different language versions of the test items (Ercikan, 2002).

Of paramount importance in the test adaptation process is the proven ability of test developers to adapt test items across groups in meaningful ways. One way investigators seek to understand the level of item equivalence on a cross-lingual assessment is to analyze items for *differential item functioning*, or DIF. DIF is present when examinees from different language groups do not have the same probability of responding correctly to a given item, after controlling for examinee ability (Camilli & Shephard, 1994). In order to detect and minimize DIF, test developers employ both statistical methods and substantive (judgmental) reviews of cross-lingual items. In the Kyrgyz Republic, item developers rely on substantive review of items by bi-lingual professionals. In situations where statistical DIF detection methods are not typically utilized, the

accuracy of such professionals in discerning differences in content, meaning and difficulty between items is especially important.

In this study, the accuracy of bi-linguals' predictions about whether differences between Kyrgyz and Russian language test items would lead to DIF was evaluated. The items came from a cross-lingual university scholarship test in the Kyrgyz Republic. Evaluators' predictions were compared to a statistical test of "no difference" in response patterns by group using the logistic regression (LR) DIF detection method (Swaminathan & Rogers, 1990). A small number of test items were estimated to have "practical statistical DIF." There was a modest, positive correlation between evaluators' predictions and statistical DIF levels. However, with the exception of one item type, sentence completion, evaluators were unable to predict which language group was favored by differences on a consistent basis. Plausible explanations for this finding as well as ways to improve the accuracy of substantive review are offered.

Data was also collected to determine the primary sources of DIF in order to inform the test development and adaptation process in the republic. Most of the causes of DIF were attributed to highly contextual (within item) sources of difference related to overt adaptation problems. However, inherent language differences were also noted: Syntax issues with the sentence completion items made the adaptation of this item type from Russian into Kyrgyz problematic. Statistical and substantive data indicated that the reading comprehension items were less problematic to adapt than analogy and sentence completion items. I analyze these findings and interpret their implications to key stakeholders, provide recommendations for how to improve the process of adapting items from Russian into Kyrgyz and highlight cautions to interpreting the data collected in this study.

ACKNOWLEDGEMENTS

I feel fortunate to have had the opportunity to pursue doctoral work in the College of Education at Michigan State University. I would like to express my sincere gratitude to my dissertation director, Dr. Mark Reckase, for his patient guidance as he mentored me throughout my doctoral studies. Special thanks also to my academic advisor Dr. Jack Schwille for his thoughtful probing and encouragement. Committee members Dr. Jim Fairweather and Dr. Ed Roeber were always accessible and provided constructive feedback. Dr. Michael Sedlak has been a consistent source of moral and financial support for all the students in the educational policy program at MSU. Though not directly involved with this dissertation, I learned a tremendous amount about leadership from Dr. John Hudzik in the Office for Global Engagement and about educational politics from the wisdom of Dr. Phillip Cusick. Thanks to Seung-Hwan Ham and Wang Jun Kim for their friendship and interest in reading my work.

Colleagues and friends at the *American Councils for International Education* in both Washington, D.C., and the Kyrgyzstan field office influenced my thinking about this dissertation. I would like to acknowledge Dr. Dan Davidson, Dr. David Patton, Michael Curtis and Kimberly Verkuilen as well as past and current members of the *American Councils* team in Bishkek for their friendship and the role they have played in my professional development for more than a decade. This dissertation would not have been possible without the support of dozens of colleagues, students and friends in Kyrgyzstan. I thank Nina Dolzhenko, my former students, the entire collective at school number one in Kant, colleagues at the Ministry of Education, and the Shirinovi and Chokubeavi families for “introducing me” to Kyrgyzstan almost two decades ago. I also thank former Minister of Education, Camilla Sharshekeeva, for her friendship, inspirational courage, and tenacious optimism.

Past and current staff of the *Center for Educational Assessment and Teaching Methods (CEATM)* in Bishkek led by Dr. Inna Valkova have not only my deepest gratitude for their enthusiastic collaboration, but my sincere admiration for the outstanding work they do in very difficult conditions: Constantine Titov, Natalia Naumova, Merim Kadyrova and Asel Bazarbaeva, study participants, and the rest of the CEATM team supported me in every possible way in the summer of 2010. This research was made possible by support from the U.S. State Department's (Title VIII) *Research Scholars* program administered by the *American Councils for International Education.*

Finally, I want to express a very heartfelt thanks to my family. Vitaly and Lubov Stolyarovi as well as their extended families in Kyrgyzstan have been constant supporters. I thank my parents, R. Wayne and Gayle Drummond, for their unwavering love and a lifetime of opportunities and encouragement. I dedicate this dissertation to the most important person in my life, my wife and best friend, Natalia, who has been with me every step of the way.

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1: Predicting Differential Item Functioning in Cross-Lingual Testing

Overview

Cross-lingual tests are assessment instruments created in one language and adapted for use with another language group. 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 (Ercikan, 2002; Hambleton, 2005). In 2002, educational policy makers in the Kyrgyz Republic (KR) changed the selection criteria for awarding state scholarships to higher education by replacing oral admissions examinations with a standardized, cross-lingual test (Clark, 2005).[1] The new test, known as the National Scholarship Test (NST), is conducted in May of each year in the Kyrgyz, Russian, and Uzbek languages (Valkova, 2004).

The introduction of standardized testing in Kyrgyzstan[2] merits scholarly attention for many reasons. In general, high stakes selection testing is a political endeavor with distributive consequences. For some students, success on the NST represents a once in a lifetime chance to access higher education (Drummond & De Young, 2004). As NST results are the sole criterion for university scholarship distribution, the public is counting on the NST to be fair to all examinees, regardless of ethnic or language background. Research is needed to determine the extent to which the NST has met its stated goal of reducing corruption in access to university scholarships. Another inquiry worthy of exploration is the extent to which the new selection

[1] Kazakhstan, Georgia, Russia, Ukraine, Azerbaijan and Uzbekistan have also replaced oral examinations with cross-lingual, standardized admissions tests since the collapse of the USSR. The primary rationale for change has been to overcome corrupt practices that have plagued university admissions in the post-Soviet era (Drummond & De Young, 2004; Clark, 2005; Osipian, 2007; Heyneman, Anderson, & Nuralieva, 2008).

[2] The "Kyrgyz Republic" is the official name of the country but "Kyrgyzstan" is also commonly used.

criterion has impacted schooling. Selection testing for tertiary admissions can impact secondary school classrooms as administrators, teachers, and pupils adjust to the incentives created by what is assessed on high stakes tests (Yeh, 2005). While the above issues are important, this study addresses key questions in cross-lingual assessment at the test item level. Though not often the focus of policy makers' attention, item level analyses are essential because valid selection inferences in cross-lingual testing must be based upon the foundation of equivalent test items (Hambleton, 2005).

The Challenge of Cross-Lingual Assessment

The validity of inferences based on the results of any assessment must be carefully substantiated (Messick, 1988). However, cross-lingual testing introduces additional complexity into measurement and interpretive processes. Inferences derived from cross-lingual test results are based on the assumption that the items are measuring the same constructs at the same level of difficulty across language groups. In fact, cross-lingual item adaptation is a highly complex task due to the myriad of linguistic, cultural and psychological differences between groups: Item equivalence, and thus comparability across groups, can not simply be assumed (Hambleton, 2005).

Successful cross-lingual item adaptation requires not only an understanding of test specifications, item aims and content knowledge, but also cultural and nuanced linguistic expertise in order to ensure that all examinees experience "the same" test items (*ibid*, 2005). The evidence from empirical studies of test adaptation across languages is that accurately adapting items is not always a straightforward task (Reckase & Kunce, 2002; Ercikan, 2002; Van de Vijver & Poortinga, 2005; Grisay & Monseur, 2007). Unintentional differences between item versions can manifest themselves in many ways: Variation in content, presentation, translation or

adaptation, format, or mistakes in one version (e.g. grammar mistakes) can all result in differential performance across groups. Even when two language versions of an item appear to be linguistically equivalent and convey similar content, meaning and difficulty, there may be less visible but critically important cultural, contextual, and psychological background differences between diverse groups that impact a group's performance on an item. For example, variation in curricular or content exposure, opportunity to learn, instructional differences or other background phenomena may impact item performance by group differentially (Gierl & Khaliq, 2001; Van de Vijver & Poortinga, 2005; Hambleton, 2005).

One way investigators seek to understand the level of item equivalence on cross-lingual assessments is to analyze items for *differential item functioning*, or DIF. DIF is present when examinees from two or more distinct groups do not have the same probability of responding correctly to a given test item, after controlling for examinee ability (Camilli & Shephard, 1994). Like factor analytic studies, the utility of DIF studies is that they provide an understanding of the measurement invariance of a test between studied groups. A large number of un-interpretable or un-rectifiable DIF items can result in invalid selection, categorization, or policy decisions and consequently have important political and social implications (Ercikan & Koh, 2005; Grisay & Monseur, 2007).[3]

Researchers conduct DIF studies on gender, racial, language and other group differences. When professional capacity and large sample sizes are readily available, such studies typically employ statistical methods to detect DIF. Sometimes, they include a substantive item review to predict or interpret DIF, *post-hoc* (Ercikan, 2002). Substantive review relies on experts' "best

[3] In theory, "rectifiable DIF" (typically the result of overt issues such as translation mistakes) can be directly addressed after DIF analysis and therefore does not represent as serious a threat, assuming that analyses are conducted and steps taken before test scoring.

estimates" to identify and/or interpret differences and estimate how groups will be impacted by those differences. Previous research has shown that substantive reviews are not consistently effective at accurately predicting or interpreting statistical DIF. In some studies, there has been a low correlation between reviewers' predictions and statistical DIF outcomes (Plake, 1980; Engelhard, Hansche & Rutledge, 1990). However, in some recent cross-lingual DIF studies, substantive review has proven to be relatively successful in interpreting DIF causes *post-hoc* (Allalouf, Hambleton, & Sireci, 1999; Gierl & Khaliq, 2001). The choice of substantive review methods, timing of review (before or after statistical analyses), knowledge of whether or not items have been flagged as statistical DIF, expertise of evaluators, and other contextual factors appear to impact the results of such studies (Ercikan, 2002). Ideally, in order to both accurately detect and interpret causes of DIF, both statistical and substantive analyses of test items are needed (Sireci & Allalouf, 2003).

However, in many cross-lingual testing contexts, there is little or no capacity to employ statistical DIF detection methods. In countries of the former Soviet Union, those charged with developing assessments rely almost exclusively on substantive methods in item review and analysis (Drummond & De Young, 2004). Historically, standardized testing was considered "ideologically incorrect" and no investment was made in the field of educational measurement. As there was no standardized testing, there was no need for statistical DIF detection methods. The application of quantitative methods to educational outcomes in general was rare and the validity of inferences was not typically empirically tested.[4] In the development of cross-lingual educational materials in general, substantive review relying on bi-lingual educators and

[4] This is primarily due to the fact that in the Soviet period educational assessment at both the secondary and tertiary levels relied heavily on oral examinations (Drummond & De Young, 2004).

translators (not necessarily panel review, sometimes a single translator) was considered to be a satisfactory verification of adaptation quality. The concept of an educator whose expertise was in "measurement" or "psychometrics" did not exist in the KR until the introduction of standardized testing in 2002 (Drummond & De Young, 2004).

Today there are still no courses offered in higher education in educational assessment and measurement in the KR and only a handful of specialists have received any training in the basic concepts of psychometric theory. To my knowledge, and the knowledge of the test center staff who conduct the NST, no educators associated with the development of assessment instruments in the republic have ever participated in a DIF analysis. In contexts such as the KR where test developers rely on substantive item review, it is essential that bi-lingual personnel be able to identify both overt item differences between language versions as well as predict how differences in examinee backgrounds will impact group performance. If bi-linguals can not detect differences or predict performance patterns with at least a modicum of accuracy, this calls into question the feasibility of accurate test adaptation across groups and hence the feasibility of cross-lingual assessments: Thus, the need to "problematize" the ability of the bi-lingual evaluator to accurately predict DIF in the republic at this time.

There are also good reasons to probe for the quantity and causes (sources) of differential item functioning (DIF) on the NST. 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, & Hambleton, 2005; Ercikan & Koh, 2005; Grisay, de Jong, Gebhardt, Berezner, & Halleux, 2006; Grisay & Monseur, 2007). The Russian and Kyrgyz languages come from very different language families, Slavic and Altaic (Oruzbaeva, 1997). In other words, while there

may be some “common challenges” to cross-lingual test adaptation in general (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.

Research Questions

In this study I explored two research questions related to the two cross-lingual item adaptation issues highlighted above. First, to what extent were bi-lingual item evaluators able to predict differential item functioning (DIF) on cross-lingual, verbal skills test items from the 2010 National Scholarship Test (NST)? This question was answered by determining how accurately evaluators predicted statistical DIF and by how accurately they estimated which language group was favored by DIF. Second, what were the causes or sources of DIF on the Kyrgyz and Russian test items? Were DIF causes related to overt item adaptation issues like poor adaptation or due to background characteristics of examinees such as cultural or inherent linguistic differences in the way a particular language expresses or represents certain meanings or constructs? (Reckase & Kunce, 2002).

To answer these research questions I designed and conducted a substantive review of thirty-eight verbal skills test items from the 2010 NST. Ten bi-lingual evaluators were selected to complete the item review process. This work took place in Bishkek, Kyrgyzstan, in June of 2010, between the time that the 2010 tests were administered and the time examinee score reports were released. The items evaluated consisted of eighteen analogy items, ten sentence completion items, and ten reading comprehension items. The item analysis rubrics developed for this study required the evaluators to: (1) estimate the level of difference(s) (if any) in content, meaning and difficulty between the two versions of each item; (2) characterize the nature of difference(s); (3) describe the difference(s); (4) estimate which group was favored; (5) suggest

improvements to make the items more equivalent; and (6) participate in a group discussion about each item pair.

Then, I analyzed the items for statistical DIF using the logistic regression (LR) method to provide empirical data about the actual item response patterns by language group (Swaminathan & Rogers, 1990). An effect size measure proposed by Jodoin and Gierl (2001) was applied to each item analysis to limit type one error in statistical estimation. With this data I was able to compare the predictions of the evaluators with actual statistical outcomes and analyze the relationship between these two estimation approaches. Data for understanding the causes of DIF came from the item evaluators' descriptions of the items on the evaluation rubrics and the group discussion of each item pair.

Utility of this Study

In general, there are relatively few studies that seek to identify the causes of DIF on cross-lingual assessments (Ercikan, 2002; Hambleton, 2005). To my knowledge, no DIF studies comparing items from the Altaic and Slavic language families have been carried out at the time of this study. An important goal of this study is to contribute to an understanding of the unique challenges to test adaptation between these two language groups: Characterizing the sources of DIF will inform the planning and design of future cross-lingual assessments in the KR (Gierl & Khaliq, 2001; Jodoin & Gierl, 2001). At present, there are large performance gaps between the Kyrgyz and Russian language groups on the NST. Thus, the study touches on not only technical but also sensitive political issues and the results can either provide support for inferences based on the NST or reveal critical areas where further work needs to be done to improve item equivalence.

Performance gaps of course do not automatically mean high DIF levels. There are urban and rural cleavages in educational outcomes in the KR that parallel the language gaps on the NST (chapter two). Despite evidence that demographics, socio-economic conditions, and selectivity bias plausibly best explain the performance gaps by language, poor test adaptation could nonetheless be a contributing factor to these gaps and any improvements in test quality would improve the validity of inferences based on test results. As noted above, the exclusive reliance on substantive review needs to be problematized until there is empirical evidence that bi-linguals can effectively adapt and analyze cross-lingual test items without the help of statistical DIF detection techniques.

The Ministry of Education in the KR has an interest in this study as policy makers seek to enact a selection policy for university scholarships that is fair to all ethnic and language groups (Presidential Decree No. 91, 2002). In the event that item adaptation in this context appears fraught with irreconcilable problems, policy makers have choices: They could consider different policy options like administering separate - not cross-lingual - assessment instruments. They could consider modifications to the NST if the results of this study indicate this might be necessary. Or, they could consider returning to oral examinations and abandoning cross-lingual, standardized testing entirely.

The *Center for Educational Assessments and Teaching Methods* (CEATM), the organization that conducts the NST, also has a stake in the results of this study. While they have procedures in place for test adaptation and item review, results could shed light on weaknesses in these processes and indicate areas for improvement. Results could demonstrate that different approaches to adaptation are necessary for some item types or that more stringent curriculum surveys or other analyses need to be carried out. They could alter they way they invest in

adaptation procedures or take new steps that would improve DIF predictability and lower DIF levels. Methods could be explored such as special equating or scaling methods that take systemic DIF into account by adding points to groups who have been discriminated against.

Finally, outside of the KR, countries continue to join international assessment regimes like the Trends in Mathematics and Science Survey (TIMSS), the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Teacher Education and Development Study in Mathematics (TEDS-M).[5] Cross-lingual testing is likely to remain a highly visible endeavor that includes more and more countries in the coming years (Hambleton, 2005). Many of the newcomers to these regimes are not from countries with high capacity in the field of psychometrics and measurement. Despite the fact that the item development protocols for the above regimes are designed in countries with a longer history in testing and measurement, all countries must still conduct much of the adaptation from core languages (English, French, in PISA for example), into other languages.

In the newly independent countries of the former Soviet Union, substantive item review is still used as the primary means of reviewing and analyzing adapted tests (Drummond, 2011). These Eurasian countries also employ cross-lingual, standardized tests in the Slavic and Altaic languages. Depending on the level and nature of DIF discovered and the efficacy of bi-lingual reviewers in this study, the results could assist policy makers in these countries develop their own assessment capacity. In the short term, the results could help them decide whether or not cross-lingual assessments should serve as the single selection criterion for high stakes university admissions tests (Clark, 2005). In the rest of this introductory chapter I situate the study, define key terms and then set the stage for what follows.

[5] For PIRLS and TIMSS information see: http://timss.bc.edu/; For TEDS-M see: http://www.iea.nl/teds-m.html; for PISA see: http://nces.ed.gov/surveys/pisa/

Situating the Study and Key Terms

Differential item functioning (DIF) is occasionally confused as synonymous with *bias* (Hambleton, 2005). However, this is not a *bias* study. In fact, there are important distinctions between DIF and bias. The term bias tends to imply inherent unfairness and the term is often used broadly in a social rather than statistical sense. Items identified as DIF however, may or may not be fair, depending on the sources of DIF. In cross-lingual studies, item pairs marked as DIF indicate only that the two versions of the item are performing differently in the two groups, not the reasons for that differential performance. Only by collecting and analyzing more information (usually through *post-hoc* substantive review) can it be determined if bias exists. In essence, DIF is an essential prerequisite for bias but is not the same thing as bias (Camilli & Shephard, 1994).

Despite the popular notion that large achievement gaps between groups are usually due to bias in testing, this is not necessarily the case. Zumbo (2003) points out that when two groups demonstrate different probabilities of answering correctly due to *true differences* in the underlying ability being measured, this indicates *item impact,* not item bias. Van De Vijver and Poortinga (2005) help clarify the distinction between DIF and bias by noting that bias occurs when one group of examinees is likely to perform less well because of some characteristic not relevant to the assessment purpose. Or, "A measure is considered to be biased if scores of different language versions of the same instrument are differentially affected by an unwanted and undesirable source of variance" (Van De Vijver & Poortinga, 2005, p. 41). Thus, bias is usually associated with the presence of some "nuisance factors" which hinders our ability to attain a closer approximation of a true score.

Bias on an assessment can be the result of some background factor that differentiates tested populations (differences population experience with testing format) or due to overt, item related issues like translation problems or unclear items due to format mistakes. An example of overt bias on a cross-lingual test is the adaptation of a word from the source language that results in the use of a word with a different or multiple meanings in the target language. The different meaning may result in an item that confuses the examinees in the target language and result in a failure to assess knowledge of the intended, original word. The resulting variance in performance between the two groups can not be said to be due to knowledge of the original word or construct but due to artificially introduced differences in difficulty of the item due to confused word meaning. In this case, the nuisance factor is the poor quality of test translation or adaptation (Hambleton, 2005).

Van de Vijer and Tanzer (1999) identify three types of bias that can impact cross-lingual tests – construct bias, item bias, and method bias. Construct bias occurs when there is an incomplete overlap of psychological or linguistic constructs between the cultural groups in question. Entire ways of conceptualizing problems can differ and existence of certain concepts and ideas might not be found to the same degree between disparate groups (Hambleton, 2005). Method bias can occur due to variation in conditions under which an instrument is administered or due to differences in exposure to certain techniques, like "filling in bubbles," on multiple-choice testing (Van De Vijver & Poortinga, 2005). Method bias was less of a concern for this study as all regions of the Kyrgyz Republic experienced the introduction of standardized testing at the same time. All test NST administrators are trained by the same central authority using detailed test administration manuals and NST testing is conducted under tightly controlled, standardized conditions (Valkova, 2004).

The term DIF is utilized in reference to statistically identified differences in item response patterns, not the claims of item reviewers as to whether or not an item has the same meaning in different groups. That is, evaluators do not identify *DIF* per se, but instead estimate the likelihood of difference in their professional opinion, or - as in the case of post-DIF analysis reviews, provide interpretations as to what might be the cause of DIF identified by statistical methods (Camilli & Shephard, 1994). While some researchers use the terms "DIF review" and "substantive review" interchangeably, in this study I use the term "DIF review" to imply statistical analysis, not the estimation of differences from a substantive review.

At the same time, statistical tests alone reveal nothing about the nature of differences between groups – only that respondents in the groups have different odds of answering an item correctly. In order to interpret DIF it is essential to conduct substantive reviews with bi-lingual expert panels or committees (Ercikan, 2002). As noted above, the reasons for differences in outcomes may be due to true differences or bias. It is possible that through the statistical DIF detection and substantive review methods employed in this study, bias on the NST items will be detected. However, in this study I sought to understand how sensitive bi-lingual evaluators were to overt item and background differences of examinees (i.e. differences from either item impact or bias) as well as how accurately they could predict which group these differences favored - not necessarily distinguish between bias and DIF per se.

Finally, strict measurement equivalence on cross-lingual tests is rare in practice. Van De Vijver and Poortinga (2005) maintain that the constituent elements of constructs like behaviors, attitudes, or norms are never identical across all cultural or linguistic groups. Representatives of different groups are likely to always be somewhat differentially impacted by certain situational types of questions, curricular coverage, and background knowledge. Thus, the finding of some

DIF on the NST does not automatically invalidate the comparative inferences based on the NST: DIF results must be put into perspective and context.

Study Limitations

Despite the importance of DIF studies, this type of analysis gathers only a portion of the validity evidence necessary to support the appropriateness of an assessment for the purpose for which it is employed (Messick, 1988). For example, the absence of DIF on a cross-lingual selection test such as the NST in Kyrgyz Republic reveals nothing about whether or not the domains covered by the test are the most useful for university selection purposes. The various language versions of the test might be equivalent but lack predictive validity. Determining the validity of inferences from any assessment is "an overall evaluative judgment, founded on empirical evidence and theoretical rationales, of the adequacy and appropriateness of *inferences* and *actions* based on test scores" (Messick, 1988, p. 35). In short, the appropriateness of the NST as a selection instrument (and overall fairness) can not be determined by a DIF study alone.

Further, the practical utility of any validity or DIF study in policy making is not always related to the meaningfulness of the study's results. As Margaret Archer (1979) reminds us, educational policy is not a natural response to evolving "societal needs," but rather the expression of the will of actors with the power and ability to influence policy and institutionalize their version of reality. Policy decisions are political and can be arbitrary or confused, or made by policy makers whose intentions are not benign; data and evidence can be utilized, or not (Archer, 1979). In countries like Kyrgyzstan where institutional corruption and test score abuse has a long history, validity issues can be peripheral or even completely irrelevant to policy outcomes (Clark, 2005; Drummond & De Young, 2004). Archer's perspective helps us maintain

realistic expectations as to the power (or lack of power) of validity and DIF studies to impact policy decisions.

Nonetheless, this study provides important foundational validity evidence because the results provide information about the challenges to item adaptation from Russian into Kyrgyz as well as the utility of employing substantive item review. As Messick (1988) emphasizes, there is no way to judge responsibly the usefulness of score inferences in the absence of evidence as to what the scores mean. Of course the overall selection inferences made from cross-lingual assessments can be valid or invalid, even when DIF levels are low. However, if DIF levels are high between various groups tested, and test developers are unable to understand why, there can be few valid selection inferences based on the test, regardless of how transparently test results are utilized by higher education institutions (HEIs).

A final general limitation of the study is that the nature of the work itself is highly interpretive. Test adaptation is a human process and evaluators bring different skills and dispositions to their work (Engelhard, Hansche, & Rutledge, 1990). Evaluators can hypothesize and provide plausible predictions and interpretations, but never be 100% certain of those claims. As will be argued in later chapters, even if the statistical DIF estimates are reasonably accurate, the determination of *the exact* DIF rate by statistical means is also influenced by contextual factors such as differences in the ability distributions of the two groups under study (Narayanan & Swaminathan, 1996). A detailed explication of the statistical limitations in DIF studies is presented in Chapters 3 and 6. This does not mean assessment practitioners and policy makers should not try however. The purpose of a DIF study is to generate empirical data in order to address the root of as many challenging issues as possible and adapt policies and methods accordingly.

Organization of the Study

As Kyrgyzstan is not well known by western scholars, brief historical context about the educational system, language politics in the Soviet period, and the politics of contemporary language issues are provided in Chapter 2. I place particular emphasis on the proportion of pupils being schooled in the various language media as well as educational outcomes by language group. I also present trends over time in enrollment by language medium since the collapse of the Soviet Union. In the last half of Chapter 2 I detail how NST results are utilized as the selection criterion for state scholarships to higher education. In Chapter 3 I review the relevant DIF literature. In Chapter 4 I present the design of the study and the methods utilized to collect and analyze data, in Chapter 5 the results, and in Chapter 6 I analyze and discuss the results as well as offer recommendations for future cross-lingual DIF studies.

Chapter 2: Education & Language(s) of Instruction in the KR

Overview

Every language is a unique system of communication that conveys meaning, ideas, and culture. However, languages evolve and develop in the context of social and political systems. The trajectories of their evolution are thus framed by social conditions and power relations (Korth, 2005). Language use also demarcates class, privilege, and social boundaries, and thus has meaning beyond conveyance of literal meaning. A language can be heavily influenced by a more "powerful" or prestigious language, the interaction with which differs through time and place. Therefore, language as a "variable" in research in multi-lingual societies needs to be understood in relation to other competing languages that co-inhabit the same linguistic and cultural space. This is especially true in societies where one language has enjoyed a hegemonic position over all others for a considerable amount of time as Russian did in the Soviet period (Grenoble, 2003).

Understanding the development and political place of the Kyrgyz and Russian languages in the Kyrgyz Republic helps set the context of this DIF study. This chapter highlights the salient historical and demographic issues related to language and schooling in the republic. After a brief overview of language politics in the Soviet era, contemporary data on school enrollment rates and quality of education by languages of instruction in the republic are presented. The chapter concludes with a discussion of the resilience of the Russian language as a language of instruction in the republic and a brief overview of the higher education system and the contextual conditions that gave rise to the new selection test, the NST.

Historical Context of Languages of Instruction

Education in the Soviet era was characterized by centralized administration and tight ideological control. This resulted in the standardization of most educational norms and practices, and in theory, an egalitarian, mass approach (Glenn, 1995). In Central Asia, as in other parts of the USSR, a success of the Soviet state was the development of a mass education system and the attainment of high literacy rates. According to Dienes (1987), the literacy rate for Uzbek males in 1926 was under 25%: By 1979, over 90% of Uzbeks had access to some form of education. Not more than 3% of the population living on the territory of what is today Kyrgyzstan was literate before the Soviet period. By independence in 1991, literacy rates were near 100% (Fierman, 1991). Despite standardization in approach to educational policy throughout the USSR, the achievement of literacy was initially made through the use of multiple languages of instruction. In the early Soviet years many citizens in Eurasia had their first exposure to formal education through the medium of their native language (Grenoble, 2003).

Some accounts of the Soviet's assumption of power in Central Asia emphasize the widespread poverty and absence of mass schooling at the time (Glenn, 1995). Indeed, the Bolsheviks faced many challenges consolidating power at the end of the Russian Revolution. In addition to establishing law and order and creating new administrative structures, schools had to be built and "new literary languages" had to be developed (Korth, 2005). The written Kyrgyz language as it exists today is a Soviet era creation. The narrative that emphasizes the "educational successes" in the early Soviet period however, has been strongly contested in recent years (Hu & Imart, 1989; Oruzbaeva, 1997; Megoran, 2002).

Whatever the claims of the early Bolsheviks, a limited number of Kyrgyz (and other Central Asian) elites did have access to the written word at that the time of the Soviet conquest

(Hu & Imart, 1989). Further, many people of the Eurasian steppe did not identify themselves as belonging to the distinct ethnic or linguistic groups that the Soviets were busy constructing (Grenoble, 2003; Korth, 2005). The common literary language used by Kazakhs, Kyrgyz and other literate Turkic peoples at the turn of the 19th century was the one learned "at the Tatar speaking medressehs of Ufa, Kazan or, to a lesser extent, in Orenburg" (Hu & Imart, 1989, p.70). While there were differences between the "oral reality and written word," the "Turkic" produced by early writers was mutually intelligible to many on the steppes and had the potential to serve as a unifying *lingua franca*: A language utilizing the Arabic script which could serve to unite, not divide the peoples of the steppes. Hu and Imart (1989) conclude:

> "Such a lofty aim maybe was surrealistic and in any case hard to attain: it demanded time and above all wide autonomy in cultural and educational matters. The impending historical events were to show that this was precisely what the Kazakh-Kirghiz intellectuals lacked, or, more exactly, what they were denied" (Hu & Imart, p. 73).

The development of a written Pan-Turkic language was not to be. Between 1926 and 1931, under the direction of the Soviet authorities, a distinct written Kyrgyz language was developed with the Latin alphabet which would be utilized until the end of the 1930s. The fact that the Soviets initially selected the Latin script for the newly codified Kyrgyz written language indicates that they perhaps felt threatened by the development of a pan-Turkic language that could serve to unify millions of Muslim subjects. Some scholars also contend that Latin (instead of Cyrillic) was selected in order to avoid being seen as "Russifying" the Kyrgyz language while at the same time it avoided the use of the Arabic script (Grenoble, 2003).

Lenin himself spoke of the need to provide educational opportunities in the native languages of the newly "liberated" peoples of Central Asia (*ibid*, 2003).[6] Scholars debate

[6] According to Grenoble (2003), Lenin clearly believed that there should be no "state language" of the USSR. Indeed, Russian never actually became the "state language" of the USSR - at least

whether he believed that the native medium was essential over the long term but there is no question that the Bolsheviks had political aims in mind when calling for education through native language media. Because the Tsar had outlawed native language schools, the Bolsheviks were sensitive to the language question and did not want to alienate Central Asians: Hence, the guarantee of the right to education through native language (Glenn, 1995). This policy of native language education fit well within the overall strategy of "*korenizatsiia*," or "indigenization," the Soviets' initial approach to institutionalizing the communist state through the appropriation and utilization of local elites in visible social, economic and political positions (Grenoble, 2003).

"*Korenizatsiia*" officially began in June 1923, and was seen as necessary to develop a strong communist movement and institutions. In December of 1923, a decree mandated that official documents be produced in the local languages in all the Central Asian Republics and Autonomous Regions. Initially, even ethnic Russian functionaries were encouraged to learn local languages (Fierman, 1991). Grenoble (2003) proposes that it was highly unlikely however, that the committees with policy making power (which were comprised of highly educated, urban, Russian *intelligentsia*) were actually willing to give up power to the "uneducated" indigenous peoples in matters of culture and education. For example, he notes that Soviet philologists, while claiming to support indigenous languages, "did much to influence them (new languages) to acquire a vast number of Russian lexical items, collocational and grammatical patterns, as well as to directly impose Russian orthography and spelling" (*ibid*, 2003, p. 37). Thus, native language development was encouraged to the extent it was politically expedient for the new regime.

---

not until 1990 when the move was more of a desperate reaction to the outbreak of national assertiveness exemplified by the 1989 language laws in many of the republics (Fierman, 1991).

Nonetheless, the Soviets saw basic education and literacy as essential for the economic, social and ideological development of the region. And, for better or worse, they were successful both in establishing new educational institutions and creating written languages for the titular majorities in the new Soviet Republics, including Kyrgyz.[7] Soviet census data show dramatic increases in literacy rates in all the new republics within the first 20 years of Soviet rule. In Kyrgyzstan, as early as 1923 there were reportedly 251 new Kyrgyz language schools out of the 357 total schools in the republic.[8] In 1913, no books had been published in the Kyrgyz language but by 1924 the first school textbooks were already being produced (Grenoble, 2003).

The early emphasis on education through native languages was to be short lived however. Despite the fact that Article 121 on the Soviet Constitution of 1936 guaranteed the right to native language education for the titular majorities in the new Soviet Republics, by the mid 1930s a change towards overt Russification was already evident. In June of 1934, the Communist Party in Kyrgyzstan promoted the maximum use of "Sovietisms" and internationalist terms in the Kyrgyz language (Huskey, 1995; Korth, 2005). In 1938 a law was passed that required all Soviet citizens to study the Russian language. According to Grenoble (2003), "the rationale for the law was the need for a common inter-ethnic *lingua franca* for communication, economic and cultural development, the need for Russian to promote science and advanced training, and defense" (p. 54). This policy had the effect of stimulating growth in the number of citizens studying in the Russian medium of instruction.

---

[7] Huskey (1995) and others have emphasized the artificiality of the creation of the "new" distinct Central Asian Turkic languages where no such delineations had existed prior to the Soviet era. For a discussion on how the Soviets "constructed ethnic identities" see Grenoble (2003), pp.30-40. For more on "constructed languages," see Hu and Imart (1989).

[8] There was considerable Russian settlement in the region before the Bolshevik Revolution of 1917 and these settlers had been sending their children to basic primary schools for some time. Grenoble (2003) notes that while 54% of eligible European children attended school, only 4% of native children did so at the time of the revolution (p. 142).

Further, in 1938 the Latin alphabet was eliminated for written Kyrgyz and replaced by the Cyrillic script. Perhaps not coincidentally, the late 1930s also mark the period of consolidation of Soviet state power and the brutal persecution of all forms of dissent, political and intellectual. Huskey (1995) believes that the Kyrgyz were in no position to resist the overt linguistic Russification of this period. Not only did ethnic Kyrgyz make up just 40% of the total population of the republic - living mostly in the peripheral regions - but the destructive Stalinist purges decimated the ranks of Kyrgyz intellectuals. A historical account of the 1937-38 mass killings of Kyrgyz elites at *Chong-Tash* notes that most of the executed were accused of have connections to "Pan-Turanic" (Turkic) parties, actively working against the USSR (Helimskaia, 1994). Huskey (1995) argues that the obvious "sycophantism" in the immediate post-repression years and the eager embrace of the Russian language on the part of local party leadership further expedited Russification.

By the end the post-war period, Russian language education was clearly the means to professional advancement in industry, agriculture, science, medicine and culture. A 1954 decree in Kyrgyzstan eliminated the requirement that Europeans[9] study Kyrgyz as a second language (Huskey, 1995). Higher education at this time was conducted almost exclusively in Russian and movement through the communist party hierarchy at any level was difficult, if not impossible, without Russian language skills (Glenn, 1995). Another blow to the Kyrgyz language came in the form of Clause 19 of the 1958 education reforms. According to Clause 19, study of the native language (for non-Russians) was no longer compulsory. In Kyrgyzstan, the result of this

[9] The term "European" in the Eurasian context is typically used to denote non-indigenous, usually non-Muslim, inhabitants of the region who are of "European" origin. While most of these groups are highly "Russified" today, the use of this term allows the inclusion of Germans, Jews (considered a "nationality" by the Soviets), Poles, Ukrainians, Belarusians, Moldovans, the Baltic nationalities, etc. into a category of peoples sharing common traits (Russian medium schooled, typically not functional in local languages, etc.).

policy was that Kyrgyz became a marginalized language, for use at the primary levels of schooling, not for secondary schooling and certainly not for those with ambitions to higher education (Grenoble, 2003, p. 57).

The Brezhnev era (1962-84) has been characterized as a time of further Russian language expansion. Propaganda, policy, and funding promoted the Russian language as the great unifier that would bind all peoples of the USSR, enable socio-economic development, allow for demographic mobility of elite cadres and ultimately serve as the language of the new "Soviet man" (Fierman, 1991; Grenoble, 2003). By the end of the 1970s, 82% of the entire Soviet population had at least basic knowledge of Russian. In 1989 a total of 35.2% of all ethnic Kyrgyz in the republic reported fluency in Russian (*ibid*, 2003). At independence in 1991, only 4% of all books in the national library and only 9% of all films produced by the state cinematography industry were in the Kyrgyz language (Huskey, 1995). Translation of educational materials and books in the sciences and industry now became almost exclusively a one way process – Russian to local languages (Grenoble, 2003).

In 1989, only 7% of all schools in the capital (*Frunze* at that time) were Kyrgyz language medium, while 54% were Russian language medium. The rest were mixed medium schools.[10] Approximately 42% of ethnic Kyrgyz who were not in a Kyrgyz medium track were also not studying Kyrgyz as a second language (Fierman, 1991). Between 1989 and 1992, the number of

[10] In a "mixed medium" school, two or more separate cohorts are taught in two or more different languages all in the same school building. Pupils may attend school at the same time of day or be organized so that one language cohort attends in the morning, one after lunch, all depending on the logistical arrangements, size of the study body and facilities available. Class cohorts in mixed schools are known as "Kyrgyz A" or "Russian A." School buildings characterized as mixed medium do not provide *bi-lingual* education. The combined attendance of various linguistic groups in the same school building is an administrative and logistical, not pedagogical, arrangement (Korth, 2004).

Russian language medium schools in the republic declined from 234 to 142. The number of Kyrgyz language schools increased from 1,018 to 1,122. However, without the total enrollment numbers for each group it is difficult to determine the actual proportion of change in enrollment in the Kyrgyz and Russian language tracks. That is because many Russian schools did not in fact close but instead became mixed medium schools by opening parallel Kyrgyz language groups. The number of mixed schools in the republic increased during this same period from 332 to 409 (Huskey, 1995, p. 562).

With *Perestroika* in mid 1980s there were calls for change in this status quo in cultural and educational affairs. The Central Committee of the Kyrgyz Communist Party issued a proclamation on "National (Kyrgyz) and Russian Bi-lingualism" in August, 1988. Supporters of the proclamation bemoaned the lowly status of the Kyrgyz language. However, the proclamation did not challenge Russian language hegemony directly but instead called for the redoubling of efforts to improve the knowledge, use and quantity of instruction in languages besides Russian, including of course, Kyrgyz. The idea was not to "bring the status of Russian down" but rather to bring other languages "up" to equal status (Huskey, 1995).

The 1989 language law however was less equivocal. Its main features included the renaming of Russian place names, the requirement that official government documentation be in Kyrgyz, and the introduction new norms about language use in the workplace. The law called for the mandatory provision of all business and social services in the Kyrgyz language by 1999 (*ibid*, 1995). According to Huskey (1995), perhaps most controversial was Article 8 which required that ethnic Russian managers be able to speak and carry on normal work activities in the Kyrgyz language. For some however, the law did not go far enough. There was a compromise provisions that noted "either language could be used" in certain official activities and that where

Kyrgyz was being used, translation into Russian was to be provided. Further, while Kyrgyz became "the state language," Russian remained "a language of interethnic communication." Critics of the law pointed out that the provision of these "Russian options" reduced the incentive to learn Kyrgyz (*ibid*, 1995).

Events over the next few years would conspire to make implementation of the 1989 language law incomplete. President Askar Akaev, despite occasional opposition, was perhaps more concerned about keeping Europeans in Kyrgyzstan than the "Kyrgyzification" of the political, economic and educational system. In the summer of 1990, deadly clashes between ethnic Kyrgyz and Uzbeks in the south of the country raised the stakes of nationalist politics. President Akaev was a moderate, committed to the development of the Kyrgyz language over the long run, but eager to avoid fanning the flames of nationalism. In 1993, President Akaev extended the time for full implementation of the 1989 language law from 1997 to 2000 (Wright, 1999).

In characterizing language politics in the republic in the 1990s, Huskey (1995) identified two political camps in addition to the centrists. One group consisted of highly Russified elites, the "internationalists," who had both personal and professional stakes in the Soviet system and the Russian language. The nationalists or "indigenizers" were primarily embedded in the various Kyrgyz language committees and enforcement agencies. They were primarily Kyrgyz speakers who sought faster movement on language reform. According to Huskey (1995), both of these groups used "alarmist tactics" to push their agendas, threatening dire consequences for failure to act decisively. President Akaev however, maintained his moderate position through the promotion of the motto "Kyrgyzstan: Our Common Home," intentionally designed to ally European fears of a nationalist resurgence that would jeopardize their livelihoods and security.

By the mid 1990s the rate of European emigration abated as much of the early nationalist fervor faded (Korth, 2005). In fact, new Slavic educational institutions such as the Kyrgyz-Russian Slavonic University were opened and other efforts were made to maintain close cultural ties to Russia. In 1996, the lower house of parliament even approved the return of Russian as an "official" language. The proposal however, was rejected by the upper house in 1998 (Wright, 1999). Nonetheless, at the end of the Soviet period and well into the 1990s, Russian remained the dominant language of higher education and continued to carry significant political and cultural capital in Kyrgyzstan.[11]

There were of course practical realities that impeded significant change to the status quo in regard to the status and use of the Kyrgyz language. The transition to the Kyrgyz language in official and educational life would have been a daunting financial burden under "normal" economic and political conditions; in the wake of the collapse of the USSR, exceedingly difficult (Fierman, 1995). In such conditions, a small, economically dependent state like Kyrgyzstan could hardly have been successful in overseeing the costly transition from Russian to Kyrgyz overnight, or in other grandiose projects such as reintroducing a new alphabet (Latin script) as some had proposed. Korth (2005) provides considerable evidence that there were not only financial challenges, but also attitudes, dispositions, and pedagogical challenges which coalesced to make the transition unsuccessful.

Whatever the dispositions of policy makers, today the Russian language and its role in education have remained relatively unchanged since the Soviet era. Indeed, not only was the

[11] The maintenance of Russian in Kyrgyzstan as the lingua franca of business, government and education well into the 1990s contrasts sharply with other countries of the former Soviet Union; the Baltic countries, the Caucasus and even other neighboring countries in Central Asia with the exception of Kazakhstan. According to Bruner and Tillet (2007), 67% of all those enrolled in higher education in the KR today are receiving it through the Russian language medium.

1989 language law not vigorously applied in practice, in 2000 the Russian language was given new life by becoming an "official" language of the Kyrgyz Republic. The resilience of the Russian language is connected to history, demographics, culture as well as contemporary political, economic, and pedagogical issues, including the relative strength of Russian medium education, a topic addressed in the next section of this chapter (Korth, 2005).

Contemporary Schooling and Language Issues

Kyrgyzstan inherited and maintained both the Soviet tradition of centralized authority and a multi-lingual system of education. The Ministry of Education in the capital, Bishkek, makes all major education policy decisions (Johnson, 2004; Bruner & Tillet, 2007). Education departments in the 7 *oblasts* (provinces), 2 cities of "republican status" (Bishkek and Osh), 40 *rayons* (regions), and 23 *gorono* (city) administrations implement policy at the local level (Census, 2010). Overall, representatives of over 100 *natsional'nosti* (nationalities) reside in the republic.[12] It is possible to receive an education from kindergarten to the completion of an advanced degree in three languages of instruction - Kyrgyz, the state language, Russian, an official language since 2000, Uzbek, the language of the most numerically predominant minority in the republic. There are also four Tajik language schools in the Batken *Oblast*.[13]

Approximately 80% of schools in the republic are designated by the ministry as rural schools. The average size of rural schools is 477 pupils, the average class size is 23.7, and the

[12] Soviet and many post-Soviet Eurasians use the term natsional'nost (nationality) as American scholars would use the term "ethnicity," not citizenship. For example, an ethnic Russian born and raised in Kyrgyzstan (and a citizen of Kyrgyzstan) would nonetheless be considered to be of "Russian nationality." I use the term ethnicity and nationality interchangeably depending on context. While the total number of different nationalities in the republic is over 100, only 15 distinct nationalities have more than 10,000 people or more in the republic today. The total population of the republic in 2009 was just over 5.3 million (Census, 2010).

[13] Opportunities for higher education in the Uzbek medium are limited. There are no higher education institutions offering study through the Tajik medium.

pupil to teacher ratio is 14.9. The average size of urban schools is 774 pupils, the average class size is 26.9, and the pupil to teacher ratio is 17.9 (Herczynski, 2003).[14] Pupils study in class cohorts which stay together as a group and move from teacher to teacher. There can be from one to seven or eight cohorts in a given grade level. Graduating class sizes can thus consist of as few as 10 students all the way up to 120-140 students, though this larger number is usually only found in the largest schools in Bishkek.

According to the data presented in Appendix A, there were 1,911 total schools in the republic in 2003. More recent figures provided by Steiner-Khamsi, Teleshaliyev, Sheripkanova-MacCleod, & Moldokmatova (2011) put the total number of schools at 2,168. Schools in all instructional languages usually contain all grades in one building (complete secondary schools). However, a small number of buildings house grades one to four, *nachal'nee shkoli* (primary schools), and *ne pol'nee sredniye* (basic secondary) grades one to nine. For the 1999-2000 school year primary schools comprised 6% and basic secondary schools 11% of the total number of schools in the republic (Herczynski, 2003).[15] However, it is also quite common that a large number of pupils attending schools with all 11 grades stop their schooling after the ninth grade.

During the Soviet period, approximately half of those completing their last year of compulsory education (*ne pol'nee sredniye*) received one to two additional years of vocational education in a professional *uchilishe* or *teknikum* instead of completing the full 11 years of secondary school which was considered necessary for university study.[16] The proportion of

[14] This is 2003 data. More recent anecdotal reports indicate that class sizes have increased in recent years, especially in Russian language tracks.

[15] While the Russian word "sredniye" literally means "middle," western observers of Soviet and post-Soviet schooling tend to use the term secondary to denote schooling across all grades 1-11.

[16] In the Soviet era mandatory education (basic) took eight years, while "complete" secondary education took ten years.

those finishing complete secondary education has risen in recent years while enrollment in professional technical schools and other specialized vocational schools has dropped dramatically (Herczynski, 2003). According to Brunner and Tillet (2007) only 10% of the 1993 graduating cohort went on to receive higher education while today approximately half the graduating cohort enrolls in some form of higher education. This is primarily due to the availability of inexpensive educational options with the growth of private institutions and the opening of more places for paying students in HEIs in general.

Table 2-1 below presents the proportion of the three largest *natsional'nosti* in the republic by proportion of the total population since 1959. 1959 was selected intentionally because it represents the peak of European habitation in the republic since data was collected in 1926 (Huskey, 1995). 1989 was the last census year before the collapse of the Soviet Union. The decline in proportion of the Russian and the "other" group between 1989 and 2009 is primarily due to European emigration and the higher rate of population growth of non-Europeans (Census, 2010).

Table 2-1: Three Largest Nationalities in the Kyrgyz Republic

Nationality	1959[17]	1989	1999	2009
Kyrgyz	40.5%	52.4%	64.9%	70.9%
Russian	30.2%	21.5%	12.5%	7.8%
Uzbek	10.6%	12.9%	13.8%	14.3%
Others[18]	18.7%	13.2%	8.8%	7.0%

Upon entering the first grade pupils select a language medium according to the options available in their communities. Russian or Kyrgyz, if not the first choice of instructional

[17] Data from 1959 and 1989 come from Huskey (1995). The 1999 and 2009 figures are from census data.

[18] The large "other" category is composed of many nationalities but includes several highly Russified European groups which were at one point quite substantial in the republic. For example, in the 1959 data the Ukrainian population of the Republic was 6.6% while the German population was almost 2% (Huskey, 1995).

medium, is then studied as a second language (Korth, 2005). Some communities provide no language options while others have two to three options. Data on the total number and proportions of pupils studying in the different language tracks in 1989 and 1999 is presented in Table 2-2. As can be seen, the overall proportion of Kyrgyz and Uzbek language enrollees grew while the proportion of Russian language enrollees declined during this period.

Table 2-2: Percentage of Students in Main Language Tracks

	Total/ thousands	**1988/89***	**1998/1999****
Kyrgyz	474	52.4%	63.3%
Russian	323	35.7%	22.7%
Uzbek	106	11.7%	13.4%
Tajik	2	.2%	.3%

**Narodnoye obrazovanie i kultura v SSSR (1989)* ** Herczynski (2003)

A breakdown of the number of schools by language of instruction is presented in Appendix A. At the time this data was collected in 2003, almost 20% of all pupils in the republic studied in mixed schools that offered both Kyrgyz and Russian language tracks.[19] Despite the overall diversity of the republic, there is some geographical clustering of the various language cohorts by *oblast*. Russian medium schools, while found throughout the republic, tend to be concentrated in urban areas, *oblast* capitals and in the north of the country. All Uzbek schools except for one small school in the northern city of Tokmok are located in the south of the country. Mixed Russian and Kyrgyz schools can be found in all *oblasts* while some southern cities and towns offer various combinations of Kyrgyz, Russian, and Uzbek tracks. Appendix B

[19] The reader should avoid extrapolating numbers of actual speakers or proportions of a *natsional'nost* in the population at large from the total proportion of schools by language track. For example, data from 1955-56 indicated that there were 324 Russian medium schools compared to 1,376 Kyrgyz medium schools in the republic at that time (Grenoble, 2003). However, 49% percent of all children studied in the Russian medium while 51% of children studied in the Kyrgyz medium. At that time only about 33% of the population was ethnic Russian (Grenoble, 2003). It is also important to note that in 1955 - just as today - language of instruction is not a marker for ethnicity.

presents the dispersion of students enrolled in the main language tracks by *oblast* in the year 2000. Osh, Djalal-Abad, and Batken are southern *oblasts*, separated from the north by the Tien-Shan Mountain range. The capital city of Bishkek is listed separately in Appendix B and in other presentations of data.

Language track availability reflects the cultural and linguistic demographics of the communities in which a school is located. In both the north and south of the country, some communities are highly homogeneous while others are quite heterogeneous in ethnic constitution. Further, while sometimes a community is ethnically homogenous today, some towns were more significantly influenced by European settlers than others in the recent past. For example, some mining and industrial towns are now mono-ethnic Kyrgyz settlements; in the past however, they had high concentrations of Russian speakers and today have both a considerable number of bi-linguals and retain some Russian language track options (Korth, 2005; De Young, Reeves & Valyaeva, 2006).[20]

The overwhelming majority of ethnic Russians study in Russian language tracks.[21] However, the Russian language tracks are diverse in ethnic composition. Koreans, Ukrainians, Germans, Dungans, Tatars, Kurds, Turks, Kazakhs, Azerbaijanis, Chechens and other *natsional'nosti* also study in Russian language tracks in large proportions (Korth, 2005). Some these *natsional'nosti* speak primarily Russian at home as well as at school. However, with the exception of ethnic Kyrgyz in these schools, functional command of the Kyrgyz language on the

[20] Uzbek communities like Uzgen and Aravan, have both Uzbek and Russian options as Kyrgyz communities in the south have Kyrgyz and Russian options. Larger cities like Osh and Djalal-Abad have all three. Russian language options can still be found in all three southern oblasts in communities such as Mali-Suu (formerly Mali-Sai), Tash-Kumyr, Kizil-Kiya, Kadamjai, etc.

[21] In the nine years since the NST was introduced, CEATM staff can recall only two or three ethnic Russians sitting for the NST in the Kyrgyz language. According to census data, 322 individual Russians claimed Kyrgyz as their native language in 2009.

part of non-Kyrgyz is quite low. In general, only 7% of all Russians over the age of 15 in the republic claim knowledge of a second language. Further, of those Russians claiming knowledge of a second language, more actually claim knowledge of English (42%) than Kyrgyz (36%) (Census, 2010).

In contrast, 42% of all Kyrgyz over the age of 15 in the republic report fluency in a second language; in 94% of such cases, that fluency is in Russian (Census, 2010). Some ethnic Kyrgyz in urban areas grew up in families where the Russian language was the primary home language.[22] However, there are also many bi-lingual Kyrgyz who are schooled in Russian for whom Russian is indeed a second language, learned primarily once schooling began, rather than from the home environment. Kyrgyz language schools are usually attended by pupils whose strongest language is Kyrgyz though they are also attended by bi-linguals with varying levels of Russian competency (Korth, 2005).

The city of Osh has the highest rate of bi-lingualism in the republic with 62% of the population reporting knowledge of a second language.[23] The Naryn region has the lowest bi-lingualism rate with 27% (Census, 2010). While Russian is studied as a second language in most Kyrgyz schools where teachers are available, it is difficult to generalize across the entire Kyrgyz school population in regard to how well these students know Russian: Exposure to language is

[22] While it is perhaps common for an ethnic Kyrgyz person (in the post-Soviet period) to identify their "native language" as Kyrgyz, this does not necessarily indicate that they have functional ability in the Kyrgyz language, especially in urban areas where many non-Russians know and speak primarily (sometimes almost exclusively) in Russian. In general, both Soviet and post-Soviet census data relies heavily on respondent self-reporting. For a discussion on interpreting language data in the Soviet and post-Soviet census see Grenoble (2003), pp. 28-32.

[23] This figure is from the total population (not just above 15 years of age). Of course the second languages known differ by region. In Osh city for example, where 49% the population is ethnic Uzbek, 33% of bi-linguals report their second language as Kyrgyz while twice that amount report their second language as Russian. Tri-lingualism is also common in places like the city of Osh.

not the same as functional capacity in a language. In areas of the country where school study plus the opportunity for daily interaction in Russian is available (primarily in the north as well as in the regional capitals) it is probably safe to assume that the percentages of those who have functional command of Russian are relatively higher than in more isolated, homogenous Kyrgyz or Uzbek communities. By any measure however, ethnic Kyrgyz in Kyrgyz schools are more likely to have command of at least basic Russian than non-Kyrgyz in Russian schools are to have command of Kyrgyz (Korth, 2005).

Regardless of the ethnic constitution of the community, Kyrgyz language schools are typically composed almost entirely of ethnic Kyrgyz pupils with the exception of a relatively small number of non-European *natsional'nosti* (Kazaks, Tatars, Uighurs, Turks, those claiming mixed heritage, or in some cases Uzbeks) who have assimilated into predominately Kyrgyz speaking communities through marriage or long time residence.[24] Many southern Uzbeks and a small number of Tajiks in the Batken *Oblast* are schooled through their native language though many members of these groups also attend Russian medium schools (Herczynski, 2003).

The Status of Russian as a Medium of Instruction

Despite the decrease in the proportion of Russian language enrollment from 1989 to 1999, data from NST participation rates indicates that the Russian language has retained its status as an important language of instruction in the republic since independence. Recall that in 1999, 22.7% of the total cohort was enrolled in Russian medium education. At that time, approximately 20% of the entire population was ethnic Russian or "others" which consisted of

[24] These ethnicities are the only ones (besides Kyrgyz) with more than 1% of their group claiming Kyrgyz as their native language: Tajiks 2%, Tatars 3%, Uighurs 4%, Turks 16%, and Kazaks 26% (Census, 2010). The native languages of these groups are all in the Turkic language family (except for Tajiki). Overall however, 98% of all Kyrgyzstanis report their native language to be that of their ethnicity.

many Russian speaking, European groups. By 2009, the total combination of Russian and "others" was only 14.7% of the total population. Yet, 36% of those sitting for the NST in 2010 did so in the Russian language (CEATM[a], 2010). CEATM estimates that perhaps half of the 10,994 Russian language examinees were ethnic Russians. Thus, Russian as a language of instruction appears to have retained popularity among non-Russians, at least for those seeking higher education.[25]

The under-representation of Uzbek language pupils in the NST is plausibly due to demographic and background factors. In 2009, in the southern Osh *Oblast* (excluding Osh city), 84% sat for the NST in Kyrgyz, 7% in Russian, and 9% in Uzbek. However, according to the 1999 data presented in Appendix B, a full 28% of pupils in the Osh *Oblast* studied in Uzbek language tracks. Demographic data indicates that the overall proportion of the Uzbek population has not declined in this period but increased (Census, 2010). Because of the lag in data points between 1999 and 2009, it is of course possible that there was a mass exodus of pupils from Uzbek to other language tracks. However, the more likely explanation for this under-representation (proportionally) of Uzbeks in the NST is simply the lower overall tertiary matriculation rates of Uzbeks in higher education. Few opportunities to receive higher education in their native language - with the exception of one or two institutions - as well as the high

[25] As students are free to select any language for the NST, NST language is not necessarily indicative of the language of schooling for a given individual. Some bi-lingual ethnic Kyrgyz for example, have received education through different languages of instruction at different times in their schooling. The researcher is aware of children of mobile parents who were sent to regions far from the capital when they were completing their secondary education. In their previous location they studied in Russian for eight or nine years but they completed their secondary education in the Kyrgyz language. Then, they took the NST in Russian. However, according to CEATM, the overwhelming majority of students sit for the NST in the language in which they completed their schooling.

concentration of Uzbeks in high poverty, rural regions in the Fergana Valley are plausible explanations for this state of affairs.

It is also possible that some Uzbeks seek higher education in Uzbekistan instead of Kyrgyzstan. Yet, the Uzbek regime has not been especially welcoming to Uzbeks from Kyrgyzstan in recent years and there are a myriad of visa and other hindrances making the border not as permeable as it once was (Megoran, 2002). In the wake of the summer 2010 violence between Uzbeks and Kyrgyz in the south of the country - and the destruction of several Uzbek higher education institutions - the higher education matriculation rate for Uzbeks is not likely to increase in the near future.

Appendix C presents the number and percentage of NST examinees by language in 2009 and 2010 for all *oblasts* of the republic. In 2010, a total of 30,264 examinees sat for the NST, just under half of all secondary school graduates.[26] In Bishkek and the northern Chui Valley the majority sat for the NST in the Russian language. In two homogenous (Kyrgyz) provinces, Talas and Naryn, the proportion of NST examinees sitting in the Russian language was 20% or more. Interestingly, the longer term trend in NST participation reveals that despite the continued emigration of Europeans out of the republic, as well as the higher growth rates of the Kyrgyz and Uzbek populations, the proportion of Russian language examinees for the NST has actually increased gradually since the NST was introduced in 2002 (CEATM, 2009; Census, 2010).[27]

[26] A caution to interpreting educational data in the republic is that the proportion of various groups in the youth population (for those under 18) is not necessarily same as in the overall population rates in general. The proportions of Kyrgyz and Uzbek populations are higher in the under 18 group than in the overall population (Census, 2010). Therefore, neither NST data, nor other data on language of instruction in general will reflect the overall proportions of an ethnicity in the republic.

[27] Between 1989 and 1993 alone, 50% of the 100,000 Germans left the republic (Wright, 1999). Between 1989 and 1999, the number of Russians dropped from 916,558 to 603,198 (Korth, 2005, p. 119).

Approximately 60% of all examinees in 2010 sat in the Kyrgyz language, down from 63% in 2009. These figures are down from highs of 71% in 2003 and 69% in 2004 (American Councils for International Education[a], 2004). In those same two years, 23% and 26% sat in Russian respectively. By 2010, the proportion of Russian language examinees reached 36% and the proportion of Uzbek examinees had dropped to 4%. From 2009 to 2010, the proportion of Russian medium examinees increased by two to three percentage points in five *oblasts* and the two major cities of Bishkek and Osh, decreased in one *oblast* (Naryn) and was the same in two others, Talas and Batken (CEATM[a], 2010).

One explanation for these percentages is simply the difference in tertiary matriculation rates for the various language groups as a whole. For example, the participation data from 2003 indicate that 93% of all eligible school leavers from Bishkek (n= 5,089) sat for the NST. Only 34% of all those eligible from the southern Djalal-Abad *Oblast* (n= 4,852) did so in that same year. The Bishkek cohort overwhelmingly sat for the NST in Russian while the Djalal-Abad cohort overwhelmingly sat for the NST in Kyrgyz. These two NST participation rates were the highest and lowest in that year (American Councils for International Education[a], 2004).

Another explanation is that Russian medium enrollment has held steady or perhaps grown since 1999, despite the continued emigration of Europeans.[28] Several scholars have suggested there was an initial "surge" in Kyrgyz medium enrollment in the wake of independence in the early 1990s and following the passage of the 1989 language law (Fierman, 1995; Korth, 2005). In the early 1990s, parents might have believed that their children would have had better life chances with a Kyrgyz language education; by the mid 1990s, this was no longer so apparent (Huskey, 1995). The 2009 and 2010 graduating cohorts began schooling in the late 1990s, after

[28] Unfortunately, I was unable to locate enrollment data beyond the 1999-2000 year.

the initial wave of nationalistic enthusiasm had subsided. The trend in NST participation data towards Russian language corroborates the assertion made by De Young (2007) in his ethnographic study of teachers in rural Kyrgyzstan. According to De Young:

> "… In each of our Chui and Naryn samples, we heard stories about initial enthusiasm for learning in Kyrgyz among Kyrgyz parents and teachers, but a swing back to Russian as at least equally important by the late 1990s. Ethnic Kyrgyz teachers teaching in Russian language schools in particular made this claim. Kyrgyz nationalism and pride led early-on to demands for Kyrgyz as an instructional language in many schools, but the lack of Kyrgyz language texts and other printed materials was one immediate problem, as was the realization among many parents that at the universities they wanted their children to attend, classes were usually taught in Russian." (p.5)

There are of course other plausible explanations for the over-representation of Russian medium examinees on the NST. First, as higher education itself has traditionally been primarily a Russian medium endeavor, it is understandable that more Russian speakers matriculate (Korth, 2005; Bruner & Tillet, 2007). Urbanites are still more likely to receive higher education in general than their rural peers (Census, 2010). More rural Kyrgyz and Uzbeks have perhaps simply become relatively more marginalized in recent years due to increasing poverty – and thus, less able to afford higher education.[29]

It is also theoretically possible that Russian language schools are actually gaining slightly in overall enrollment proportion as they attract more Kyrgyz and Uzbek youth due to the increasing disparity (or perception of disparity) in the quality of education between the different language tracks (Herczynski, 2003). In a series of interviews with key stakeholders, Toursunov

[29] Data on language enrollment represent the overall number of pupils enrolled throughout all grades, 1-11. A grade level breakdown of data might be revealing as those eligible for higher education typically have "complete" secondary education (11 years of schooling). A large number of pupils finish schooling after grade nine. It is quite possible that the proportion of Kyrgyz language track pupils who exit schools after grade 9 is higher than for the Russian tracks. That is, Russian track enrollment in grades ten and eleven might be 27- 30% of the total, not 22%. I was unable to find data to confirm this hypothesis.

(2010) provides anecdotal evidence from a credible source at the Ministry of Education that the demand for Russian-language education has grown in recent years. According to one of his respondents, the head of the secondary education department at the Ministry of Education:

> "There is a definite trend where the number of children at Russian schools is increasing … there are many overcrowded schools. Every year, 5,000 to 6,000 children start attending Russian schools … even many children who attended Kyrgyz primary schools switch to Russian secondary schools when the time comes…" (Toursunov, 2010).

Indeed, since NST inception in 2002, highly publicized test results have indicated that the Russian medium examinees have consistently and significantly outperformed their Kyrgyz and Uzbek peers. In 2003, expressed in *z*-scores, the average difference between Russian and Kyrgyz mathematics test scores was .94, almost a full standard deviation in difference. The Russian-Uzbek difference was just over one standard deviation at 1.04 (American Councils for International Education[a], 2004). The urban-rural divide parallels the language divide. The data from 2010 (below) reveal similar gaps.

Table 2-3: NST 2010 Scores by Language of Instruction

Test Language	No. Participants	Mean Score	Std. Dev.	Cronbach's Alpha
Kyrgyz	18,270	103.4	24.6	.89
Russian	10,994	131.4	38.4	.95
Uzbek	1000	100.8	23.3	.87
Republic	30,264	113.5	33.2	.93

(CEATM[a], 2010) [30]

[30] Data is not aggregated by gender in this study. However, females have consistently both outperformed males on the NST and received more scholarship places than their male counterparts. In 2003, the first year data was collected, females captured over 60% of the scholarship places. The gap is most in favor of females from rural areas. In 2003, almost 66% of all winners from rural regions were young women (American Councils for International Educationa, 2004).

Quality of Education by Language of Instruction

By many accounts there is a crisis in education in the republic today. There is a shortage of funding for education, a shortage of quality teachers, textbooks, teaching materials and a crumbling infrastructure (OSI, 2002; De Young & Santos, 2004; Korth, 2005; Silova, 2009; Shamatov & Niyozov, 2010). However, the "crisis within the crisis" is the state of education in the Kyrgyz medium schools (Toursunov, 2010; Shamatov, 2011).[31] While the dramatic differences in schooling outcomes seem clear, the reasons for these gaps in some ways are straightforward, and in other ways complex and multi-faceted.

NST results do not provide data about the quality of education in the republic because the NST was not designed to assess educational quality. Further, NST examinees are not fully representative of the student cohort as the NST is an optional, tertiary admissions test. However, in recent years representative studies of educational quality have been conducted in the republic. In 2006 and 2009, with support from the World Bank, Kyrgyzstan participated in the *Programme for International Student Assessment* (PISA). In 2007 and 2009, a nationally representative evaluation of educational quality was also conducted (CEATM[b], 2010).[32] Both studies utilized sophisticated sampling designs which adequately covered all three languages of instruction and demographic regions of the country. While the purposes of these assessments differed, the results of both studies indicated wide performance gaps by language tracks.[33]

[31] And, one could of course contend, the Uzbek medium. However, for obvious political reasons, the public and government focus has been on the state of Kyrgyz language education.

[32] During Soviet times, standardized assessments for determining education quality were not conducted (Bereday, 1960).

[33] See www.testing.kg for the technical reports of both NAEQ and PISA. For general data on PISA, see http://www.oecd.org/dataoecd/15/13/39725224.pdf.

PISA assesses the "life skills" of fifteen year olds in reading, mathematics, and science. One hundred and one schools and 3,412 pupils participated in PISA Kyrgyzstan 2006, including 54 Kyrgyz schools, 34 Russian schools, and 13 Uzbek schools. Kyrgyzstan showed the poorest results in all three subjects from all participating countries. The average mathematics score for the KR was 311. However, when aggregated by language of instruction, the average Russian track score was 331.5 while the average Kyrgyz track score was 286.7. Table 2-4 below presents a breakdown of the Russian and Kyrgyz cohorts by percentages in the various mathematics score ranges.

Table 2-4: PISA 2006 Mathematics Scores by Language of Instruction

	Percentage of Examinees in Score Range					
	100-180	180-240	240-300	300-360	360-420	420-500
Kyrgyz	6%	12%	33%	42%	7%	0%
Russian	3%	4%	17%	42%	27%	7%

www.testing.kg

The results from the *National Assessment of Educational Quality* (NAEQ) by language of instruction were similar. NAEQ assessed knowledge and skills at both the 4th and 8th grade levels. Unlike the NST and PISA, the NAEQ was explicitly intended to assess how well students were mastering national standards in mathematics, reading comprehension and science. Over 3,000 pupils in schools across the country participated in both the 2007 and 2009 test administrations. Cleavages in grade eight results by language of instruction were large (see Table 2-5).

Table 2-5: NAEQ 2007 Reading Scores by Language of Instruction

Levels Achieved	Kyrgyz	Russian	Uzbek
High	.8%	5.8%	0%
Above Base	4.5%	15.6%	3.0%
Base	12.3%	23.4%	7.8%
Lower than Base	82.4%	55.1%	89.1%

(CEATM[b], 2010)

The first qualification to interpreting the data about educational quality is the aggregation of the data itself: Recall that the language gap closely parallels the urban-rural divide. Most Kyrgyz and Uzbek schools are concentrated in rural, resource-poor areas while Russian schools are more typically found in urban areas. Rural pupils miss more school hours at harvest time, have teachers who are less educated, and in general face greater poverty levels (Herczynski, 2003). Urbanites are two times more likely to have higher education than their rural counterparts (Census, 2010). This makes disentangling the various explanations for the gaps challenging though it is probably safe to assume that socio-economic conditions, rather than language of instruction itself, is of primary importance. Appendix D presents data on poverty levels, levels of higher education, NST score averages, and percent sitting for the NST in the Russian language for each of the *oblasts* and Bishkek. The poorer southern regions (except for the city of Osh) have the highest poverty levels, the lowest levels of higher education per capita, and the lowest NST scores from the entire KR.

While demographics plausibly explain most of the disparities in educational outcomes, there is evidence that Kyrgyz language schools face unique challenges, regardless of location (Korth, 2005). According to Toursunov (2010), though the Russian Federation provides 40,000 to 70,000 school textbooks a year to the republic, only 60% of all Russian schools have enough. The state of textbook provision to Kyrgyz schools is worse with only 39% of schools having adequate textbooks. This is especially challenging for teachers who by tradition are accustomed to teaching with textbooks (De Young et al., 2006). Further, according the Asian Bank's 1997 School Mapping Project, teachers in Russian medium, urban schools, have significantly more contact hours (seven more per week) with students than their urban or rural Kyrgyz school counterparts (Herczynski, 2003).

Finally, in 2010, many policy elites raised in the Soviet era continue to send their children to Russian schools (Korth, 2005). Those who are responsible for improving the situation in education are not necessarily personally affected by the low quality of Kyrgyz schools. In essence, there may be a class element to choice of language of instruction and some "selectivity bias" at work. De Young (2007) presents data from ethnographic research in the Naryn *Oblast* in which participants associate Russian medium schooling with modernity, sophistication and cosmopolitanism. The loss of opportunities in the Russian language in this rural region was perceived by some to be a serious problem, despite the fact that 98% of the provinces' population is ethnic Kyrgyz (Census, 2010).

Several of De Young's (2007) respondents also noted differences in classroom cultures between Russian and Kyrgyz schools. The acquisition of Russian was linked by some to active and independent learning. According to a teacher from At-Bashy:

> "… If you tell something to a kid, he will obey without delay; and kids will not express their opinions or defend their opinions, just do what you told them, and that's it. (But) in Russian language schools, kids defend their points of view; they can even add something better, or even change the direction of an assignment… in sum, they are more or less - how to say - maybe more democratic? … I would say (this) is partly to do with the community where they live. You know, when we teach Russian, and start teaching Russian literature and Russian lyrics of freedom, it has its impact in child development. (Meanwhile), Kyrgyz (stories) also has the same freedom lyrics and also the same democrats and fighters (akyns and writers), but (it is not the same)…." (p. 9).

De Young (2007) summarizes his interviews with teachers at a Russian school in a Kyrgyz community:

> "All the staff and all the teachers we interviewed at Kazybek claimed that their school was the best in the raion (*region*), and that Russian as the instructional language was a primary reason for their success. Importantly, almost every school in our study, including Kazybek, gauged school success in terms of how many graduates went on to the universities in Bishkek (and secondarily to Naryn among schools in that *oblast*), as a result of the education they received" (p.5).

Toursunov (2010) concluded that the core issue is simply the failure of a corrupt, authoritarian regime to care for its citizens by providing quality education in their native language. In a series of interviews with parents and educational administrators, he found that parents send their children to Russian language schools simply because they believe the quality of education there is better. One respondent, a 30-year-old teacher and ethnic Uzbek from Osh whose daughter attends a Russian school, argued:

> "…migration is not the key issue. The main reason why parents take their children to Russian schools is that they offer a better education than Uzbek and Kyrgyz schools" (Toursunov, 2010).

And, from a sociologist in southern Kyrgyzstan:

> "Since Kyrgyzstan obtained independence after the collapse of the Soviet Union in 1991, the quality of services provided by secondary schools has been declining… the situation is most alarming at Uzbek and Kyrgyz language schools. Services are better at Russian schools, which attract more and more parents seeking better education for their children" (Toursunov, 2010).

Other explanations for the persistence of the Russian language are that TV and media available in Russian is seen as superior to local equivalents (Huskey, 1995; Korth, 2005). Increased labor emigration on the part of ethnic Kyrgyz to Russia (and hence the need for Russian language) has also been noted by some.[34] Finally, many post-Soviet Kyrgyz elites might simply still strongly identify with Russian culture and language as an inherent part of their own identity. Identity formation is a complex phenomenon and it is not necessarily the case that all Kyrgyz feel the need to be educated through the Kyrgyz medium in order to "feel Kyrgyz." There is evidence that many ethnic Kyrgyz identify strongly with Russian culture and language (Faranda & Nolle, 2010).

[34] According to some estimates, over 500,000 Kyrgyzstani citizens currently work in Russia (Podolskaya, 2011).

Whether it is ineffective governance, sensitivity to economic ties with Russia or the domestic European population, utilitarianism, the need to be seen as modern, cultural affinity, or simply the lack of motivation and interest on the part of highly "Russified elites" to address the issues, the evidence is clear that there are serious problems with the provision of quality education through the Kyrgyz medium. Education through the Russian language medium is still perceived as higher quality, both at the secondary and tertiary levels. In the next section of this chapter I turn to a discussion of tertiary education and the NST.

Tertiary Education and the NST

There were only two higher education institutions (HEIs) in the Kyrgyz Republic in 1932. By the early 1980s there were 10 with 57,109 students enrolled (Soktoev & Usubaliev, 1982). The total number of first year students enrolled in 1988 was 12,106 (National Statistical Committee of the USSR, 1989). Eight of the 10 HEIs in the republic were located in the capital at that time, Frunze. Soktoev & Usubaliev (1982) record 87 degree options in the 1980s and lists economics, engineering, pedagogy, medicine, and agronomy as popular specializations (majors). According to official statistics there were five general HEIs with humanities, pedagogy, and natural sciences programs - including one "university" - one medical academy, one institute of physical education, one arts academy, one agricultural institute, and one building and construction institute in the republic in 1988 (National Statistical Committee of the USSR, 1989).

The provision of tertiary education in the USSR was funded entirely by the state. Not only were the operating budgets and fixed capital provided by the state, but all students also received full state funding for the duration of their studies (Bereday, 1960). As in other sectors of the economy, centralized planning characterized all aspects of higher education provision: The

allocation of resources for the support of operations and facilities, academic programs and materials, the number of professorships and student places available, and even curricula were all determined according to state planning needs (Reeves, 2005). Upon completion of a course of study, graduates were typically assigned a job based on the needs of the relevant scientific, economic or social sector at the time of graduation. As a key provider of human resources for the state planned economy, HEIs did not have the institutional authority to enlarge their faculties or student bodies, significantly alter program offerings, create their own curricula, or make other major institutional decisions without direction from the central Ministry of Education (Bereday, 1960; Reeves, 2005).

Another characteristic of the Soviet higher education system was that HEIs were not all subordinate to the Ministry of Education. For example, the medical institute was under the Ministry of Health, the agricultural institute was under the Ministry of Agriculture, and the military and police academies were under the Ministry of Defense and Internal Affairs, respectively (Bereday, 1960). Higher education in the republic was (and still is) also distinguished from some western systems by the institutional separation of research and teaching. Scientific research is the responsibility of the Academy of Sciences, not HEIs. Another distinction is that academic programs are characterized by the high number of contact hours compared to their western peers. In some courses of study, students are in the lecture halls as much as 35-40 hours per week (Reeves, 2005). Once enrolled, students do not select classes and elective options but follow a prescribed course of study. As in secondary education, they move through their courses in cohorts (groups) which attend all classes together throughout their years of study.

With independence, most HEIs opened Kyrgyz language tracks for degree courses which had previously been taught only in Russian (Korth, 2005). While there are now both Russian and Kyrgyz groups for most fields of study, it is widely considered that for fields like medicine and the sciences the Russian groups still have better access to quality materials and teachers. Today, 67% of students overall continue to receive their tertiary educations through the Russian language medium (Bruner & Tillet, 2007).

Higher education in the republic has been dramatically affected by the collapse of the USSR. Educators have struggled to define the mission of higher education which had been so tightly coupled with state planning in the past. Today many remain proud of the Soviet era accomplishments in science and research and there is little agreement on whether reorienting the purpose of higher education away from "the needs of the state" to the needs of the market or the individual is either needed or appropriate (Reeves, 2005).[35] The biggest change however, has been the decline in state financial support and its impact on the higher education system. According to data provided by Bruner & Tillet (2007), by 1994 only 61.2% of all funding for higher education was from the national budget while by 2005 it had declined further to 30.4%.[36] These are tremendous decreases considering that just a few years ago 100% of all HEI funding had come from the state.

[35] This can often be seen in the contradictions between stated intentions and actual policy implementation. Rhetoric about the market aside, in regard to the way the budget places in HEIs are assigned, individuals do not select how to use their scholarships but instead choose from places available according to the needs of the state: Over half of all scholarship places are for teaching positions; i.e., the budget places are used to provide incentives to fill positions the state has prioritized (Silova, 2009).

[36] Anecdotally, depending on the institution, some university rectors would argue that today the actual state funding levels are around 10% but it depends on how "budget funding" is defined, i.e. figures vary depending on whether or not the value of buildings and other fixed capital inherited from the USSR is included in estimation. The main point is the dramatic decrease from the Soviet era.

Despite this decline in support, or perhaps because of it, the number of HEIs in post Soviet Kyrgyzstan has grown. Private institutions as well as institutions partly sponsored by foreign governments are now prevalent in the republic.[37] The founding of regional state universities in Naryn, Talas, Djalal-Abad, Kara-Kol, and Batken is also a post-Soviet development. In the Soviet Era, the only non-capital institutions of note were Osh State University and the pedagogical institute in Kara-Kol (Soktoev & Usubaliev, 1982). With the 1992 *Law on Education*, HEIs now have the discretion to collect tuition and fees from students and engage in other revenue generating activities. In some cases, entrepreneurial rectors have created "for profit" departments or institutes within state institutions as a way to generate resources though the quality of many of the new programs has been questioned (Bruner & Tillet, 2007).[38]

Bruner and Tillet (2007) report that in 2005 the number of HEIs in the republic was 49. This figure includes both the new state HEIs as well as many smaller, private institutions. Many of these institutions receive no state funding to support budget (scholarship) students however, and therefore are not obligated to accept NST results for admissions. Each HEI negotiates with the Ministry of Education the exact number of scholarship places it makes available every year. In 2010, 21 state-funded institutions enrolled budget students according to NST results

[37] Kyrgyz Russian Slavonic University (KRSU), The American University in Central Asian (AUCA), and Kyrgyz-Turkish Manas University are three of the most popular HEIs in the republic today. All three are partially funded by partnering countries.

[38] I have argued elsewhere that power relations between HEIs and the Ministry of Education have also changed (Drummond, 2011). While formally the ministry has retained many of their Soviet era oversight prerogatives, in reality, the funds generated by HEIs themselves have empowered them relative to other state institutions.

(CEATM[a], 2010).[39] Of course, along with the increase in institutions there has been an increase in overall student enrollment, a subject to which I turn in the next section after a brief review of Soviet HEI selection policy.

Student Selection in the Soviet Period

Though overall HEI admissions policy was made at the ministerial level in the Soviet period, each institution selected *abiturients* with internally created and administered examinations.[40] Examination scores served as the single criterion for selection for the majority of *abiturients*: The exception being special conditions for "gold medal" winners (perfect marks throughout the school career), winners of academic "Olympiads" and quotas for disabled, orphaned, or other special categories of students granted special admissions privileges. HEIs administered oral examinations in subjects deemed necessary for a particular course of study plus a written essay in the Russian language for humanities majors. Mathematics also required a written exam in addition to an oral exam (Clark, 2005). School transcripts, interviews, portfolios, *abiturients*' community or social activities and other criteria were not utilized as selection criteria.

---

[39] Private institutions are not required to take scholarship students. Foreign sponsored organizations are also not required but some have complicated admissions arrangements. For example, KRSU has over 500 scholarship "budget places" provided by the Russian Federation and only around 120–150 provided through the Kyrgyz republican budget. This means that admissions requirements are different depending on which budget supports the particular student. For the Russian Federation funded budget places, NST results are not considered in admissions decisions. The American University in Central Asia also has its own admissions requirements though it has traditionally provided considerable scholarship support for high scorers on the NST.

[40] The term abiturient most likely entered Russian from German (in the German system, an abitur has completed the type of secondary education that allows a pupil to apply to a university). In Russian, the term is commonly used to denote an HEI applicant, or entrant (лицо, поступающее в учебное заведение). Of course, there is also the assumption that applicants have completed secondary education necessary to apply to an HEI, i.e. an abiturient is no longer a pupil, but not yet a student.

Abiturients prepared for admissions exams by studying many "topics" or questions about a particular theme or subject area relevant to their desired major. At the appointed time in July of each year, *abiturients* then went to each HEI to which they were applying to sit for examinations. *Abiturients* came before an examination committee and randomly selected one or more of these topics, turned face down, on small cards or strips of paper (Drummond & De Young, 2004). After *abiturients* demonstrated their knowledge of the selected topic, admissions committees composed of specialists in the subject area being assessed asked the *abiturients* questions relevant to the topic. The examiners assigned marks on a scale of two to five, five being the highest mark. After the completion of examinations, the admissions committees forwarded their lists of recommended *abiturients* up the institution's chain of command for official approval and eventual enrollment (Drummond & De Young, 2004).[41]

After the breakup of the Soviet Union some HEIs continued to select *abiturients* through these procedures. However, with the loosening of bureaucratic controls, many institutions throughout Eurasia began to introduce written, multiple-choice tests. Representatives of some HEIs believed that multiple choice testing was more efficient in handling the increasing numbers of *abiturients* to higher education. According to one recent analysis, the total HEI enrollment in the 19-24 year old cohort went from 14% to a 36% in Kyrgyzstan from 1989 to 2001 (Bruner & Tillet 2007). Using multiple-choice tests, HEIs could screen larger numbers of *abiturients* than with the more time consuming individual based, oral exams (Drummond & De Young, 2004). Some perhaps saw such testing as representing a more "modern" approach to student selection (Valyaeva, 2006). Others however, saw standardized testing as a threat to their own educational heritage and as unwanted "Americanization" of their education system (Reeves, 2005).

[41] Ministerial approval of lists of recommended abiturients was a formality but necessary in order to enable state funding for the abiturients selected.

The new HEI-administered tests in the Kyrgyz Republic were not standardized in the sense that one test was utilized to assess all *abiturients* throughout the country; each HEI required *abiturients* to sit for their own tests. Critics pointed out that the multiple-choice tests administered by the universities were of low quality and could be easily manipulated by test administrators for *abiturients* willing to pay for the service (Clark, 2005). In the era of instability that followed the Soviet collapse, throughout Eurasia evidence that HEIs began to abuse their power in the admissions process through both oral examinations and the new multiple-choice tests mounted (International Crisis Group, 2003; Osipian, 2007; Heyneman et al., 2008).

By 2000, ministerial bureaucrats and even some university officials in Kyrgyzstan and other Eurasian countries began to propose major changes to their admissions systems (Valkova, 2001; Valyaeva, 2006; Osipian, 2007). While the timetable for reform was different in each country, the focus on corruption in selection was a common rationale for policy change across Eurasia. According to a report by the Russian Ministry of Education and Science and the Moscow School of Economics, *abiturients* in Russia were allegedly paying the equivalent of several years of tuition in illicit payments to enter higher education (Clark, 2005). At a September 19, 2006, address to participants at a conference on university admissions and examinations, the Minister of Education of the Georgian Republic, Alexander Lomaia, claimed as many as 80% of students admitted to Georgian HEIs in late 1990s, were enrolled for non-academic reasons (Conference Program, 2006).

Fighting corruption emerged as the primary rationale for the move to standardized testing in Kyrgyzstan as well (Drummond & De Young, 2004). President Askar Akaev issued his first decree in support of admissions reform on April 18, 2002. The decree called for the introduction

of the National Scholarship Test (NST) in June of that same year.[42] The decree explicitly eliminated all HEI discretion in selecting budget students: All scholarship (budget) places were to be allocated strictly according to test results (Presidential Decree No. 91, 2002). The Presidential Decree also called for public observation of the enrollment process, similar to the kind of monitoring that accompanies major political elections.

The NST is a high stakes selection test used for the distribution of over five thousand full university scholarships. The purpose of the National Scholarship Test (NST) is to determine which examinees have the scholastic aptitude and academic skills for study at the tertiary level (Valkova, 2004). The NST has mathematical and verbal reasoning domains.[43] Table 2-6 below highlights the differences between the Soviet examination system and the new NST in the Kyrgyz Republic.

Table 2-6: Soviet and Contemporary Selection Procedures

Country	Administered	Oversight	Purpose	Format
USSR	HEI-administered	Ministry of Education	Selection	Oral Exam, Subject Based Achievement
Kyrgyzstan (2002)	Non-governmental organization	Board of Trustees	Selection	Multiple Choice Test, Scholastic Aptitude

Presidential Decree (2002), Drummond & De Young (2004)

Acrimonious struggles over which institutions should have the discretion to conduct the NST and select students have accompanied the NST reform since inception. HEIs initially

[42] In the Russian language, the NST is known as "Общереспубликанское тестирование" which translates literally as "General Republican Testing." However, I use the name "National Scholarship Testing" as it captures the idea that results are used for HEI scholarship allocation.

[43] Additional subject tests (scored separately) are required for examinees seeking certain academic majors such as medicine and foreign languages (Valkova, 2004).

strongly resisted the introduction of the NST in 2002. In 2003 and 2004, opposition to the NST came from the Ministry of Education itself which sought to usurp the right to test from the non-governmental *Center for Educational Assessment and Teaching Methods* (CEATM) (Drummond, 2011). In 2002, the Minister of Education and Culture, Camilla Sharshekeeva, had insisted that the new testing center be a non-governmental organization, overseen by a board of trustees, not the ministry (Drummond & De Young, 2004). Minister of Education Ishengul Boljurova, who replaced Sharshekeeva in June of 2002, initially supported the new test center (Mambetaliev, 2003; Boljurova, 2003; Drummond & De Young, 2004). However, at a White House presentation for university rectors on November 24th, 2003, the minister articulated a new vision for admissions testing starting in 2004. That new plan entailed the Ministry of Education owning the rights to the student testing databases, ministerial oversight of test scoring and the ministerial production of test score certificates (Drummond, 2011).

The politics of NST implementation have been addressed elsewhere and will not be analyzed in detail here. However, it is important to highlight the fact that while HEIs are essential stakeholders in the NST, their representatives played (and continue to play) no formal role in deciding admissions policy for budget students: Neither in terms of determining the selection criteria nor in terms of how the enrollment process is organized.[44] Due to the publics' loss of trust in HEIs in the 1990s, they were cut out of the policy making circle on university

[44] Note that the selection reform was implemented with a Presidential, not ministerial decree. That is, while the Ministry of Education has formal power to make selections policy, major decisions need to be backed by the President of the Republic. As I have argued elsewhere, HEIs are in fact quite powerful in relation to the Ministry of Education (Drummond, 2011).

admissions both by Sharshekeeva and the ministers who followed her since 2002 (Drummond, 2011).[45]

The introduction of the standardized NST and elimination of HEI administered oral examinations represents the reassertion of administrative control by the Ministry of Education over HEIs which were perceived to be out of control in terms of selection corruption. However, this renewed ministerial oversight is not indicative of total control. As noted above, the majority of students now enrolled in higher education are those who pay for their educations, so-called "contract students." The majority of contract students are still selected primarily according to HEI examinations.[46] Thus, the impact of the NST on the students and the HEIs it is designed to assist has been different for different student cohorts. On the one hand, the NST is a high stakes test in that there are no other criteria utilized for scholarship distribution for full state scholarships. On the other hand, scholarships to study academic majors designated by the state are not necessarily that popular and there are a myriad of low cost, easily accessible opportunities for some kind of higher education in the event that NST results for a given *abiturient* are low (Bruner & Tillet, 2007).

In general, even with the decline in funding, the number of students in the post-Soviet era has sky-rocketed since independence. In 1992, there were only 53,670 total students enrolled in higher education in the republic, almost 100% of which were budget students on a full

[45] However, there is evidence that the NST is meeting HEI needs in terms of student selection. A predictive validity study of the NST demonstrated reasonable correlations between NST scores and academic achievement of students at the completion of one year of course work (Davidson, 2003).

[46] However, in 2010, the MOE required HEIs to accept 50% of their contract students based on NST results.

government scholarship.[47] By 1998, 120,986 students were enrolled, but only 27.5% received full state support (Bruner & Tillet, 2007). While higher education was free in the Soviet era, access to higher education was competitive throughout the USSR. At the end of the Soviet period, there were approximately three applicants for every available place in the Kyrgyz Republic (National Statistical Committee of the USSR, 1989). Considering that up to fifty percent of the entire age cohort left school after the eighth form, this three people for one place ratio is competitive. Today, while many students must pay, there are more access points to higher education available than ever before. [48]

Since 2002, approximately 5,200 to 5,700 full state scholarships have been allocated annually based on NST results. During this same period, the average size of the cohort graduating from secondary school has been between 72,000-82,000 pupils per year. Approximately 40-48% of the graduating secondary cohort (30,000-36,000) sits for the NST each year (CEATM[a], 2010). The enrollment data indicate that most participants go on to some form of higher education whether or not they win a scholarship place (Bruner & Tillet, 2007). It is estimated that another 10,000-15,000 enroll in some form of correspondence education, putting annual matriculation at approximately 40,000-50,000 and total number of students at over 200,000 in the entire system any given time (Bruner & Tillet, 2007). It can be deduced from these figures that only around 10% of all entering students currently receive full scholarships.

[47] This number includes both "day students" and "zaochnoye" (correspondence students) which are about 20% of the total student population.

[48] Anecdotally, many of the new institutions are popularly perceived to be little more than "business ventures" (providing low quality education) though it should also be noted that three of the most popular HEIs in the republic were also all founded in the 1990s, all with outside support (KRSU, AUCA, and Turkish Manas).

While there are more opportunities for higher education today, there are of course few HEIs with "elite status." An indicator utilized by the public to assess the prestige of HEIs is the annual NST report which contains the average NST scores of the entering scholarship classes. A full list of those institutions enrolling scholarship students can be found as Appendix F along with the NST average scores and number of budget entrants for each of these state HEIs. Note the wide dispersion in average scores across HEIs in Appendix F. The average 2010 scores for those entering with scholarship support at the prestigious Kyrgyz Russian Slavonic University and the Kyrgyz-Turkish Manas University were 182.2 and 182.1 (about two standard deviations above the mean) respectively. The Medical Academy average was also high at 177.7. At the same time, regional HEIs such as Talas State and Naryn State awarded scholarships to abiturients whose NST scores averaged just 116.4 and 115.4, respectively, barely above the NST average score of 113 for the nation as a whole. One can conclude the competition for budget places at elite HEIs is fierce (average score of entering cohort more than two standard deviations above the national average) while for "middle of the road institutions," hardly competitive at all.

One reason for the low competitiveness of places at the middle and lower tier HEIs is related to the purpose of the scholarships. An issue that is not visible in the data presented in Appendix F is what kind of budget opportunities are offered (by department). Recall that the scholarship does not "follow the student" but rather the student "follows the scholarship." The ministry has traditionally allocated approximately half of all budget places for "pedagogical faculties" and other specializations needed by the state, which are less popular than subjects like economics, international relations, and computer science (Drummond & De Young, 2004; Silova, 2009). The result is that many high NST scorers do not take part in the scholarship competition and prefer to pay for their educations and study the major of their choice. In a study

of 2007 NST results by Silova (2009), she found that the dispersion of NST average scores by faculty is as great as the geographical divide noted above. That is, those enrolling in areas of study like international relations had considerably higher NST scores than those enrolling as pedagogy majors.

The National Scholarship Test and Language Politics

Despite the vast performance gaps by language of instruction on the NST, the introduction of the NST has been relatively non-controversial, even popular among rural and non-Russian speaking cohorts.[49] There are several reasons for this. First, despite some initial resistance from elite HEIs, all examinees are allowed to sit for the NST in the language of their choice, regardless of the language on instruction in the HEI department to which they are applying (Drummond & De Young, 2004). This policy was introduced in 2002 in order to ensure that the brightest rural students were not denied educational opportunities at elite institutions due to a lack of language knowledge. More specifically, so that graduates of Kyrgyz language schools in rural regions could not be denied access to elite universities like the Kyrgyz-Russian Slavonic University because they didn't speak Russian. Minister Sharshekeeva argued that if examinees could score high enough on the NST in their native language, they were capable of learning Russian in a year of pre-enrollment language preparation (Drummond & De Young, 2004).

[49] In the winter of 2003 and spring of 2004, over 900 school directors were surveyed on their attitudes towards the new selection system (American Councils for International Educationb, 2004). The overwhelming majority of school directors favored independent testing, with only 5.6 % noting that universities should conduct selection testing. According to survey results, school directors believed that the motivation to learn had increased among pupils due to the introduction of the NST.

Perhaps the primary explanation for the continued support of the NST might have more to do with the fact that rural, Kyrgyz speaking students are winning scholarships in equal proportion to their Russian language counterparts, despite their overall lower average NST scores. This is because scholarships are awarded according to a quota system that places examinees in competition only with those from within the same quota category. Bishkek *abiturients* compete only against *abiturients* from Bishkek, not rural regions, and rural *abiturients* compete only against other rural *abiturients* for scholarship places. Each *abiturient* is assigned one of four possible demographic categories depending on the location of the school from which they graduated. Each village, town, or city has its own official designation. The purpose of the quota system was (and is) to "level the playing field" between rural and urban examinees (American Councils[a], 2004).

The result of this quota system is close proportional representation from each of the demographic and language categories in the overall proportion of scholarships awarded. This proportional representation persists, despite the fact that urban and Russian track examinees score almost a full standard deviation above the other groups on the NST. As can be seen from Appendix E for example, 66% of 2010 total scholarship winners were from Kyrgyz language tracks while these tracks represented only 60% of the total test takers. Note that the average score of this group was 125.6 while the average for the Russian language track examinees was 153.9.

In fact, the two most rural and impoverished quota categories - "village" and "high mountain" - are actually over-represented in the proportion of scholarships received (Table 2-7). While the quality of higher education varies between regions of the republic, it appears that not only are "village" and "high mountain" winners well represented in scholarship winnings overall,

they are also well represented in urban institutions.[50] The trend between participation and winnings is fairly consistent throughout each *oblast* and it is likely that without the quota system, this proportional representation in winnings would not be occurring.

Table 2-7: Scholarship Winners by Quota Category (2010)

	Republic	Bishkek	Towns	Village	Mountain
% Participation	100.0%	21.0%	14.7%	49.8%	14.5%
% Scholarships	14.8%	14.6%	13.5%	52.4%	19.5%
Avg. Score	113.5	135.4	122.7	104.4	102.2
Avg. Scholarship Score	134.9	158.1	146.4	127.9	125.6

(CEATM[a], 2010)

[50] See CEATM's 2010 Annual NST report for the demographic breakdown of scholarship winners at each urban university. www.testing.kg.

Chapter 3: Literature Review

Since the 1960s, a considerable amount of both applied and theoretical research on DIF, bias, and item equivalence has been conducted (Holland & Wainer, 1993; Camilli & Shephard, 1994). Studies have analyzed for racial, gender, and language differences on a variety of assessments and tests. DIF studies vary in purpose: Some are designed to assist practitioners identify and interpret causes of DIF while others compare the efficacy of DIF detection methods. In regard to statistical DIF, there has been considerable comparative research on various item response theory models, Mantel-Hanszel chi-squared and logistic regression methods (Clauser & Mazor, 1998). Much of the statistical DIF detection research has utilized simulated data sets in order to create experimental conditions for testing hypotheses (Hambleton, Clauser, Mazor, & Jones, 1993).

Most DIF studies conducted in the USA have focused on racial or gender differences (Holland & Wainer, 1993). However, there is a growing literature on DIF in cross-lingual assessments (Hambelton, 2005). This literature review focuses on the DIF research most relevant to this study: Studies of item reviewers' ability to predict DIF through substantive review and studies of causes of DIF on cross-lingual, verbal assessment items. The last section is an analysis of the literature that addresses how the particular statistical methods employed to detect DIF can impact DIF detection results.

Substantive Review and DIF Prediction

Studies of the relationship between substantive review and statistical DIF detection methods have been conducted in various contexts for some time (Mazor, 1993). Some early analyses in the USA focused on racial, gender, or group other differences (Plake, 1980; Engelhard, Hansche & Rutledge, 1990). More recently, Gierl & Khaliq (2001) have conducted

research on Canadian achievement tests in which the relationship between substantive and statistical methods was assessed. However, the overall number of studies is quite small and there have been very few studies, if any, on congruence of these methods on cross-lingual assessments in developing country contexts.[51]

Cross-lingual substantive analyses often employ *post-hoc* review in which linguists, translators and content specialists analyze items flagged as DIF by statistical procedures (Joldersma, 2008). However, as in this study, it is also possible to work in the opposite direction and collect substantive data first, then statistically analyze the items in order to understand how well item reviewers predict DIF. Various protocols, coding guides and rubrics, questionnaires and focus groups have all been employed to collect and analyze such data (Engelhard, Hansche, & Rutledge, 1990; Allalouf, Hambleton & Sireci, 1999; Gierl & Khaliq, 2001; Ercikan, 2002).

Early studies of substantive reviews reflect the socio-political issues important in those times (Holland & Wainer, 1993). In the 1970s and 1980s, analyses typically focused on whether minority groups and women were represented in a positive light on educational and professional tests, whether they were represented at all, and whether the content presented in tests and items would be equally familiar to all examinees across groups (Tittle, 1982). In the early literature, the term "bias" was often used in a broader way than is currently accepted in the psychometric literature. Bias was sometimes used in reference to any test with "poor representation" of minority groups or for test items that appeared to place women or minorities in stereotypical

[51] Cross-lingual testing is increasingly entering the domain of US policy makers. In particular, members of the Obama administration often reference the US's "poor performance" on such cross-lingual assessments as the Programme for International Student Assessment (PISA) and the Trends in Mathematics and Science (TIMSS) assessment programs. See 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.

roles. Much early writing about substantive review also had a highly prescriptive character with "how to" type recommendations for test developers and reviewers (Holland & Wainer, 1993).

In 1982, Carol Kehr Tittle presented a comprehensive plan for how substantive reviews could be employed at all stages in the test development process – planning, specifications development, item try outs, post-test review, etc. She provided recommendations and detailed rubrics for scoring and collating problematic items. Other work from this period also has a highly prescriptive character for how to ensure item fairness. According to Scheuneman (1982), Coffman (1961), Donlan (1971) and Dwyer (1979) conducted studies of the relationship between substantive reviews and statistical outcomes, but all three focused on gender differences on the *Scholastic Aptitude Test* (SAT). Medley and Quirk (1974) studied black-white differences on the *National Teacher Examination*. Such studies were viewed as important not only for political purposes but because the computational costs of statistical analyses were high at that time (Plake, 1980; Holland & Wainer, 1993).

Further, as Plake (1980) argued, in many testing situations the number of examinees was sometimes too low to conduct statistical analyses, even when technology was available. Thus, it was considered important to find ways to improve the quality of the substantive evaluation process as test developers did not always have the luxury of statistical DIF detection methods. In cross-lingual testing today, high profile cross-lingual assessments like Trends in Mathematics and Science (TIMSS) and the Programme for International Student Assessment (PISA) rely on sophisticated, quantitative methods for item analysis. However, not all multi-lingual countries have the financial and personnel resources to conduct such sophisticated DIF detection techniques. Thus, in some ways, many developing countries still face the same challenges to DIF detection that many western analysts and researchers faced in the 1960s and 1970s.

The results of many of the early studies on reviewers' ability to predict DIF were mixed at best. Tittle (1982) found that overall, the outcomes were inconsistent, with results highly dependent on methods employed, type of prediction study, and expertise and background of item reviewers. A decade later, Mazor (1993) stated more conclusively that the cumulative result of the early research was that accurate DIF prediction by substantive review was the exception rather than the rule. In order to understand the challenges involved in substantive DIF review it is necessary to present several of these studies in greater detail. In the next section I present findings from some of the more well known studies of the efficacy of substantive review.

Using data from the *Iowa Test of Basic Skills,* Plake (1980) analyzed whether raters could identify DIF for students from the 4^{th} through 8^{th} grades who all had 5^{th} grade skill levels in mathematical concepts. In order to control for ability level as a confounding factor, she paired examinees with like ability from 5^{th} grade with similar ability from the other grades and created separate test groups. Three specialists in elementary math education then predicted which items would be easier or harder for non-5^{th} graders. When two out of three specialists selected an item, it was deemed a DIF item. Plake utilized ANOVA to analyze the test results and compared these to her panel results. The result was that raters predicted twice the amount of DIF than the statistical procedures yielded. In terms of direction of DIF (which group was favored) one third of the items favored the opposite direction that was predicted by the specialist raters. The raters also differed greatly in the number of DIF items they identified at 41, 38, and 16 cases.

Engelhard, Hansche & Rutledge (1990*)* analyzed the ability of item raters to predict DIF between blacks and whites on a series of three different teacher education examinations. Forty-two judges examined 40 test items from teacher certification test batteries. Twenty four evaluators were black and 16 were white. The judges were divided into three separate review

committees - one for early childhood, one for administration and supervision, and one for middle childhood examinations. All participants in the study were experienced members of previous bias review committees. They received 45 minutes of training and written guidelines for identifying potential problems. They categorized items as "favors blacks, no difference, or favors whites."

From the results of the review, the researchers created a categorical index called the "Judged Category Index" with categories coded as -1 = favor blacks, 0 = no difference, 1 = favor whites. They then compared the results from this index with results from a statistical DIF detection method, the Mantel-Haenszel (MH) chi-square method, a method commonly used by the ETS (Camilli & Shephard, 1994). The MH procedure tests whether the odds of success on a given item are proportional for both groups across levels of the matching criteria (ability). The null hypothesis is that the proportion of examinees answering correctly in the reference group is the same as the proportion for the focal group. The MH method employs a 2 X 2 contingency table for each item where item response data (correct/incorrect) is entered along with group membership for those examinees with the same ability (Mazor, 1993).[52]

Engelhard et al. (1990*)* then computed the correlation between the substantive estimates and empirical estimates that were calculated using MH. The result was little agreement between the two estimation methods. The correlations ranged from .00 to .11 for the three different tests. However, they did find significant individual differences between reviewers with one having a

[52] Swaminathan & Rogers (1990) argue that the MH procedure is best conceived as a special case of the logistic regression method (LR). The main difference is that MH treats ability as a discrete variable while in LR ability is treated as continuous. They note that having the variable treated as continuous enables analysis of an interaction effect between ability and group. Mazor et. al (1992) argue that this is important because in the MH model, if an item favors one group at one end of the ability distribution but the other group at the other end, key information gets canceled out and no DIF is reported (Mazor, Clauser & Hambleton, 1992).

.52 correlation to the statistical results. At the same time, another one of the reviewers in the administration and supervision group had a negative correlation of -.36.

The two main results of this study were that (1) there was significant variation in the ability of the item raters to accurately detect differences and (2) as a group- raters were not able to predict DIF very well. From each of the three analyzed groups of items, only one or two evaluators (from 42) demonstrated better than chance agreement with statistical data. Engelhard et al. (1990*)* concluded that item reviewers could not predict which test items would perform differently for black and white examinees when they had no empirical data. They argued that a primary reason for low agreement between the two indices was the infrequent use of the category "favors blacks." They proposed that because many reviewers were asked to represent the interest of their social category (race) in a high stakes situation, this might have influenced their estimations.

Another conclusion they drew from this study was the need to conduct experimental research (using simulated data) which would allow them to compare how well reviewers could identify flaws with test items. The authors argued that the practical utility of an experimental study would be useful in selecting quality reviewers for review committees. Engelhard, Davis, and Hansche (1999) conducted such an experimental study with thirty-nine reviewers on a state-wide student assessment program in the state of Georgia. The reviewers were practicing elementary teachers and administrators, were diverse in age and all had experience either writing test items or participating in bias review committees.

Before beginning, the evaluators received a sixty minute training session which included the overall purpose of the assessment system and guidelines for identifying flaws. The key to this study was that some of the over seventy test items had known flaws which served as the

criteria (a baseline) with which to assess the accuracy of the judges' ratings. Test items came from a variety of content from grades three through eight. Twenty-eight of the items had no known flaws while 47 items had flaws; nineteen had one flaw, 22 had two flaws, 5 had three flaws and 1 item had four flaws. The flaws were broken down into cultural flaws and technical flaws. After reviewing the items, the reviewers responded to 16 questions. The questions on cultural flaws had to do with gender, race, handicaps, socio-economic status, demographics (rural vs. urban) etc. For the technical flaw category, reviewers answered questions about the comparability of the difficulty levels in format, language, prior knowledge, grammar, typographical errors, item content, appropriateness of topic, etc.

Each reviewer then spent two to three hours evaluating the 75 items. Reviewers marked "yes" if they believed the items exhibited any flaws. They left the questions blank if they found no flaws. The accuracy of their estimations was determined by the agreement of their marks and the predetermined *a priori* classification of the items. They utilized a logistic transformation of ratios to determine the probability of accuracy vs. inaccuracy of evaluators' predictions. False positives and false negatives were scored as inaccurate. The most accurate reviewer was 94% accurate while the least accurate was 83% accurate. Overall, accuracy rates were higher on the cultural flaws than technical flaws. The study demonstrated that substantive committees could be quite accurate in detecting various item flaws. However, as the authors noted, identifying flaws is not the same thing as predicting DIF.

In recent cross-lingual studies of DIF prediction, bi-lingual reviewers have been more successful than in Engelhard et al.'s 1990 study. Gierl and Khaliq (2001) conducted a study with eleven reviewers analyzing French and English social studies and mathematics tests at the 6^{th} and 9^{th} grade levels. Their method included having the reviewers generate *a priori* hypotheses

about types of DIF and which groups might be favored. Cognizant of the potential for multi-dimensionality (addressed below) evaluators attempted to discern not only primary traits assessed by the items, but also what secondary traits these items might be assessing and how these traits might impact the two groups differentially. It was a matter of judgment as to whether the secondary dimension was benign or adverse. Items with similar characteristics were organized into "bundles" for analysis. They utilized the Simultaneous Test for Bias (SIBTEST) to test for statistical DIF.

Across both grade levels, the evaluators predicted the direction of DIF correctly in 7 of 8 times for the mathematics items and 8 of 13 for the social studies items. Intuitively, the results of Gierl and Khaliq's (2001) are plausible as differences between languages may at times be more explicit and somewhat easier to detect than predicting how racial groups will respond to items that are in the same language. Overt mistakes like poor translation or typographical errors might be easier (on average) to detect than say how females or males might react to different kinds of items in the same language. However, as Ercikan (2002) points out, the raters in Gierl and Khaliq's (2001) study also knew in advance which items had been flagged as DIF. Thus, they were not so much "predicting DIF" and DIF direction as "interpreting DIF direction" with *known* DIF items. They employed "a consensus-building model wherein the reviewers worked as a group and focused on standardizing interpretations and ratings across reviewers, which may have contributed to high success rates of explaining DIF" (Ercikan, 2002, p. 201). Thus, it would appear that method of DIF evaluation, whether DIF is known or not, and whether individual or group analyses are employed might also contribute to success rates.

In addition, Ercikan (2002) argues that it makes a difference whether a cross-lingual DIF study is on test items utilized within a single country study or across countries. This is because

the potential number of DIF sources is higher in cross-country studies. In cross-country analyses there is greater potential for variation in opportunities to learn or curricular coverage to cloud reviewers' estimations. Further, with within-country studies, there is a larger pool of potential reviewers with intimate linguistic knowledge and cultural understanding that may not always be available for the cross-country study. Languages, conditions and cultures in different language groups can be relatively well understood by within country bi-lingual reviewers who not only have life long experience with both languages, but in many senses may consider themselves to bi-cultural as well.

Item evaluators in the Kyrgyz Republic might also be expected to do relatively well in prediction in comparison to some other types of DIF studies. It is possible to find item reviewers who are themselves the products of both Russian language and Kyrgyz language educations and many have intimate knowledge of the cultural differences between the two groups. As the NST is an aptitude test, curricular differences are not expected as they might be with achievement tests (Ercikan, 2002). School teachers and other educators from both language groups are trained in the same institutions, sometimes with the same materials. School textbooks for both languages are often the same (translated from Russian to Kyrgyz) or at least have historically been so for the generation of evaluators participating in this study (De Young, Reeves, & Valyaeva, 2006).

On the other hand, accuracy in DIF detection and prediction might be more of a function of evaluators' expertise and experience. Gierl and Khaliq's (2001) study involved highly trained and experienced reviewers. Lack and training and experience in sophisticated evaluation techniques may present challenges to accurate DIF identification in some contexts. In the Kyrgyz Republic there are few (if any) specialists with the experience in undertaking such

analyses. Further, at this point, no comparative research has been conducted on test items produced in Russian and the Turkic languages.

Ercikan (2002) also contends that the results of substantive review depend on whether or not the reviewers know the DIF statistics before or during their analyses. When evaluators have knowledge that DIF has been indentified by statistics, some such "DIF cause" may always be found – whether accurate or not – which can lead to an inflated success rate. She also argues that it makes a difference as to whether items are evaluated individually or as item pairs (reviewed simultaneously). When both items are presented at the same time, evaluators tend to focus on the comparability of details like format, content, and language use. Reviewers of a single item focus more on context and content that might make the item biased for a particular group. Ercikan (2002) proposes that the single item review approach leads to a more nuanced analysis of item content and context and the consideration of different cognitive processes among comparison groups.

In the section that follows I first review various studies that have determined the amount of DIF on cross-lingual assessments. I then address studies that sought to determine the causes of bias or DIF on cross-lingual assessments. This review will set the context for the second research question in this study in regard to DIF sources.

Levels of DIF in Cross-Lingual Testing

Several cross-lingual DIF studies have reported large percentages of items as DIF. Gierl, Rogers and Klinger (1999) found that 52% of English–French item pairs on a Canadian elementary social studies test exhibited DIF. Ercikan and McCreith (2002) discovered DIF rates of 41% on TIMSS science items. 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%. They go on to say, "By any reasonable criterion for interpreting the delta-DIF statistics, the DIF results reveal major problems with the translation/adaptations with the Altaic versions of the exam" (Robin et al. 2005, p. 15).

On the verbal section of a university admissions exam in Israel, Russian-Hebrew DIF rates on the test were 34% (Allalouf, Hambleton & Sireci, 1999). On Programme for International Student Assessment (PISA) reading items, Grisay and Monseur (2007) found DIF rates of 25%-30% on European to European language comparisons but the rates increased to 45% when the items were from highly dissimilar language groups. Interpretation of any given DIF result in light of other DIF studies however, is not necessarily straight forward. Different studies use different criteria to define DIF levels. Therefore, determining how much a given DIF level threatens test comparability is not simply determined by percentages of DIF or DIF item counts as these figures mean different things in each study (Grisay & Monseur, 2007). For example, how one employs an effect size measure to distinguish between statistical DIF and practical DIF (or does not) impacts how one classifies items as DIF.

Grisay and Monseur (2007) evaluated PISA data from the 2000 reading assessment to determine item performance across various groups. Utilizing data from 47 countries, they analyzed 32 reading passages with a total 132 test items. They found that adapting a test from a source version always had at least a basic cost in terms of loss of equivalence. They found that using tests in the same language (but developed differently or in another location as in several Spanish language tests developed in each of the Spanish speaking PISA countries) is not as

valuable as using identical (twin) tests.[53] This is because any translated version is just one in an infinite number of potential "sister versions." Reckase and Kunce (2002) also found that different translators produce highly variable results in terms of accuracy and quality of translation.

In Grisay and Monseur's (2007) study DIF levels increased when comparisons were made to "cousin versions," or different language versions within same Indo-European family (German to English for example), with, on average, 25%-30% of items displaying DIF. However, the most fascinating finding was the comparison across language families. When examining the Indo-European and Asian language groups, the level of average DIF was around 45%. In other words, it was difficult to interpret whether or not about 45% of the total items were actually measuring the same way in say, the English and Japanese versions of the PISA reading section. Grisay and Monseur (2007) also found:

> "A highly positive correlation between communality and test reliability (.72), as well as the negative correlation between reliability and Asian country (-.70). This suggests that some non-random factor affecting the geographic or cultural distribution of DIF items was deteriorating, to some extent, the reliability of the scale in a number of countries" (p. 76).

Their study was not the first to show the lack of construct invariance across European and non European languages in DIF studies. A study by Grisay, de Jong, Gebhardt, Berezner, and Halleux (2006) with TIMSS data also found a high level of DIF between Indo-European languages and non-Indo-European languages.

[53] For example, the English version used in five countries was more or less the same test, slightly adjusted for local differences. Each version of the Spanish version however, was adapted from English and French in complete isolation from all other Spanish versions and these different Spanish versions were not compared with each other prior to test administration. The result was higher levels of DIF in the Spanish versions.

PISA's 2003 Technical Report also suggests that despite expertise and highly developed protocols for item adaptation, some versions have higher percentages of what they claim to be "weak items" than others; for example, 18% of the items for the Japanese test and up to 32% for the Arabic language, Tunisian version (OECD 2003, pp. 77-79). They note that one explanation may be the overall instability of the scale as these language groups tend to be located on either the upper or lower extremes of the scale. However, they also offer this potential explanation for the larger portion of weak items for the non-European languages:

> … a second possible explanation might be of some concern in terms of linguistic and cultural equivalence, i.e. the fact that the group of outliers included all but two of the ten PISA versions that were developed in non-Indo-European languages (Arabic, Turkish, Basque, Japanese, Korean, Thai, Chinese and Bahasa Indonesian)…and, finally, a third explanation may well be that competent translators and verifiers from English and French are simply harder to find in certain countries or for certain languages than for others (OECD 2003, p. 79).

Thus, an important finding of recent cross-lingual DIF studies is that DIF levels appear to vary depending on the relationship between the two language groups in question. That is, while there may be common challenges to all cross-lingual adaptation, not all languages "compare" across these commonalities in the same way. In particular, assessments involving languages from within the same "language family" tend to exhibit lower DIF levels than when assessments involve languages from more disparate language families (Grisay & Monseur, 2007). These are significant findings for this study as Russian and Kyrgyz come from very different language families, Slavic and Altaic (Turkic) (Oruzbaeva, 1997).

Causes of DIF in Cross-Lingual Testing

Several studies have focused on the causes or origins of DIF on cross-lingual assessments. Studying an intelligence test with German and English language examinees, Ellis (1995) concluded that most of the DIF was due to translation error. Van de Vijver and Poortinga

(2005) argue that the most significant sources of item bias in cross-lingual testing is poor test adaptation resulting from poor translation, careless work, lack of subject knowledge, or lack of understanding of the principles of test development. Hambleton (2005) lists five general sources of item bias – the test itself, selection and training of translators, the process of translation, poor protocols for adapting tests, and poor data collection designs and data analysis for establishing equivalence.

In Gierl and Khaliq's (2001) study with data from several content areas, his 11 member review committee found four sources of adaptation/translation DIF: (1) omissions or admissions of words that effect meaning, (2) differences in the words, expressions, or sentence structure that are inherent to the language and or culture, (3) differences in the words, expressions, or sentence structure of items that are not inherent to the language or culture, and (4) differences in item format. Several other studies have concluded that the issue of word difficulty (inability or failure to use words of equal difficulty) is a common cause of DIF. Schmidt and Belistein (1987), Bejar, Chaffin and Embertson (1991), and Roccase and Moshinsky (1997) all found word difficulty to be problematic.

However, not all DIF on cross-lingual assessments is caused by translator-related adaptation error. In Gierl, Rogers, and Klinger's (1999) study of French and English examinees, only 2 of 7 math items detected as DIF were found to contain translation errors after substantive review. Only 6 of 26 DIF items on that same test (social studies items) were found to have translation errors. Similarly, Ercikan and McCrieth (2002) found large levels of DIF on the TIMSS Science section but poor adaptation was the cause of only 22% of the mathematics items and 40% of the science items flagged as DIF.

Other hypotheses for explaining DIF causes have been proposed. Less visible psychological factors like different student response strategies used by examinees might also cause DIF (Gierl et al., 1999). By looking at the distribution of DIF items by curricular topic area, Ercikan, Gierl, McCreith, Phan, & Koh (2004) discovered that different opportunities to learn could lead to DIF. Even word count might be an important DIF issue. In the exam that Gierl et al. (1999) examined, there were 24% more words on the French exam than the English exam. They concluded that the longer test length might make the exam more difficult for the French examinees.

In regard to causes of DIF specifically on cross-lingual verbal assessments, Agnoff and Cook (1988) discovered that in some cases, additional text size is sometimes a good thing. They hypothesized that longer texts are sometimes necessary in one of the languages in order to provide enough context and sufficient explication of meaning. This is related to the idea that sometimes inherent linguistic differences can make some item types more conducive to item adaptation than others. They found greater DIF in antonym and analogy (shorter) items and less in sentence completion and reading comprehension. Beller (1995) and Gafni and Canaan-Yehishafat (1993) also found that DIF was greater in analogy items than sentence completion and reading passages.

Using data from the Israeli Psychometric Entrance Test (PET), Allalouf, Hambleton and Sireci (1999) examined the causes of DIF between Russian and Hebrew examinees on verbal test items. They concluded that analogies were problematic with 65% of items demonstrating DIF. Reading comprehension items showed a small amount of DIF. Through *post-hoc* substantive review, they found the primary causes of DIF to be: (1) Changes in difficulty of words or sentences – i.e., the translation was accurate and meaning relatively intact, but words or

sentences became easier or more difficult after adaptation for one of the languages. This could also be due to how literal or symbolic the meaning of questions is presented; (2) Changes in content – i.e., an item lost its meaning for one of the languages after adaptation; (3) Changes in format – i.e., an item became much longer, shorter or more awkward for one of the languages after adaptation; (4) Differences in cultural relevance – i.e., items contained meaning, symbols, norms, content, or expressions that had no equivalent connotation in the other language group. The findings from these four studies indicate that there are characteristics of item types like length of items that seem to a have similar impact on DIF levels across a range of language groups.

There are a myriad of explanations as to why it is difficult for evaluators to predict DIF: Lack of training and experience, poorly designed procedures and protocols, lack of time and resources to do the evaluations, personal dispositions or pressures for certain outcomes, and of course simply the difficult task of trying to anticipate how the background and psychological make-up of any given group will impact how they respond to any given set of items. Whatever the care and the methods employed, identifying the sources of DIF through substantive studies is a challenge. Mazor (1993) argues that the failure of substantive studies (on both real and simulated data) to consistently identify DIF also challenges some of the fundamental assumptions that researchers make in DIF studies. In the next section I turn to some of the most important of those assumptions.

Statistical DIF detection methods like the Mantel-Hanzsel (MH) and logistic regression (LR) often condition on total test scores as a proxy for examinee ability. The items that compose these scores are typically hypothesized to be tapping into a single trait or skill. However, for some time it has been known that test items are often not uni-dimensional. Items may be multi-

dimensional which means that they are measuring more than one latent trait (Reckase, 1985). As Ackerman (1992) noted in his definition of DIF, we should keep in mind that the single test score, typically used as the proxy for ability, is an *alleged* conditional ability. This is important to keep in mind for any DIF study, especially for those involving "real world items" like the NST in the Kyrgyz Republic (Kok, 1988).

Several early DIF studies demonstrated that the uni-dimensionality assumption of DIF detection methods was untenable (Birbaum & Tatsuoka, 1982; Subkoviak, Mack, Ironson, & Craig, 1984). A commonly cited example of an item with a high probability for multidimensionality is the mathematics word problem that demands considerable reading or verbal ability (a secondary trait) in addition to mathematics skills (primary trait) in order to solve the item correctly. The ramification for DIF studies is that if there are underlying differences between groups on secondary traits, DIF could actually be the caused by this multi-dimensionality (Kok, 1988). Interpreting DIF results can become ambiguous if these secondary traits are not identified and parceled out (Shealy & Stout, 1993).

Ackerman (1992) calls the primary ability the *target ability* and the secondary ability *nuisance* ability. Ackerman contends that all test items tap into at least some level of nuisance ability. He believes that small amounts of DIF are likely in conditions where a secondary trait is tapped and the distribution on that secondary trait across groups differs. At the same time, an item may be multidimensional but not DIF if the groups involved have equal distributions on all the traits assessed. Kok (1988) cites background knowledge, language skills and "test wiseness" as examples of secondary skills and knowledge upon which populations may differ. He notes that sometimes multidimensionality is unavoidable when tests employ complex items with the

intent to approximate situations in which many skills need to be applied simultaneously by an examinee.

Douglas, Roussos, and Stout (1996) make a useful distinction between types of secondary abilities: *Auxiliary* abilities are those that can be legitimately a part of the construct measured, while *nuisance* abilities are those not related to the construct of interest in any way. They thus conclude that DIF arising from auxiliary abilities is *benign* DIF, while DIF arising from nuisance ability is *adverse* DIF (Douglas et al., 1996). They note that in practice, in order to determine which kind of DIF is prevalent, substantive reviews *a priori* are needed to hypothesize which item bundles might exhibit multidimensionality.

The low correlations between statistical DIF and substantive review methods in some studies could be related to this unaccounted for multi-dimensionality because it is hard to identify. Assessing for multidimensionality on cross-lingual test items requires evaluators to know about more than just test adaptation processes and linguistic issues; knowledge of examinee exposure to a broad variety of content and their cognition as they engage with content is also important. It may be difficult in many instances to find reviewers who are both bi-lingual and equally knowledgeable about the nuances of item response.

One way to try and minimize the effect of multi-dimensionality in DIF estimation has been to condition on sub-scores rather than total scores during statistical DIF detection. Theoretically, this provides analysts with a cleaner estimate of the particular ability under study (Clauser, Mazor, & Hambleton, 1991; Ackerman, 1992; Mazor, 1993; Clauser, Nungester, Mazor, & Ripkey, 1996). For example, using the MH method on 91 items, Clauser, Mazor, and Hambleton (1991) examined a sample of 1,000 examinees from two subgroups (Anglo-Americans and Native Americans), with an average test score difference of about one standard

deviation. The test had four sub domains - mathematics, reading, prior reading, and charts. They discovered that choice of conditioning variable made a difference in the level of DIF identified. Twenty two items were identified as DIF when conditioned on total score. When they conditioned on a sub-scores alone, the amount of DIF identified fell by one third and reduced the overall type 1 error. At the same time, however, when then they conditioned only on sub-score, some items emerged as DIF that had not been previously indentified.

Mazor, Kanjee, and Clauser (1995) conducted a study on two achievement tests. They compared males and females but also took into consideration knowledge of English (those who reported it as their best language vs. others who reported some other language as their best). They used both logistic regression (LR) and the MH procedures and first conditioned on total score. Then, using LR they added the sub-scores SAT-verbal and SAT- math to their model. They found that with the LR procedure the number of items identified as DIF was reduced when conditioning on sub-scores.

Clauser, Nugester and Swaminathan (1996) employed a logistic regression model and conditioned on both a total score and educational experience (area of specialization of medical students) as a secondary variable. As men and women tend to generally select different areas of specialization (on average, more men in surgery and more women in pediatrics), the authors hypothesized that males and females may differ in ability distributions across background. They believed that the conditioning variable would reduce the number of flagged items as it would partially account for those differences in group performances on this secondary ability. When conditioning only on total score, 30% of items were identified as DIF. When the background variable was added, the number of DIF items was reduced to 19%. Although the main result was a reduction in the total number of items identified, some new DIF items were identified when

using the background variable that were not identified using the total test score alone. Nonetheless, these studies that address the multi-dimensionality issue have informed the design of this study and the methods that I present in the next chapter.

DIF as Statistical Artifact

A final and very important consideration in disentangling the sources of DIF is the extent to which the methods employed are themselves producing reliable DIF estimates. That is, it is possible that one reason why substantive evaluators sometimes can not identify DIF is because there may in fact be no DIF: Items that have been identified as DIF may simply be the result of statistical artifacts (Mazor, 1993; Gierl et al. 1999; Ercikan et al., 2004). For example, inflated type one error due to a poor choice of conditioning variable may muddle statistical outcomes. It can not be assumed that DIF levels indicated by a particular statistical method are infallible. In research conducted by Jodoin and Gierl (2001) power rates for a host of real world DIF conditions using LR methods were only 70-80%; thus the interpretation of DIF statistics needs to be made with caution.

Fortunately, there is a way to gain a general understanding of the effectiveness of various detection methods utilized in varying conditions. Much of what we know about the efficacy of DIF methods comes from simulation studies because simulations allow a comparison of the efficacy of different approaches under controlled conditions, i.e. we know "how much DIF actually exists *a priori*" (Hambleton et al., 1993). Through simulation studies, researchers can create DIF levels or other necessary experimental conditions by adjusting the difficulty and discrimination parameters of artificially generated item data. Thus, they can compare the effects of large and small sample size, variation in the ability distributions of examinees, item types, DIF

levels per test, test length and dimensionality among other factors (Hambleton et al., 1993; Narayanan & Swaminathan, 1996; Jodoin & Gierl, 2001).

In general, it should be noted that comparative research done on various methods consistently shows that IRT, MH, and LR methods are equally effective in the identification of uniform DIF (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Roussos & Stout, 1993; Narayanan & Swaminathan 1994). Nonetheless, there are differences between methods. While both the MH and LR consistently show similar results in their capacities to detect uniform DIF, the MH method has not been able to identify non-uniform DIF (Swaminathan & Rogers, 1990; Narayanan & Swaminathan, 1994; Hambleton et al., 1993). While non-uniform DIF is less common than uniform DIF, it does occur in practice. Thus, one possible advantage of logistic regression over other DIF methods is that it can assess interaction of group membership and examinee ability (Swaminathan & Rogers, 1990; Gierl et al., 1999). On the other hand, they also found that while type 1 error rates (identifying items as DIF when they were not) were within expected limits for the MH, they were a bit higher for the LR procedure.

The size of the examinee sample is also important. The converging evidence from DIF detection studies is that a larger sample sizes allow for more accurate DIF detection. In Rogers and Swaminathan's (1993) comparison of MH and LR methods, they discovered that the detection rates increased by 15% when the sample size was increased from 250 to 500 for both methods. In Mazor et al.'s (1995) study, various sample sizes were created from 100 to 2,000 per group. The study demonstrated that a small sample size (100) was not adequate but sizes of 200 to 1,000 were satisfactory. Hambleton et al.'s (1993) review of simulation studies also indicates that these findings about sample size are true across combinations of item types, ability distributions and other experimental conditions.

On the other hand, while it would appear that large sample sizes are necessary, there is evidence that type 1 error (over-identification of DIF) increases with larger sample sizes. Thus, one concern with overly large samples is that even the most trivial differences between groups can be identified as statistically significant even though they are of little practical significance. Hambleton (1989) argues that while small sample sizes fail to capture much DIF, with sample sizes around 5,000 it is conceivable that much of the DIF detected will be of no practical significance. Jodoin and Gierl (2001) have proposed that DIF detection methods using chi-squared tests must have a reliable measure of effect size.

Another factor that can impact DIF results is related to the ability distributions of the two groups under study. In cross-lingual DIF studies ability distributions are often not the same: In fact, gaps may be quite large which is why several simulation studies have created experimental conditions in which the groups tested differ by as much as one standard deviation. Narayanan & Swaminathan (1996) found that DIF detection rates were higher when examinees were sampled from the equal ability groups for the MH, SIBTEST, and LR methods. The differences in detection rates dropped when two differing ability distributions were analyzed but not equally across all methods. The biggest drop was 14% for the LR method. For all three procedures, the type 1 error rates were higher for the unequal ability distribution than those for the equal ability distributions. At the .05 level, they were 4.1% for MH and 6.1% for LR. At unequal, they were 5.5% for the MH and 9% for the LR method (Narayanan & Swaminathan, 1996).

Simulation studies also tend to agree that item characteristics can impact DIF results. In Rogers and Swaminathan's (1993) study, the items with DIF that were most easily detected by both the LR and MH procedures were items of moderate difficulty and high discrimination. For these items, detection rates were as much as 15% greater than for other types of items (Rogers &

Swaminathan 1993). Hambleton et al.'s (1993) study also indicated that items with lower discrimination were associated with items that were likely to be missed with MH, regardless of differences in difficulty. They also found that very difficult items were more likely to be missed in DIF detection methods, regardless of ability level.

The statistical issues raised through the literature review above are directly relevant to this DIF study in the Kyrgyz Republic. Sample size is not likely to be a problem. However, the difference in ability distributions between the two groups under study is large and the item characteristics for the Russian and the Kyrgyz items do differ. In Chapter 6 of the study, I will return to these issues and discuss them in relation to the findings of this study. I now turn to a presentation of the study's methods, Chapter 4.

Chapter 4: Methods

Multiple research methods were employed in this study. Before reviewing each of them, I first introduce examples of the item types analyzed along with descriptive statistics from the 2010 NST. I then highlight the statistical DIF estimation method, logistic regression, utilized in the study. Next, I discuss the purpose and design of the individual item analysis rubrics employed in the substantive review, the process for selecting item evaluators, the steps in administering the rubrics, and the use of group discussion for each item. In the last two sections I present the methods used for determining the inter-rater reliability of the evaluators' marks and the rank order correlation estimation procedure for determining the relationship between the statistical DIF and evaluators' predictions.

Content and Development of the 2010 NST

The NST is administered at the end of May in all regions of the republic over a two week period. Examinees receive their NST score reports at the end of June. The NST lasts 3 hours and 35 minutes and in 2010 had 150 test items (CEATM[a], 2010). The items in this study were taken from the NST *verbal reasoning* (словесно-логический) domain. This domain 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 items are multiple-choice with three distractors and one answer key. The verbal reasoning format of the NST contrasts with what was historically assessed for university entry, native language and literature which focused on knowledge of grammar and literary works (Drummond & De Young, 2004). Descriptive data from the test variant analyzed is presented below. Reliability estimates are presented for the full complement of items from all variants, all verbal items, and for the test items analyzed in this study.

Table 4-1: Descriptive Data from the NST 2010

Descriptive Statistics for Test Variant Analyzed						
	N	Min	Max	Mean	Std. Error	Std. Dev.
Russian (All items)	2,850	47	241	137.20	.743	39.652
Kyrgyz (All items)	1,557	24	204	102.7	.658	25.973
Russian (Verbal)	2,850	10	119	69.35	.388	20.716
Kyrgyz (Verbal)	1,557	10	96	49.18	298	11.745

Reliability Estimates		**Cronbach's Alpha**	**N Items**
All Variants/Math and Verbal Items	Russian	.956	150
All Variants/Math and Verbal Items	Kyrgyz	.896	150
All Variants/Verbal Items only	Russian	.907	60
All Variants/Verbal Items only	Kyrgyz	.702	60
Analyzed Variant/Studied Items only	Russian	.871	40
Analyzed Variant/Studied Items only	Kyrgyz	.660	40

The last two reliability estimates given above are based on 40 verbal items. However, I analyzed only 38 of these items from the analogies, sentence completion and reading comprehension sections because two of the item pairs in fact contained different items: 18 analogy items, 10 sentence completion items, and 10 reading comprehension items were analyzed in total. 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:

> "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, pp.14-16).

In regard to sentence completion items:

> "Sentence completion checks (a) the ability to understand logical connections between different parts of verbal expression, (b) vocabulary richness" (CEATM, 2007, pp.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. This 60 minute section has 30 items" (CEATM, 2007, pp. 14-16).

Below are two English language versions of the type of items analyzed in this study. These are example items from a previous year as items from the 2010 test remain secret. Due to the length of the reading comprehension texts I did not translate items from that section 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) or CEATM (2007).

Table 4-2: 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 Item Adaptation Process

In 2010, the source language of all sections of the NST except 'grammar use' was Russian.[54] As highlighted in chapter two, a large percentage of non-Russians in Kyrgyzstan, speak, read, and write in the Russian language with native proficiency. Therefore, finding personnel to adapt items is not difficult. CEATM test developers rely primarily on peer review

[54] According to Dr. Valkova, the director of the testing center, CEATM has experimented with developing different test sections in various languages in the past. In other words, test items have not always been developed in Russian first and then adapted into Kyrgyz.

and substantive methods of item evaluation to determine adaptation quality and equivalence of the adapted test forms. However, they also calculate p-values (difficulty) and discrimination coefficients of items in order to get a more complete understanding of how items are performing. CEATM reports that in addition to the use of back translation (see Table 4-3 below), close cooperation between item development groups and translators is maintained to ensure adherence to test specification(s) in all language versions as well as consistency in item construction and adaptation. Table 4-3 presents the item adaptation process for the 2010 NST. As can be seen, the Russian items are evaluated both substantively and statistically while the Kyrgyz and Uzbek items receive primarily substantive review.

Table 4-3: Flow Chart for Test Item Adaptation

1. Russian items created

↓

2. Russian items pre-tested

↓

3. Russian items are reviewed and revised based on pre-test results

↓

4. Russian items adapted into Kyrgyz and Uzbek

↓

5. K/U translator compares own version to Russian version

↓

6. K/U translator verifies the translation with the author of the R items by looking at the K/U items but reading them in Russian as the item author checks original meaning (oral, back translation)

↓

7. Specialists in Kyrgyz and Uzbek grammar review the target language translations

Interview with CEATM's head of test item development

Statistical DIF Detection Method

Investigators employ a wide variety of statistical DIF detection methods depending on study aims, skill level of the researcher, resource constraints, and nature of the specific tests and items examined. The most commonly utilized methods are Item Response Theory (IRT)

methods, the Mantel Hanzsel (non-parametric) chi-squared method and logistic regression (LR). Because it can not be assumed that statistical DIF indices are always correct (i.e. serve as a 100% reliable baseline from which to compare substantive evaluations) it is necessary to carefully select the statistical approach to be used in any DIF study and qualify any findings based on the analyses (Jodoin & Gierl, 2001).

As noted in the literature review, one challenge to DIF estimation is that the ability distributions between compared groups are typically not the same, especially in cross-lingual DIF studies. In selecting the appropriate statistical method for this study an important consideration was the large difference in ability distributions between the Russian and Kyrgyz groups. Russian examinees on average have performed consistently better on the NST since inception in 2002 (Valkova, 2004). Narayanan & Swaminathan (1996) found that DIF detection rates were more accurate when examinees were sampled from the equal ability groups than when unequal distributions were examined. However, if large enough sample sizes are used - and access to large sample sizes was not a problem in this study - this challenge can be addressed to some extent (Hambleton et. al., 1993).

After a careful review of methods, I elected to utilize the logistic regression (LR) method for DIF detection as articulated by Swaminathan & Rogers (1990). The LR model is easy to implement for the novice researcher, flexible, can detect both uniform and non-uniform DIF, and has power comparable to other DIF detection methods (Swaminathan & Rogers, 1990; Zumbo, 1999; Gierl et. al., 1999; Jodoin & Gierl, 2001). The LR method is a non-parametric probabilistic approach to DIF detection. In the LR method examinees must represent the

complete population of interest because non-representative samples will impact the results (Hambleton et. al, 1993).[55]

Unlike IRT models, non-parametric models utilize observed scores to test for the likelihood of difference in group performance on an individual item after conditioning on ability. The LR approach to DIF analysis relies on a chi-squared test of statistical significance and has an established measure of effect size.[56] In most non-parametric DIF studies, the total test score or sub-score on the instrument examined serves as a practical proxy for ability (Sireci, Patsula & Hambleton, 2005).[57] Considering the issues highlighted in the literature review about dimensionality, I elected to condition on verbal scores rather than the total NST score for this study. The logistic regression model for predicting the probability of a correct response to an item is based on (Swaminathan & Rogers, 1990):

$$P(u = 1|\theta) = \frac{e^{(\beta_0+\beta_1\theta)}}{[1+e^{(\beta_0+\beta_1\theta)}]},$$ Where:

u = the response to the item
θ = the observed ability of an individual
β_0 = the intercept parameter, and
β_1 = the slope parameter

[55] One, two and three parameter IRT models have been used to estimate for DIF. Each allows for estimation of item characteristic curves (ICC) which specify the relationship between the probability of success on the item and the underlying ability or trait. A key assumption in IRT models is that the estimates are invariant and do not depend on the sample. This differs from non-parametric models which utilize observed scores and thus to some extent depend on the samples utilized. For this reason, some have argued that IRT methods are superior because they allow for conditioning on true ability, not observed scores which are at best proxies for true ability (Camilli & Shephard, 1994).

[56] This was not the case with LR originally until Zumbo (1999) and Jodoin & Gierl (2001), introduced pseudo R-squared measures of effect size.

[57] 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 et al., 1993).

According to Swaminathan and Rogers (1990), by specifying separate equations, the probabilistic model presented above can be adapted for two separate groups of interest as:

$$P(u_{ij} = 1 \mid \theta_{ij}) = \frac{e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}}{[1 + e^{(\beta_{0j} + \beta_{1j}\theta_{1j})}]}, \qquad i = 1, ..., n_j,\ j = 1,2.$$

Where u_{ij} = the response of person i in group j to the item

β_{0j} = the intercept parameter
β_{1j} = the slope parameter for group j, and
θ_{ij} = the ability of individual i in group j.

This model can also be formulated as (Swaminathan & Rogers, 1990):

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

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

In simple terms, the use of this equation allowed the determination of whether or not to reject the null hypothesis of "no difference" in item response for two groups (Kyrgyz and Russian) on the particular item under study. In this study, 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. It is important to remember that the DIF analysis with LR proceeds at the item level; the data was entered into the equation for each test item individually. In this study that meant thirty-eight separate analyses for determining uniform DIF and thirty-eight separate analyses for determining non-uniform DIF.

For each item analysis, the dependent variable was a dichotomous variable - either a "1" for a correct item response, or a "0" for an incorrect response. On the right hand side, θ was a measure of examinee ability - observed sub score (verbal scores in this case). Language group membership was a categorical variable "G" and was coded "1" for Kyrgyz or "0" for Russian, sometimes called the reference and focal group, respectively. The term θG represented an interaction between these two independent variables. In DIF studies using LR methods, a significant interaction means there is evidence of "non-uniform DIF." Non-uniform DIF occurs when differences between two groups are not the same across all ability levels (Swaminathan & Rogers, 1990). For example, Russian examinees might perform better at the upper ability levels, but worse at the lower ability levels on the same item, or vice versa. In sum, the parameters β_0, β_1, β_2, and β_3 represented the intercept followed by the weights for ability, language group, and ability by language group interaction term respectively (Jodoin & Gierl, 2001).

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 recommend, I assessed each item in a two step process. In order to assess for uniform DIF, two models were identified. The first "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, 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 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) ($z = \beta_0 + \beta_1\theta + \beta_2 G + \beta_3 (\theta G)$). In other words, the chi-square value from the "full model" ($z = \beta_0 + \beta_1\theta + \beta_2 G$) was subtracted from the chi-square value from the third model with the interaction term and compared to the test statistic for significance (Jodoin & Gierl, 2001). In practical terms, for both the uniform and non-uniform tests, chi-square values lower than 3.841 indicated a very close correspondence in item response patterns between the two groups: I.e., I did not reject the null hypothesis of "no difference".

Further, such "identical" items had a β_2 (group) value at "0," or close to it. The Exp (B) or odds ratio for "no DIF" items was at or close to "1." In the LR model, understanding which group is favored is determined by the sign of the β_2 value. When $\beta_2>0$, the uniform DIF favored the reference group (Kyrgyz language). When $\beta_2<0$, the uniform DIF favored the focal group (Russian language). In general, non-uniform DIF is present when $\beta_3 \neq 0$, regardless of the value of β_2. When $\beta_3>0$, the item favored high ability Kyrgyz and low ability Russian. Items with negative values for β_3 favored high ability Russian and low ability Kyrgyz (Jodoin & Gierl, 2001).

An early criticism of the LR approach was that it did not have a measure of effect size (Kirk, 1996). This was considered a weakness as the power of the statistical test is somewhat dependent on sample size and large samples have a tendency to generate high type 1 error (over identification of significance). In recent years this problem has been addressed by Zumbo (1999) and Jodoin and Gierl (2001) with the introduction of an $R^2\Delta$ (r-squared delta), a weighted least

squares effect size measure. In this study I utilized this $R^2\Delta$ effect size measure proposed by Jodoin and Gierl (2001). Below I outline the steps I took to test for "practical DIF significance."

After testing for the statistical significance of each item, it was essential to interpret the results in terms of practical significance through the effect size measure. If the null hypothesis of "no DIF" was not rejected, there was no need to employ the effect size measure. However, if the chi-square test was significant, the $R^2\Delta$ needed to be assessed. For example, for determining the magnitude of significance for an item identified as statistically significant uniform DIF, the R^2 for the test score term (θ, compact model) was subtracted from the R^2 for the group membership term (G, full model). For determining the magnitude of significance for an item identified as non-uniform DIF, the R^2 for the group membership term (G, model 2) was subtracted from the R^2 for the interaction term (θG) model.

The resulting $R^2\Delta$ levels were then interpreted in light of Jodoin and Gierl's (2001) $R^2\Delta$ effect size measures. In simulation studies, Jodoin and Gierl (2001) demonstrated that this approach results in more powerful detection and lower type one error. These effect size measures were effective in trials with both simulated data and real data (Zheng, Gierl, & Cui, 2005). In a study by Gierl, Rogers and Klinger (1999), the $R^2\Delta$ effect size measure utilized in the LR method correlated at .91 with the MH effect size measure for an analysis of math items and .93 with the MH effect size measure for a social studies test.

The values utilized to classify the practical significance of DIF were the following:

- 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

In order to demonstrate how I utilized the logistic regression method and effect size measure proposed by Jodoin and Gierl (2001) in this study, I present two example item analyses here for items 7 and 32 (uniform DIF).

For item 7, the chi-square value for the compact model ($z = \beta_0 + \beta_1\theta$) was 159.771. The chi-square value for the full model (language group added, $z = \beta_0 + \beta_1\theta + \beta_2 G$) was 161.089. The difference in these two chi-square values was 1.318, lower than the test statistic of 3.841 at the .05 significance level. The β_2 (group variable) was low, estimated at .122. Recall that a β_2 value at zero or very close to zero indicates no difference. The odds ratio for this item, Exp (β), was 1.13 and odds ratios at 1 or close to 1 indicate the same odds of response for both groups. For item seven, the null hypothesis of no difference was not rejected and the response patterns to the Russian and Kyrgyz versions have a close one to one correspondence after controlling for ability; i.e., there was "no DIF" for this item pair, neither statistical nor practical.

For item 32, the difference in chi-square values between the compact model and the full model was 96.334, statistically significant and much higher than the test statistic, 3.841. The r-$R^2\Delta$ (effect size) difference was .057. The β_2 was also far from zero at 1.101. Further, the odds ratio was not near 1, but 3.007. This meant that the Kyrgyz group was just over three times more likely to answer the question correctly than the Russian group (recall that a positive β_2 means the item favors the Kyrgyz group). All of the 38 items were analyzed and interpreted in turn per the above steps for both uniform and non-uniform DIF. The results of these analyses are presented in the next chapter and can be found in full in Appendix K (uniform DIF) and Appendix N (non-uniform DIF).

Preparing for the Statistical Analysis

Before analyzing each item for DIF with the LR method I took several preliminary steps. First, I physically examined the test booklets from both languages to ensure that the items were indeed the same for both language versions. The result of this investigation revealed that from the 40 items initially selected for analysis, two item pairs (items 1 & 6) actually contained different test items. Based on their own preliminary analyses, the test center believed that the original items were not satisfactory and resolved to utilize two completely different items. I thus removed these two items from the analysis.

After confirming that the rest of the items were in fact the same, I requested the item response data from the test center for the test version under study. Data was provided in Excel format and included an indicator for the language version of the test, an item response matrix which included a dichotomous "1" or "0" (correct or incorrect) for each item, and a verbal score (scaled) for each student from the analogy, sentence completion, and reading comprehension sections. Each student was denoted by an eight digit identification number which was tied to the students' test registration center.

Sample Selection

There is converging evidence from DIF detection studies that larger sample sizes enable more accurate DIF detection (power). In Rogers & Swaminathan's (1993) comparison of MH and LR methods, they discovered that the detection rates increased by 15% when the sample size was increased from 250 to 500 for both methods. In Mazor et al.'s (1992) study, various sample sizes were created from 100 to 2,000 per group. They found that when the smaller sample sizes were used perhaps only 45-65% of DIF items were being correctly identified while in the larger samples, 65-85% of DIF items were being correctly identified. Hambleton et al.'s (1993) review

of simulation studies confirmed that these findings about sample size are true different across combinations of item types, ability distributions and other experimental conditions.

However, there is also evidence that type 1 error (over-identification of DIF when none is actually present) can increase with larger sample sizes. Thus, one concern is that even the most trivial differences between groups can be identified as statistically significant even though they are of little practical significance. While small sample sizes fail to capture much DIF, with sample sizes around 5,000 it is conceivable that much detected DIF will be of no practical significance, i.e. have an unacceptable Type 1 error rate (Hambleton, 1989). Thus, the need for the use of large (but not too large) sample sizes of 200-1,000.

In 2010, 30,264 examinees sat for the NST; approximately 18,720 in Kyrgyz, 10,994 in Russian and 1,000 in the Uzbek languages (CEATM[a], 2010). However, there were several versions of the NST, each with about 4,000-6,000 examinees. The test version provided by CEATM had a total 4,407 examinees and it was administered to both rural and urban participants throughout the country, including examinees from the capital and surrounding areas. 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.[58]

The Individual Item Analysis Rubrics

In order to answer the research questions, item analysis rubrics had to be developed that would capture not only the evaluators' estimations of content, meaning and difficulty differences between item pairs, but also elicit hypotheses about the cause or source of those differences. They needed to be short enough to allow efficient administration but thorough enough to ensure

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

that essential data was captured to facilitate interpretation. I designed the rubrics based on insight gleaned from similar studies (Allalouf et al., 1999; Reckase & Kunce, 2002; Ercikan, 2002; Ercikan et al., 2004). An overall path model for the process of collecting data through the individual rubrics is provided as Appendix G.

After consultation with the director of the *Center for Educational Assessment and Teaching Methods* (CEATM), the items selected for analysis came from the NST 2010. The items to be analyzed were collated in test booklets. As the construction of the test item booklets required access to the test items, this booklet was put together only after I arrived in Bishkek. The test booklets (rubric 1.a) consisted of each of the 38 item pairs, one item pair per page. There was also space to write notes and a place to mark whether or not items were identical or exhibited differences. Rubric 1.b was a graphic organizer which required evaluators to provide an initial categorization of the type of differences (if any). English versions of rubrics 1.a and 1.b are presented together as Appendix I.

For rubric 1.a the evaluators first attempted to correctly answer all the items in both the Kyrgyz and Russian versions in the test booklet. This was a "blind review" in the sense that the evaluators did not know which items had been identified as DIF by the statistical methods (Ercikan, 2002). Evaluators took notes only on the most important problems that arose. After going through all items, item pairs coded as "identical" on rubric 1.b were set outside as they were not needed for the completion of rubric 2. I developed and translated rubric 2 before I arrived in country. Rubric 2 had the following sections: (2.1) estimation of the level of difference(s) in content, meaning, or difficulty (if any) between the two items in the pair; (2.2) the specific nature of the difference(s); (2.3) description of the difference(s) in detail; (2.4) estimation of which group might be advantaged (favored) by differences; (2.5) suggestions for

improving equivalency of the item pairs. Rubric 2 was printed in three colors for three categories 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 the issue during later analysis. English and Russian versions of rubric 2 can be found in Appendix J.

Section 2.1, level of difference(s), required evaluators to classify each pair of items as "somewhat similar," "somewhat different," or "different" in meaning, content or difficulty. A coding scheme, adapted from both Ercikan's (2002) and Reckase and Kunce's (2002) work, defined these terms as follows:

0- Identical: no difference in meaning, content, or difficulty between two versions;[59]
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.

Before presenting the process for the administration of the item analysis rubrics, I first review how the participating bi-lingual item evaluators were selected.

Selecting the Evaluators

Recall from Chapter 1 that there are no professional pyschometricians in the Kyrgyz Republic. There are many educators with experience adapting textbooks and other educational materials from Russian into Kyrgyz, but few with experience in cross-lingual standardized test development.[60] As there has been no standardized testing until recent years, there is not a large pool of human resources from which to draw upon that has experience with DIF studies or item

[59] The actual choices on rubric 2 were only somewhat similar, somewhat different, and different because for the items they selected as "identical" on rubric 1b, they did not fill in rubric 2.
[60] With the exception of a few CEATM employees and ministerial assessment specialists who have been receiving training since 2002.

review. Since 2002, the test center has relied upon bi-lingual educators and translators in test development (item writing) and adaptation. Through experience, the test center has gradually identified those who have shown ability in this area and they maintain a small pool of personnel with whom they work on a short term basis as needs arise.

It was important that the pool of selected evaluators be as skilled as possible, bi-lingual, preferably with some experience in testing, but at the same time not have direct experience with the particular 2010 items. In other words, the challenge was to select a pool of evaluators who were a proxy for "as qualified as any other feasible sample" of the potential evaluators, but not have a conflict of interest (inability to evaluate objectively) due to experience working with the 2010 items. It was decided that eligible candidates could be those with experience writing or adapting NST test items in previous testing years, item writers who worked on other sections of the NST, translators with good reputations and content specialists who were known to be bi-lingual and relatively knowledgeable about assessment issues. Ultimately, four of the ten evaluators selected had never written nor adapted test items at any point in their professional careers, two had been item writers for previous iterations of the NST, three had been item writers for NST 2010 items not evaluated in this study, and one evaluator was selected who had participated in the adaptation of the NST 2010 items under study.

Selection of competent bi-linguals was essential. Perhaps the biggest challenge in selecting the evaluators was ensuring that all participants were as close to being as purely bi-lingual as possible. While finding bi-linguals was not difficult in Kyrgyzstan, pure bi-lingualism is rare as bi-linguals are usually stronger in one language than in the other (Korth, 2005). I included not only linguists and translators in the evaluation process but also teachers. This is because the item review process requires not only the identification of linguistic differences in

the two language versions, but also a judgment as to whether these differences might lead to performance differences (Mazor, 1993; Ercikan et al., 2004).

As highlighted in Chapter 2, it is primarily ethnic Kyrgyz who are bi-lingual in Russian and Kyrgyz as Russian speakers of other nationalities tend not to know Kyrgyz (Korth, 2005). However, there is a wide spectrum of skills and knowledge amongst those who claim to be Kyrgyz-Russian bi-lingual. Table 4-4 below presents an approximate typology of Kyrgyz and Russian knowledge levels found in the ethnic Kyrgyz population.

Table 4-4: Typology of Ethnic Kyrgyz Russian Language Knowledge

Kyrgyz Primary (Basic Russian)	**Bi-lingual**	**Russian Primary (Basic Kyrgyz)**	
Kyrgyz Only	**Kyrgyz Primary (Good Russian)**	**Russian Primary (Good Kyrgyz)**	**Russian Only**

Potential evaluators were identified with the assistance of test center employees. Each prospective candidate was contacted and provided with information about the study. If they agreed to participate, they first completed a brief 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, participants were informed in an interview that they would be required to use both Russian and Kyrgyz equally, not only on the 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, evaluators 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 they learned about this requirement.

Through the survey, each candidate provided information about his or her professional background and language ability. I then selected the ten evaluators that provided a balance in terms of competency levels in both languages. All the evaluators 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 translation and linguistics in the republic (De Young et al., 2006). The majority of participants selected more than one profession. This is because in Kyrgyzstan bi-lingual educators often serve in many capacities: As translators, teachers, test item writers, consultants, or work in other capacities on a short term basis in addition to their primary place of work (*ibid*, 2006). This broad spectrum of professional experience was beneficial as bi-linguals who know the school program or educators who have experience creating test items can approach the evaluation task from a multitude of perspectives and with practical experience in a relevant discipline. None of the selected evaluators had ever participated in a formal DIF study before. The table below presents the characteristics of those selected to serve as evaluators.

Table 4-5: Background Characteristics of Selected Evaluators

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

In terms of schooling, half the evaluators completed their secondary education in the Russian language medium and half in the Kyrgyz language medium. Three evaluators received

higher education in both languages while only two completed their higher educations in the Kyrgyz language medium. Seven evaluators reported using both languages at work and six of them reported using both languages in the home. None of the evaluators reported that Russian was their primary home language. Interestingly however, four evaluators 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 four marked that they were equally literate in both languages. All participants signed consent forms and were compensated for their work.

Administering the Rubrics

The administration of the item analysis rubrics and group discussion required three half-days of work to complete. Prior to convening, each evaluator received a glossary of technical terms which defined all key concepts (English version, Appendix H). Evaluators familiarized themselves with this material prior to coming to the analysis on June 19th. On June 18th I conducted a pre-test of the rubrics with one evaluator in order to determine if adjustments were needed to the rubric or glossary. The pre-test yielded important results: In addition to the discovery of some minor formatting and typographical mistakes, in a debriefing the pre-test evaluator reported that the most challenging aspect of the rubric was interpreting the coding categories in section 2.2. Although definitions of "adaptation, translation, format and cultural issues" were provided in the glossary, the pre-test participant claimed 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 2.3).

As the main purpose of the rubrics was to get an estimation of differences and gather good descriptive data about each item, I instructed the other nine evaluators to focus on sections 2.1, 2.3, 2.4 and 2.5. Emphasis was placed on section 2.3, description of the issues that they discovered with each item. It is indeed 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, section 2.2 was left intact. In the section that follows I present the steps of the data collection by each day's activities and tasks.

The evaluator panel was convened at 98 Tynustanova Street at the *Center for Educational Assessment and Teaching Methods* (CEATM) at 9:00 am on June 19th, 2010. All ten evaluators came on time and participated in a forty-five minute overview of the item evaluation process. Evaluators were then split into two groups and each group started with different item numbers. One group started with item 2 while the other started with item 20. This ensured that all items received at least a minimum amount of coverage. Then, evaluators were seated in individual work stations and began their individual analyses.

The first task was for the evaluators to answer and analyze the thirty eight test items on rubric 1a (test booklet). Then, they provided an initial mark as to the nature of difference (if any) on rubric 1b. Each evaluator completed the analysis individually. Evaluators wrote their comments in the rubrics in Kyrgyz and Russian. This process took approximately three and one half hours. All rubrics were collected at approximately 13:00 and stored in a secure location until the continuation of work the next day.

On Sunday, June 20th, all ten evaluators arrived again at 9:00 am and worked until lunch time. Their task was to complete rubric 2 for each item they had marked with any rating other than "0 = identical" the day before. This step required the evaluators to take their notes from day

one and code their comments on the four sections (2.1, 2.3, 2.4 and 2.5) presented above. This stage of the process took approximately four hours to complete. A fifteen minute coffee break was organized after the second hour. At the end of this session, the booklets and rubrics were collected and analyzed in the evening for key patterns and issues.

I reviewed the rubrics on June 20th because the time allocated on day three for discussion was three hours: it was essential to make sure that items were prioritized for discussion. The initial review focused on their estimated "the level of differences" (section 2.1) and "description" (section 2.3) for each of the items. If certain items elicited high marks, much commentary or varying views, it was essential that the group discuss these issues on day three. As it turned out, the time for the group analysis on day three was adequate to cover all the items.

Group Item Analysis

A three hour group discussion was held on Monday, June 21st. Including this discussion time, the total time spent with evaluators was approximately ten and one half hours. I facilitated the discussion in the Russian language and a note taker from the test center recorded the conversations. As facilitator, I allowed the conversation to flow but on occasion needed to intervene to keep the discussion on track. Areas of agreement and disagreement were noted and recorded. Evaluators shared their thoughts and feedback freely about each item. Data from these discussions were later 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. The English version of group discussion for each item is presented in the summary rubrics in Appendix W.

While evaluators marked each item individually, it was important to come to agreement about how to interpret their total marks as a group for each item. In order to establish an

operational definition of group "DIF prediction" each evaluator stated their opinion on how to best interpret their marking scheme. In simple terms, it was necessary to determine how many marks by evaluators would serve as "a vote for DIF" from the group. Several opinions were stated but ultimately they agreed that four total marks in any combination from the two "upper categories" of "somewhat different" or "different" would be considered as a vote for DIF. Recall that these are the marks that received 2 or 3 points for DIF.

While the term "group discussion" has been used up to this point, the term "group analysis" will be used going forward. The term group "analysis" underscores the point that throughout the discussion process, evaluators continued to analyze, study, and process the items. In the discussion of some items, evaluators changed their minds, saw the items in a different light, debated, argued, or discovered nuances of the item pairs that they had not noticed during their individual analyses. Thus, group *analysis* better characterizes what actually happened during the discussion of each item.

Summary Rubric

Descriptive data from the individual analyses and discussion notes provided data about evaluators' predictions of DIF levels and information about causes of DIF. As over 150 individual rubrics were filled in, it was important to have a way to collate this data in summary form. All data from each evaluator were thus recoded onto one summary rubric. For each of the individual 38 items, the full range of commentary from all ten item evaluators is coded in one place (Appendix W). For example, under section 2.3 for each of the items on the summary rubric, each bullet point and comment represents a statement from a different evaluator. All comments from the individual rubrics were translated verbatim without editing or synthesis on the summary rubric. This presentation of the full data allows the reader to see the entire scope of

comments for each item. Further, it allows the reader to see 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 summary rubrics presented as Appendix W differ from the individual rubrics completed by each evaluator in a few important ways. On the summary rubric section 2.2, the "nature of difference" data was not recoded from each of the individual summary rubrics. Recall that after the pre-test, evaluators were instructed to focus on item description in section 2.3 and not to worry about the accuracy of their coding in section 2.2. The *a priori* coding categories under section 2.2 were used to guide evaluators' thinking in how best to characterize the differences between the item versions.

The "level of difference" on section 2.1 of the summary rubric was coded under the color-coded categories (content, cultural/language, for format) as submitted by each evaluator. I used these categories as a way to collate the data but did not focus on the consistency of the evaluators in marking these categories. Notice in the summary rubrics that evaluators' comments about the same issue often fell under the different headings. A difference that was defined by one as "cultural" for example, might have been characterized by another as a "content" issue. The important data for analysis was the totality of the description, not how the issues were coded according to each individual evaluator. Otherwise, the summary rubric in Appendix W reflects the same organizing principles and data as collected from each of the individual rubrics.

Estimating Inter-Rater Reliability

After collecting data from the evaluator rubrics, group discussion and statistical analyses, there was significant data about both the perceived DIF and as well as actual DIF levels based on the statistics. An important question on the use of evaluator/raters in any study is the extent to which their estimations can be considered reliable. As bi-lingual evaluators represent a sample of a larger population of possible evaluators, it was necessary to see how much measurement error existed. Thus, the first step of analysis was to determine the inter-rater reliability of the evaluators' marks and how much variation there was in their estimations.

In order to do this, an inter-class correlation coefficient was estimated with SPSS software. Inter-class correlations are ratios of rating variance to total variance and can be used as reliability coefficients for assessments of raters that are deemed to be in the same category or class (McGraw & Wong, 1996). In order to estimate this coefficient I had to first develop a scoring system that would allow the coding of the evaluators' marks for each item. Recall that on section 2.1 of rubric 2 each evaluator estimated the level of difference between the item versions under study. The coding scheme was "0" (identical), "1" (somewhat similar), "2" (somewhat different), and "3" (different).

In order to estimate reliability, I produced a matrix of their scores for each item in an Excel file. Each column represented an evaluator and the thirty-eight rows represented each of the items analyzed. In the matrices I placed their marks of 0, 1, 2 or 3 in each cell based on their perception of the level of differences in item pairs as defined above. Before conducting this analysis I reviewed the data from their individual evaluation rubrics and decided to drop two evaluators from the analysis. The one evaluator who had worked as a translator on the NST 2010 filled out only six total rubrics and his rubrics contained a considerable amount of missing

values.[61] A second evaluator filled out the rubrics incorrectly using the same single rubric to record marks for many different items. This led to confusion and I could not determine 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 evaluator. Nonetheless, after dropping these two evaluators, a group of eight evaluators remained to provide an ample number marks for each of the items.

After these two evaluators' data were removed, the marks from the eight remaining evaluators were examined for missing data. There were 13 missing entries from a total of 304 possible entries (38 items x 8 evaluations). I imputed data for these missing scores by entering the average scores from the other seven evaluators into each cell where data was missing. I then calculated Pearson's reliability in SPSS. Two-way random effects models are used where people effects and measures effects are random. I selected "two-way ANOVA, random" and selected "consistency." I then selected "absolute agreement" to see if there would be differences in these estimates. I report the results of these analyses in the results chapter.

Estimating Evaluators' Accuracy in DIF Prediction

The key question of interest in this study was the extent to which evaluators could accurately predict statistical DIF. Therefore, the relationship between evaluators' DIF predictions and the statistical outcomes needed to be established. Evaluator DIF predictions consisted of two separate steps. First, they had to estimate the extent of differences between the

[61] While it might be expected that one of the specialists working on the NST 2010 adaptation was not likely to offer critical commentary, it was nevertheless important to include one of them in the study. And, this individual, having more experience with the items, made an especially valuable contribution to the group discussion. Note that on the rubrics in appendix W the total number of marks comes from the eight retained evaluators: That is, all estimations utilized only the marks from the eight evaluators. However, the commentary under section 2.3 and group discussion comes from all ten evaluators.

items in the pair. Second, they had to predict which group, if any, would be favored by these differences. In order to assess the relationship I conducted a rank order correlation analysis between the marks of the evaluators and the chi-square difference values for all 38 items. Recall that as the chi-square values ascend, they move towards DIF, away from item equivalence that is indicated by values below the test statistic, 3.81. That is, the DIF items have the highest chi-square difference values while the non-significant items have very low chi-square difference values.

The correlation was estimated in the following manner. Recall that in order to quantify the meaning of the distinct "levels of difference," I assigned points for different levels of categorization (0, 1, 2, and 3). These scores were totaled across all eight evaluators to produce a combined total score for each item: Higher scores thus represented a stronger belief in DIF while lower scores represented a weaker belief in DIF. After calculating the scores for each item, I conducted the rank order correlation analysis using Spearman's rho in SPSS. I employed Spearman's rho because it is thought to be less sensitive to outliers in the data than Pearson's coefficient (SPSS User's Guide, Version 16). The results of this analysis are presented in the next chapter.

Chapter 5: Results

DIF Detection Results

As explicated in the methods chapter, logistic regression was utilized to analyze each item. The uniform DIF statistics with all chi-square values, effect sizes, significance levels, β_2 values and odds ratios for each item are presented in Appendix K. Recall that the sign of the β_2 value indicates which group is favored (positive Kyrgyz, negative Russian). In all, a total of six items had no statistical DIF (items 9, 2, 24, 7, 17, and 29). This indicates a very close correspondence in response patterns between the two versions of the items in these pairs. These six non-significant values are italicized in Appendix K. Twenty-eight items had negligible DIF, three items had moderate DIF (13, 19, and 32), and one item had large DIF (item 3). Throughout the rest of the study, I refer to these four DIF items (three moderate, one large) together as "practical DIF" items when distinguishing them from the statistically significant but "negligible DIF" items.

The data in Appendix K are presented in order of *ascending chi-square difference* values. Recall that this chi-square difference value is the difference between the chi-square value of the compact model from the chi-square value of the full model (with the group variable added for uniform DIF). This chi-square value is checked against the test statistic of 3.814 with 1 degree of freedom to assess for significance. Thus, in the table, the lowest values (and non-significant items) come first while the four practical DIF items have the highest values and are the four last items in order of ascension. Of the four items identified as practical DIF, three favored the Russian group while one favored the Kyrgyz group. Of the 32 items classified as negligible, moderate or high DIF, 18 items favored the Kyrgyz group and 14 items favored the Russian group. The statistical results for the four items classified as practical DIF are presented

separately in Appendix L. The six items with no statistically significant DIF are presented in Appendix M. Recall that there was no need to report an effect size for non significant items since all the chi-square difference values were below the test statistic, 3.841.

All items were also tested for non-uniform DIF in the same two step process. This time the compact model included the group (language) variable and the full model included an interaction term, $\beta_{3,}$ (θG). Twenty-one of the items had no statistically significant non-uniform DIF. Seventeen items had statistically significant chi-square values but all were classified as "negligible DIF." The largest effect size (r-squared delta) was .018. Thus, there were no practically significant non-uniform DIF items. The full results of this analysis are presented in Appendix N. These items are also arranged by chi-square difference values in ascending order.

In Table 5-1 below I present the percentage of items in each uniform DIF category by item type. As is evident, the majority of items under study fell into the negligible DIF category. In order to be classified as negligible DIF (statistically, but not practically significant) the effect size for the item had to be $0.0 \leq .035$ (Gierl & Jodoin, 2001). The effect sizes of the 28 negligible uniform DIF items ranged from .003 to .031: The higher the effect size, the closer to the cut off for moderate DIF at .035. For enhancement of interpretation, I split the group of negligible DIF items into halves. This is because there appeared to be "clustering" by item type along the effect size distribution. The median effect size value was .009. When put in rank order, an item with a .009 effect size is the 14^{th} item in the range of 28 negligible items. Note in table 5-1 below that the analogy items were spread throughout all classification categories relatively evenly. The sentence completion and reading comprehension items were concentrated more heavily in particular categories.

Table 5-1: Items (%) by Effect Size Levels and Item Type

Item Type	Non-DIF	Neg. Low	Neg. High	Practical DIF	Total
Analogy	22%	28%	33%	17%	100%
Sentence Completion	20%	20%	60%	0%	100%
Reading Comprehension	0%	70%	20%	10%	100%

Of the reading comprehension items, 90% were categorized as negligible DIF. Of the sentence completion items, 80% were categorized as negligible DIF. However, the reading comprehension items tended to cluster with lower effect size values (below the .009 median) while the sentence completion items tended to cluster with the higher effect size values (above the .009 median). Fifty percent of all the items below the median in the negligible DIF category were reading comprehension while only 14% of them were sentence completion items. At the same time, only 14% of all the negligible DIF items above the median were reading comprehension items while 43% were sentence completion items. In other words, there were proportionally more sentence completion items closer to moderate DIF levels than reading comprehension items. Appendix O presents each item by item type and effect size level in order to demonstrate this distribution across effect size levels.

Overall, from the perspective of those who adapted the items, having only 4 of 38 items classified as "practical DIF" is a positive result. This is a considerably lower percentage of DIF items than have been found in many cross-lingual DIF studies as noted in the literature review (Chapter 3). As I will argue below, however, these estimations might be a bit conservative in terms of the actual number of items that merit further review by CEATM. There were other items that received both criticism from evaluators and had relatively high effect size values near

the .035 cutoff. I return to the issue of effect size categorizations and reasons why some non-practical DIF items might be problematic in Chapter 6. Next I present the results from the inter-rater reliability estimation, rank order analysis, and evaluators' predictions about DIF direction.

Inter-Rater Reliability and Rank Order Estimations

The average number of rubrics filled out per evaluator was 17. The most active of the evaluators filled in 31 rubrics while the least active filled in 6 rubrics.[62] Such wide variation in the number of rubrics completed was also reported by Plake (1980). The least active evaluator was on the team of translators who worked on the NST in 2010. As highlighted in the methods chapter, I conducted an analysis of inter-rater reliability using marks from eight evaluators. The evaluators and measures were both considered random. The inter-rater reliability coefficient when I selected "consistency" was .66 with a 95% confidence interval of .473 to .804. The inter-rater reliability coefficient when I selected "absolute agreement" was .66 with a 95% confidence interval of .462 to .796. These modest, positive correlations are indicative of a fair amount of agreement between evaluators. The full matrix with the evaluators' marks used in the statistical analysis can be found in Appendix P. The SPSS output from these analyses for "consistency" can be found in Appendix Q.

Recall that the rank order correlation estimation assessed the relationship between the evaluators' total score for each item and the chi-square difference value for that item. After summing the individual marks for each item, the total item scores ranged from 0 to 16 total points per item; the higher the number – the stronger belief in difference by the evaluators. The mean score for the 38 items was 6.62; the mode was 5; the median was 5.5; the standard

[62] Appendix U charts the number of distinct item issues with each individual test item noted by the evaluators. It is important to keep in mind that in many cases, these should be understood as "alleged" item issues, not necessarily proven issues as some issues were clearly disputed during group analysis.

deviation was 4.48. Using Spearman's rho in SPSS, the result of the rank order correlation was a significant, positive relationship of .45, .004 significance at the .01 level. The two columns with item scores and their corresponding chi-square difference values used in the analysis are presented in Appendix R. The SPSS output from this analysis is presented in Appendix S. This modest correlation indicates that as evaluators' total scores for the items increase, so do the chi-square difference values. These results provide support for a modest correlation between evaluators' DIF predictions (of difference estimations, not DIF direction) and statistical DIF outcomes.

This relationship between the evaluators' marks and the chi-square difference values is also visible through graphical representation. Appendix T presents the evaluators' marks in one column next to the chi-square difference values arranged in ascending order. Instead of using item sum scores, to enhance visual representation I simply entered an "X" for each mark of "2" or "3" that the item received from evaluators. Any marks of "somewhat similar" (1 point) for example, were not included as an X in this table. Recall from Chapter 4 that evaluators created an operational definition of what total quantity of evaluator marks indicated "belief in statistical DIF." It was decided during the group discussion that a total of four marks in any combination of "somewhat different" (2 points) and/or "different" (3 points) would be considered a vote for "probable DIF." Thus, in Appendix T, each item with four Xs represents a vote for DIF from the evaluators for that item. In essence, four total marks serves as a "cut score" for DIF from the perspective of the evaluators; less than four total marks for any item pair means evaluators (as a group) believed statistical DIF unlikely.

Note that as the chi-square difference values ascend in Appendix T, the items with a larger number of evaluator marks tend to cluster in the bottom half of the table. While there are

a low number of evaluator marks for some items with high chi-square difference values (e.g. items 20, 28 and 13), for the first sixteen items with low chi-square values, only one of those items has four or more marks (item 7). Eleven of these initial items have a total of 0 or 1 mark and four items have a total of two marks. Looking at the very bottom of the table however, it is also apparent that three of the four practical DIF items did not in fact receive four or more marks from the evaluators. Only item 3 exhibited a high statistical DIF level *and* received many marks (six) from evaluators as probable DIF. Thus, the positive rank order correlation can not be attributed to the close correspondence between the four practical DIF items and evaluators' predictions for these particular items but rather to the general tendency for clustering near the bottom. Five of the six items with five or more DIF marks from evaluators are located in the lower half of the table.

In total, eight items received four or more marks from the evaluators. Seven of the eight items predicted by evaluators were statistically significant and most were located in the lower part the table in Appendix T with relatively high squared values. The only item predicted to be DIF by evaluators that turned out to be not statistically significant was item 7 (four marks). From the eight items they predicted as DIF, two items received four marks, four items received five marks, and two items received six marks. The eight items predicted as DIF and their effect size values are presented in Table 5-2 below.

Table 5-2: Evaluator Marks and Statistics for Predicted DIF Items

Item	Evaluators' Marks	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

Several of the items that received high marks from evaluators were negligible DIF items that had relatively high effect sizes. Five of their eight predictions had effect size values above the effect size median of .009. For example, item 21 had a .024 effect size and received six marks.[63] Item 11 received five marks and had a .028 effect size. Item 33 received five marks and had a .027 effect size. In other words, several negligible DIF items that were very close to the "cut-off" of moderate DIF (.035) were also marked as probable DIF by evaluators. It seems that evaluators' moderately accurate estimations in the middle to higher part of the effect size order best explain the positive rank order correlation of .45. There were of course outliers in terms of correspondence between the two indicators which plausibly kept the overall correlation from being high. For example, item 15 received five marks from evaluators but had a fairly low effect size measure of .008. Items 7 and 18 also demonstrated little correspondence between evaluators' marks and the DIF statistics (many evaluator marks but non-significance or a low effect size value).

[63] Note that the chi-square difference values and r-squared values (effect size) are very closely (though not perfectly) correlated.

Direction of DIF

Despite the reasonable inter-rater reliability and modest correlation between the evaluators' predicted differences and statistical differences for the item pairs, the inference that evaluators had a reasonably good understanding of DIF would be tenuous: That is because the data collected from section 2.4 of the item rubrics indicate that evaluators did not correctly predict the "direction of DIF" (which group was favored by differences) on a consistent basis. For the eight items they predicted as DIF, they correctly predicted the direction of DIF only 29% of the time (2 of 7 statistically significant items). The data in Table 5-3, arranged in order of chi-square difference values highlights this fact. Note the difference between their predictions of direction and actual DIF direction in columns five and six. Five of the seven items favored the Kyrgyz group. The evaluators were only correct in their predictions with the one practical DIF item (item 3) and with item 21.

Table 5-3: Prediction of DIF Direction for Items 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
** Item 21 also received 1 vote for favoring Kyrgyz.

Note that from the eight items they predicted as DIF, with the exception of one lone vote, they predicted DIF to favor the Russian group for each item. Thus, their only two correct

predictions were when the Russian group was actually favored. In fact, from the total pool of 38 items assessed, evaluators 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). Table 5-4 below presents a breakdown of the evaluators' marks for all 38 items in response to section 2.4 of the rubric - "which group is advantaged (favored)?"

Table 5-4: Prediction of DIF Direction for All Items

		Number of Marks				
	Total	**1**	**2**	**3**	**4**	**5**
Favors Russian	26	22, 27,29	2,4,9,12,13, 17,24,25,28, 31,32,36	7,10,11, 23,26,33	3,18,19,38	15
Favors Kyrgyz	2	5,30				
Mixed Vote	4	8 (1R,1 N/A), 16 (1R,1K), 21(5R,1K), 35 (1R,1 N/A)				
No Advantage	1	37				
No Estimation	5	14, 20, 34, 39, 40				

From Table 5-3 it is clear that sometimes the number of marks in section 2.4 (which side is favored) was sometimes less than the total number of marks for DIF, section 2.1. It would appear that in some cases evaluators were a bit more confident that there were differences in items than they were in which group might be advantaged by those differences. However, there were also cases when items received very few "different marks" in section 2.1 of the rubric but several for "favoring the Russian group" in section 2.4. For example, only three evaluators marked item 19 as "somewhat different" or "different" yet four total evaluators marked it as favoring the Russian group. Items 9 and 28 received no marks for "somewhat different" or "different" but still received two marks per item as "favoring Russian." Of course there is no

reason why evaluators could not have selected the category "somewhat similar" and also selected a group to be favored. As the majority of items marked "favoring Russian" received only 1, 2, or 3 total marks as such, I re-examined each individual rubric to check if this result might be a function of the dispositions of certain evaluators. I discovered this not to be the cases as there was a roughly equal distribution of "favoring Russian" marks across all evaluators.

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. Items 3 and 19 received four marks in favor of the Russian group, while item 13 received two marks in favor of Russian. Item 32, which advantaged the Kyrgyz group, was not predicted to be a DIF item but still received two marks as "favoring Russian." This apparent lack of accuracy (overall) in predicting DIF direction was similar to results from Plake (1980) as well as Engelhard et al. (1990) with black and white group differences. Plake (1980) found that the raters scored twice the amount of DIF than the statistical procedures yielded. In this study, evaluators also scored two times more DIF than the DIF statistics indicated. In Plake's study, one third of items favored the opposite direction that was predicted by the raters while in this study the evaluators (while accuracy rates differed by item type) overall were only 52% accurate when including all their predictions in the analysis (including negligible DIF items) and only 29% accurate for those items they predicted as DIF.

At the same time, these results contrast with Gierl and Khaliq's (2001) study of cross-lingual DIF that found Canadian evaluators to have better than random prediction rates for DIF direction for French and English versions of mathematics and science items. Methodological approaches are perhaps important in understanding the accuracy of their substantive review however (Ercikan, 2002). In the Gierl and Khaliq (2001) study, the evaluators had knowledge of

the statistical data and they set out to classify DIF direction on item pairs they knew had been flagged as DIF. Perhaps it was therefore a bit easier for them to estimate DIF direction than in the Kyrgyz situation where evaluators had no knowledge about statistical DIF beforehand.

The larger point is that if evaluators can not accurately predict who is advantaged by differences in the two versions of an item it is difficult to determine how well they actually understood alleged item differences, regardless of the inter-rater reliability and rank order outcomes. It also underscores the difficulty of the task that substantive committees face in item analysis in general. I now turn to a presentation of the data by item type.

Reading Comprehension Items

Analysis of the reading comprehension items entailed the analysis of a reading text (195 lines in Kyrgyz, 165 lines in Russian) in addition to 10 individual item pairs, item numbers 31-40. Nine of the reading comprehension items were classified as negligible DIF and one, item 32, was moderate DIF. As noted at the beginning of the chapter, 70% of all the reading comprehension items had effect size values lower than .009 - the median effect size value of all the significant DIF items. Six reading items had chi-square difference values less than 10.30 and effect size values at .006 or less. In the rank order of negligible DIF items from lowest to highest, these items occupied the 1st, 2nd, 3rd, 7th, 8th, 10th, 14th, 17th, and 26th places respectively (see Appendix O). Only one negligible DIF reading item, 33, had a relatively high effect size value at .027. It also received five marks as "DIF" from the evaluators.

Overall, as can be seen from the rubrics in Appendix W, these items generated the least discussion in comparison to the sentence completion and analogy items. The reading comprehension items had the lowest average number of distinct issues per item at 1.5, and received the lowest average number of evaluator marks for DIF per item at 1.6. Eight items

received 0, 1 or 2 total marks as DIF from evaluators. The highest number of marks for any reading item was five (item 33) and three (item 38). The average effect size value for the statistically significant items was .0129. The most commonly noted issue for the reading comprehension items was format mistakes in the Kyrgyz language (5 times).

In Table 5-5 below, the reading comprehension items are presented in order according to ascending chi-squared values. "Marks" indicate DIF votes from evaluators, while "predicted" indicates which group evaluators believed the item favored. The numbers in parenthesis indicates the number of evaluators who voted for the predicted DIF direction. The last column- "statistics"- indicates the statistical direction of DIF.

Table 5-5: Statistically Significant Reading Comprehension Items

Item	Marks	$R^2\Delta$	DIF Category	Predicted	Statistics
39	0	.003	negligible	None	Kyrgyz
35	X	.003	negligible	No Adv. (1), R(1)	Russian
36	XX	.003	negligible	Russian (2)	Kyrgyz
31	XX	.004	negligible	Russian (3)	Russian
34	X	.006	negligible	No Est.	Russian
40	0	.006	negligible	None	Russian
37	0	.009	negligible	No Adv. (1)	Kyrgyz
38	XXX	.011	negligible	Russian (4)	Russian
33	XXXXX	.027	negligible	Russian (3)	Kyrgyz
32	XX	.057	moderate	Russian (2)	Kyrgyz

Five reading comprehension items favored the Russian group and five items favored the Kyrgyz groups. In the six cases in which evaluators predicted advantage,[64] only three were correct (50% accuracy). The conversation around the reading comprehension items was perhaps tempered by the nature of the task. Evaluators had to not only to read and analyze the items, but compare the texts as well. For a full list of comments about the reading comprehension text, see the last page of Appendix W. The analysis of reading comprehension items generated commentary about issues of adaptability in general but few strongly supported and highly agreed upon hypotheses about problems with specific items. This was not the case for the sentence completion items to which I now turn.

Sentence Completion Items

No sentence completion items were classified as moderate or high DIF. Two items, 24 and 29, were non-DIF items. As noted above however, six of the eight statistically significant sentence completion items had effect size values higher than the median effect size value of .009. One item had a value of .016, two items had values of .019, one of .024, and one of .029, all somewhat close to the negligible-moderate DIF border at .035. In the rank order of twenty eight negligible DIF items by effect size value, these items occupy the 4th, 20th, 21st, 22nd, 23rd, 24th and the 28th highest positions respectively. This can be seen from the rank order of values in Appendix O. The item data for the sentence completion items is presented below in Table 5-6.

[64] "Predicted advantage" includes only items where there was at least one prediction but no cases of "split decisions" (one vote for one group, another vote for the other group).

Table 5-6: Statistically Significant Sentence Completion Items

Item	Marks	$R^2\Delta$	DIF Category	Predicted	Statistics
27	X	.003	negligible	No Est.	Kyrgyz
30	X	.004	negligible	Kyrgyz (1)	Kyrgyz
22	XX	.013	negligible	Russian (1)	Russian
25	XXXXX	.016	negligible	Russian (2)	Kyrgyz
26	XX	.019	negligible	Russian (3)	Russian
23	XXX	.019	negligible	Russian (3)	Russian
21	XXXXXX	.024	negligible	Russian (5)*	Russian
28	0	.029	negligible	Russian (2)	Kyrgyz

* There was also one vote for favoring Kyrgyz for this item

One item received six marks from the evaluators, one item five received marks and one item received three marks. The average number of evaluator marks for DIF per item for the sentence completion items was 2.5. The average number of issues per item was 3.1, twice the amount of the reading comprehension items. The average effect size value was .0159, higher than the reading comprehension value of .0129. Interestingly, compared to the other item types, evaluators correctly predicted the direction of DIF at a greater than random rate for the sentence completion items (5 of 7 times, or 71% correct). The most commonly cited problem for these items were the lack of syntaxical equivalence between the Russian and Kyrgyz items which made these items difficult for the Kyrgyz group (more below). As will be seen in the individual item analysis section, no items generated more discussion than the sentence completion items, especially items 21, 23, 25, 26 and 28.

Analogy Items

There were almost twice as many analogy items (18) examined than reading comprehension (10) and sentence completion items (10). Fourteen of the 18 analogy items were statistically significant; 11 were negligible DIF; 3 items were practical DIF; 4 items were non-DIF. Unlike the sentence completion items which tended to cluster at the higher end of effect size values, the effect size values for these items were evenly spread across the whole range of effect size values. For example, in the lower range there was one item at .004, one at .006, two items at .008, one each at .009, .010, and .011. There were also several middle level values as well as two negligible items with very high effect size values, item 11 (.029) and item 16 (.031). The three practical DIF analogy items of course had effect size measures over .035. The average effect size value was .0225 making it the highest effect size average of the three item types.

The average number of evaluator marks for DIF was 2.6. The average number of distinct issues per item was 2.0 placing it between the other two item types. The dispersion of evaluators' marks was wide with four items receiving only zero or one mark for DIF while three items received marks of two. Item 3 received the most marks at six. Two other items received five marks, one item received four marks, and three items received three marks. Table 5-7 presents the data from the analysis of the analogy items.

Table 5-7: Statistically Significant Analogy Items

Item	Marks	$R^2 \Delta$	DIF Category	Predicted	Statistics
14	0	0.004	negligible	No est.	Russian
12	XX	0.006	negligible	Russian (2)	Kyrgyz
18	XXXX	0.008	negligible	Russian (4)	Kyrgyz
15	XXXXX	0.008	negligible	Russian (5)	Kyrgyz
10	XXX	0.009	negligible	Russian (3)	Russian
8	X	0.010	negligible	Rus (1), Kyr (1)	Kyrgyz
4	XX	0.011	negligible	Russian (1)	Kyrgyz
20	0	0.015	negligible	No est.	Kyrgyz
5	XXX	0.015	negligible	Kyrgyz (1)	Kyrgyz
11	XXXXX	0.028	negligible	Russian (3)	Kyrgyz
16	XX	0.031	negligible	Rus (1), Kyr (1)	Kyrgyz
19	XXX	0.048	moderate	Russian (4)	Russian
3	XXXXXX	0.050	moderate	Russian (3)	Russian
13	X	0.072	high	Russian (2)	Russian

While the direction of DIF for the three practical DIF analogies was predicted correctly, overall, the evaluators correctly predicted the direction of analogy DIF 40% of the time (4 of 10 predictions correct for which there were no split estimates). The negligible DIF analogy items tended to favor Kyrgyz (9 of 11 items) while the practical DIF items all favored the Russian group. Overall, the evaluators overwhelmingly selected the Russian group as favored for the analogies items. There was no consistently marked "typical problem" for the analogies items: A wide variety of translation and adaption, cultural and format issues were noted as problematic. As will be highlighted below, many predictions about DIF for analogy items did not come to fruition. Sometimes issues like *loan word* use or social-cultural issues were projected to cause DIF on analogy items but did not. However, plausible causes were identified for the three items

flagged for DIF. For these three items, mistakes in "key location of meaning" such as answer keys (3, 13) and item stems (19) all plausibly led to DIF (Ercikan, 2002).

Overall, the reading comprehension items had the lowest effect size measures, the lowest number of distinct issues identified by evaluators, and the lowest average number of DIF marks from evaluators. They also generated the least amount of discussion. Only one item was practical DIF (32), though one other item was very close to moderate DIF (33). Evaluators were not however, able to offer an explanation for DIF for item 32. The analogy items demonstrated the most variation, both in terms of evaluator marks and the DIF statistics across the statistical distributions (effect sizes). They also had the highest average effect size values.

There were no practical DIF sentence completion items but many of these items were concentrated near the moderate DIF cut off of .035. These items also received the second most marks for DIF on average (2.5, while analogies received 2.6), the highest number of distinct issues per item, 3.1, and generated the most discussion. The table below presents summary information about the evaluators' marks by item type. These results by item type are consistent with other cross-lingual DIF studies of verbal items. For example, Agnoff and Cook (1988) argued that longer texts (reading comprehension) allow for the flexibility necessary in item adaptation to more accurately convey meaning. Indeed, if there are inherent differences between languages, they are perhaps less constraining when longer texts are involved.

Table 5-8: Summary of Evaluators' Marks by Item Type

Data by Item Type	**Ana**	**Sent Com**	**Read**	
	(n = 18)	(n = 10)	(n = 10)	
Evaluators:				**Data source:**
Avg. Number Issues Per Item	2.0	3.1	1.5	all items
Avg. Number of Marks for DIF	2.6	2.5	1.6	sig. items
% Correct DIF Direction	40%	71%	50%	sig. items

As in this study, other researchers also found greater DIF in analogy items and less in reading comprehension items. In this study, three of the four practical DIF items were analogy items. Beller (1995) and Gafni and Canaan-Yehishafat (1993) found greater DIF in analogy items than in reading passages. Using data from the Israeli Psychometric Entrance Test (PET), Allalouf, Hambleton and Sireci (1999) also concluded that analogies seemed most problematic with 65% of items demonstrating DIF. Reading comprehension items showed the smallest amount of DIF in their research. In this study reading comprehension items were also concentrated in the lowest range of effect size values. In the final section of this chapter, I turn to the individual item analyses with a focus on specific sources of DIF in Russian and Kyrgyz language items.

Sources of Difference

The second goal of this study was to determine the source (cause) of DIF and the specific challenges to item adaptation from Russian into Kyrgyz. As presented in the methods chapter, data from each item pair was collected on the item analysis rubrics completed by each evaluator and from the group discussion. Though only eight total items were predicted as DIF, there was at least one distinct issue or problem noted with all but 2 of the 38 items (items 20 and 39). Of course the identification of an issue or problem does not mean that the issue was in fact widely agreed upon or correctly identified and characterized. In Appendix W the reader can get a sense of just how much agreement there was on any particular issue raised.[65] Further, most of the issues or problems did not lead to DIF as indicated by the overall low number of DIF items. Nonetheless, test developers need to consider the evaluators' full spectrum of comments on item

[65] While the marks under 2.1 which were used for inter-rater reliability and rank order scoring came from only eight evaluators, the commentary under section 2.3 contains the comments from all ten reviewers (Appendix W).

quality because their comments can assist evaluators improve the quality of their work. Recall that Engelhard et al. (1999) carried out a study in which evaluators tried to locate technical and cultural mistakes in items. The most accurate reviewer was 94% accurate while the least accurate was 83%. Thus, there is some evidence to believe that whatever their accuracy in DIF prediction, evaluators can be reasonably accurate in identifying mistakes in substantive review.

According to the analysis of the 157 individually completed rubrics, there were 82 distinct issues raised with the 39 items: 53 related to adaptation/translation issues, 17 related to Kyrgyz grammar, 8 related to item format, and 4 related to socio-demographic or cultural issues. The number of distinct issues per item ranged from zero issues (two items) to five issues (one item). Ten items were marked as having one issue, 11 items had two distinct issues, 11 items had three distinct issues, and 3 items had four distinct issues (see Appendix U). The average number of issues per item was 2.15. Eight of the 36 items with comments received no suggestions for how to improve the items (section 2.5) while the remaining 28 items received suggestions for how to make the item pairs more equivalent.

All 17 grammar issues were related to Kyrgyz grammar. Not a single issue was raised with Russian grammar for any of the 38 items. Further, despite the potential impact of background factors (e.g. cultural or curricular) to impact test results, except for the four issues raised related to socio-demographic/cultural issues, commentary and item discussion focused on overt language and format issues between the two versions of the items examined. This is perhaps explained by several factors. First, the NST is an aptitude test, not directly tied to specific school curricula. Second, as a test from within a single country the NST developers likely did a reasonably good job of considering variation of conditions, culture, and content across groups. Third, as will be seen below, many issues related to the quality of the Kyrgyz

items kept attention and debate squarely on the more overt item characteristics of that language group. As will be demonstrated, the few hypotheses generated about background factors that might have led to DIF were not tenable as 3 of the 4 practical DIF items had overt technical flaws due to poor adaptation or typographical mistakes.

In the following section I break down the data by sources of difference identified by item evaluators. In some areas of concern like poor translation, many pairs of items exhibited similar issues or elicited similar debate. In those cases, two or three items are presented below as examples of recurring themes and additional examples are referenced in the summary rubric. Parts of conversations that seemed to be especially insightful are presented in the text as quotations from evaluators. In order to facilitate a coherent presentation of results, I developed the following system of references to the individual item rubrics and group analyses. I reference the two data sources as either "IA" for data coming from individual analysis rubrics and "GA" for data coming from the group analysis. For example, IA12 indicates that the data came from the individual analyses from item 12. GA33 indicates that the data came from the group analysis of item number 33.

In order to avoid confusion I have kept the original answer key and distractor names (letters of the alphabet) in the Russian style as presented in the summary data: (А), (б), (В) and (г), (A, B, V, and G in English), which is similar to the American style of distractor labeling (A), (B), (C) and (D). As noted in Chapter 4, each item has four possible answer choices. The term "item stem" is used to denote the prompt or question, and "answer key" to denote the correct answer choice while "distractor" denotes any one of the three incorrect choices. Item evaluators are referred to by two initials - MD, CJ, AB, etc.

Translation and Adaptation Issues

In the western literature, the term item *adaptation* is generally preferred to *translation* because of its connotation of flexibility in conveying meaning (Sireci & Allalouf, 2003). Adaptation implies that as long as the essential meaning, nuance, and difficulty level is kept intact and conveyed, words and phrases appearing in the source language may be changed as necessary for the target linguistic/cultural group. This use of *adaptation* is intentional - as a way to distinguish it from *translation* which implies a more literal, word for word approach (Hambleton, 2005). In this study however, I subsume translation under the larger umbrella of adaptation as I do not believe the issue is "either – or." Usually, flexibility is needed to make sure that nuance is accounted for in the target language version. However, as the data in this study will reveal, literal translation is sometimes more appropriate than misguided or overly creative attempts at adaptation. Therefore, I present issues of adaption and translation together and distinguish the nuances in the application of the terms as necessary in the context of each item under discussion.

A myriad of adaptation and translation issues arose during item analysis**.** If an item was adapted but not directly translated, evaluators tended to make a note of it as potentially problematic. In the ensuing group analysis, it was then debated whether or not the adaptation was appropriately done. Sometimes, evaluators argued that flexible adaptations were necessary while at other times they claimed that such adaptations were problematic. In analogy item 2 for example, the Russian stem was given as "chef: borscht." The logical relationship was "maker, preparer: something made/prepared by him/her." The evaluators noted the literal translation in the two versions did not correspond. In the Kyrgyz stem, the second word given in the pair was "шорпо" (broth), which in Kyrgyz can have a wide meaning and imply not only "something

liquid" but also "something eaten as a first course." In the Russian pair, "борщ" (borscht) is the name of a particular kind of soup (IA2). The evaluators noted:

> **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. **KM**: Yes, they are different, but I don't think the differences affect the relationship of the words in the analogy pair (GA2).

The evaluators agreed with KM as three of them marked the item as "identical" while four of them marked it as "somewhat similar." Their rationale was that although the second words in the item pairs were different, this did not impact essential meaning as the primary relationship was still "chef: *something prepared by a chef*" in both versions of the item. In fact, this item displayed no statistical DIF with a chi-square difference value well below the test statistic of 3.841 at .733.

In analogy item 5 the item was adapted (not directly translated) but the relationships in the two different versions were maintained. It was noted that in distractor (A), the Kyrgyz pair of words was "бут: из" (leg (*foot*): track) while the Russian version was "палец: отпечаток" (finger: fingerprint) (IA5). This time both the first and second words in the pairs were different. Evaluators were more divided over this item as three of them marked the versions as either "somewhat different" or "different" while four of them marked it as "identical." However, this item was also not a DIF item but displayed negligible DIF with an effect size of .015 (greater than the median effect size). There were other examples of analogy item adaptation that did not result in changes in the relationships between word pairs. In item 8, one Russian distractor was "ladle: pour" while the Kyrgyz version was "bucket: pour." The item received only one vote as DIF from evaluators and was in fact negligible DIF. In both language versions, this particular

distractor was the least popular: It was selected by 7% of the Russian group and 1% of the Kyrgyz group.

In distractor (B) of item 9, the Russian version was “мокрый: сушить” (wet: to dry) while the Kyrgyz version was “суу: кургатуу” (*water*: to dry). There were differences of opinion about the appropriateness of this adaptation. One evaluator noted, “If these words were used in context (in a sentence) then it would be okay. For example – ‘I was in the rain and got wet.’ However, when no context is given, this is a problem and literal translation is necessary” (IA9). The concern expressed was a minority opinion however, as no evaluators marked “somewhat different” or “different” and four evaluators marked it as “somewhat similar.” The difference in words in fact did not impact examinees as this was a non-DIF item with almost perfect one to one correspondence. The item had the lowest chi-square difference of all 38 items.

Not all adaptations of analogy items maintained essential meaning however. The evaluators noted cases of poor adaptation and outright translation mistakes. In item 10 the second word in distractor (B) of the Kyrgyz version was mistranslated from the Russian version and resulted in the Kyrgyz word having the opposite meaning than was intended. The word used in the Russian item was “Ярче” (brighter) but the Kyrgyz distractor (B) was translated incorrectly as “карараак” (darker). The Kyrgyz word for ‘brighter’ that was needed was “ачыгыраак” (IA10). Three evaluators marked this item as DIF but the DIF level was negligible at effect size .009. The mistake was located in the distractor that was least attractive to both the Russian and Kyrgyz group, which might explain why the mistake did not seem to have an impact on item responses.

Five evaluators believed that multiple translation errors on item 11 would lead to DIF. And, this item had an effect size of .028, putting it right at the verge of moderate DIF (recall the cutoff of .035). Two of the Kyrgyz distractors - "г" and "б" - had translation problems that changed the meaning of the distractors. One evaluator noted that the mistakes made the item difficult and confusing for the Kyrgyz group (IA11). During group analysis however, a different evaluator claimed that distractor б (which was not the answer key) was an attractive distractor in the Russian version, but not in the Kyrgyz version (GA11). Nothing was found to be wrong with the Russian version. Interestingly, while three evaluators marked the item as favoring the Russian group, the item in fact favored the Kyrgyz examinees. Because one evaluator noted that the number of attractive distractors in the Russian version was greater than in the Kyrgyz version, it seems plausible that mistake in the distractor of one language version could perhaps have made the odds of correct selection *greater* by reducing the total number of viable distractors for this group, assuming of course that the mistake was obvious to examinees.

Item 18 had translation problems in distractors (б) and (г) and a comment that one of the words in the Kyrgyz answer key pair was "used in simple speech" not literary language. According to evaluators:

> In (б), "шашма" (hurried), does not correspond to the Russian version "откровенный" (open) and "шамдагай" (dexterous) is not the same as "болтливый" (talkative). In other words, neither word in this pair corresponds well to the pair in the other language. This distractor does not work. (г) also has an incorrect adaptation as "колдойгон"k. (attack) which is used in simple speech, not as a literary term. Further, the meaning of the pair of words in Kyrgyz does not correspond well to the meaning of the words in Russian" (IA18).

And during group analysis:

> **ZS**: There is incorrect, inaccurate translation in several of the distractors in this item and the use of the incorrect meaning of some words. **NO**: the problem is related to the specifics and nuances of the Kyrgyz language. The thing is, some words can be used only

> in combination with each other; in certain contexts they can't be used individually. Therefore, this issue is poor adaptation.
>
> **RM**: Yes, the problem is that some words must be used in combination. **MD**: The use of some words out of context makes them impossible to understand, they can not be used individually, so the problem is adaptation. (GA18).

Four evaluators believed that item 18 would be a DIF item. In fact, item 18 was categorized as negligible DIF with an effect size value of .008, right at the median effect size level. As with item 11, item 18 also favored Kyrgyz respondents, the group in which these mistakes with word combinations were allegedly occurring. However, as in item 11, the four votes cast for DIF direction were for "favoring Russian" (IA18). This pattern of identifying mistakes in the Kyrgyz version and voting for "favoring Russian" occurred frequently, especially with the analogy items. As presented in the beginning of this chapter, 9 of the 11 "negligible DIF" analogy items actually favored the Kyrgyz group but all were predicted by evaluators as "favoring Russian." This was not the case with the sentence completion items for which prediction rates for DIF direction were better.

Another important adaptation issue that arose several times was the use of Russian *loan words* in Kyrgyz versions of the items.[66] For example, on item 4 most evaluators noted that a commonly known Russian *loan word* should have been retained in one of the distractors (IA4). Instead, the Russian word "шарф" (scarf) was adapted into the Kyrgyz "моюн жоолук" (*lit.* neck wrap) (IA4). There was consensus on this point:

> **ZS:** I think foreign words should stay in their original form. **AA:** I agree; if there are no commonly used equivalents for foreign words, use the commonly used version. **MK:** It is best to use active, commonly used, words (GA4).

[66] The term cognate means a word that is the same in several languages (i.e. shares the same origin). I use the term loan word here to emphasize that the most cognates were introduced through Russian in the 20th century.

> **MD:** the problem here seems to be a too literal translation; sometimes there is no reason to translate.
>
> **NO:** Actually, I think there is a Kyrgyz equivalent to “шарф” (scarf) but it is not used very often. **MD:** Well… how can we say what “often” is – how do we know this? (GA4)

While recommending the maintenance of a *loan word* for this item, the evaluators were divided about whether this was a DIF item or not. Two marked it as “somewhat different” with another four marking it as “identical” and two as “somewhat similar.” In fact, it was a negligible DIF item favoring the Kyrgyz group. The Kyrgyz stem of item 24 also contained an adaptation of a Russian word that several evaluators noted should have been left in the Russian original (IA24).

In a different discussion just a few minutes later, evaluators made the opposite recommendation in regard to *loan word* use. Six of the evaluators noted that several Russian *loan words* in item 7 would not be understood by Kyrgyz speakers (IA7) and four of them marked the item as “somewhat different.” The Russian words “терапевт” (therapist), “слесарь" (metalworker), “Адвокат” (advocate) were all identified as problematic, especially for rural (Kyrgyz) students. An alternative Kyrgyz word was proposed for ‘metal worker’ - “темир уста”, which means literally “мастер по железо” (master of iron) in Russian (IA7). Item 7 however, showed no significant DIF.

In item 2 the Kyrgyz version of distractor (Г) also employed a *loan word* from Russian, “деталь” (detail). Four evaluators proposed that the Kyrgyz equivalent “тетик” (detail) be used instead. It too, however, showed no indication of DIF with a chi-square difference value well below the test statistic. It would appear that in general, the Kyrgyz examinees were not troubled by the Russian *loan words* in most cases. However, it would be incorrect to say that the evaluators strongly believed that they would be; based on their marks, they were divided or had mixed feelings about how *loan word* use would or would not impact item response patterns.

Sometimes they felt they should be used, and sometimes not, depending on the item under evaluation. In any event, there is no evidence that the use of *loan words* on Kyrgyz versions caused DIF on any of the items analyzed with the possible exception of sentence completion item 23 (discussed below).

Group analysis of item 23 raised the question of how to deal with *new words* and concepts and the fact that their incorporation into the Russian and Kyrgyz languages (from other foreign languages) might proceed at an unequal pace, especially considering the demographic distributions of the two populations within the KR. One of the Kyrgyz words utilized, "камсыздандыруу" (insurance, provision) in the item stem "has no meaning in Kyrgyz" in the context of this item according to one evaluator because the concept of "insurance" is unknown (IA23). According to another, the word is technically correct, but only understood by a small number of specialists. Yet another opinion was that the Kyrgyz word "камсыздандырылган" (to be guaranteed) might fit the item but that its meaning has a wider connotation than the Russian word utilized in the item, "застрахование" (insured). It was generally agreed that the item had an urban (pro-Russian) bias and three evaluators believed it would be a DIF item favoring the Russian group. While not a practical DIF item, in fact this item did have a relatively high level of negligible DIF at .019 and the Russian examinees were favored. In the discussion of whether the word for "insurance" was known or unknown by Kyrgyz examinees, evaluator ZS noted:

> "… Many new terms are constantly being formed all the time in Kyrgyz while in Russian the concepts are well known. For example, in Kyrgyz there are four or five completely different ways to say "entertainment center." People do not know which is correct at this point. Therefore, Kyrgyz people often use Russian loan words. Some use Kyrgyz words that are not known. As teachers we see this on a regular basis. Many words are 'created' but not yet well known. In this item not everyone knows how to say "uninsured," especially in rural areas where there is no such thing as insurance" (GA23).

Despite the lack of strong evidence for differential impact of *loan/new words* on item response patterns overall, the discussions of the above items demonstrate the difficulty of the *loan word/new word* issue in standardized testing situations. In GA4 evaluator MD raised the core issue: Given the demographic diversity of the republic, how does an item developer get a handle on the "commonality" of a given *loan word*? Or, as the evaluators often stated, the "activeness" of a given word. The evaluators noted that there were differences in the extent to which Kyrgyz speakers lived and interacted daily with Russian speakers and thus differences in the extent to which they would be exposed to *loan words* and new words.

Recall that the examinee sample in this study comes from a broad, representative slice of the population, including the capital city of Bishkek. This may mean that a moderate or large part of the Kyrgyz sample is relatively well acquainted with Russian *loan words* which might explain the lack of DIF on the items presented above. That is, there may be some Kyrgyz speakers who *are* penalized by the use of Russian loan words but this doesn't show up in the overall statistics because they are small proportion of the examinees in the sample. One way to get a better understanding of this issue would be to conduct experimental DIF studies with two Kyrgyz language groups – one from an ethnically mixed area and one from a more isolated area where Russian is not well known or had less penetration historically.[67] I will return to this issue in the discussion in the next chapter.

Another challenge for evaluators was rectifying the *multiple meanings* of individual words in many of the items. Apparently, there were cases where a word had a clear, singular meaning in either Russian or Kyrgyz but several meanings in the other language, thus

[67] On the other hand, many non-Russian speaking Kyrgyz would still likely know many loan words that found their way into Kyrgyz usage decades ago and, in a sense, "became Kyrgyz words," regardless of their knowledge of Russian.

complicating the logic/meaning of certain items for one group. This finding of multiple meanings as a potential DIF cause is consistent with other DIF studies of verbal test items (Gierl & Khaliq, 2001; Sireci & Allalouf, 2003). Evaluators speculated that examinees would be confused on analogy item 3 as they wouldn't know which of the meanings in Kyrgyz was needed to solve the item:

> **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 the Russian "сад: яблоня" (orchard: apple trees) into Kyrgyz is incorrect. The given Kyrgyz version is – "бак: алма" (tree: apple). **NO**: Yes, but in Kyrgyz "бак" can mean tree 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 trees." In other words, "from what/where" (material) comes (GA3).
>
> **MD**: I agree, "бак"k. (tree) is "сад"r. (orchard) *and* **"**дерево"r. (tree). The word "бакча" k. is "огород"r. (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 (GA3).

Item 3 was in fact a practical DIF item and it received six DIF marks from the evaluators. However, there was also a serious typographical error in the answer key (discussed below under format) that most believed was the cause of DIF because there was no viable answer key in the Kyrgyz version (GA3). Multiple issues within the same item occurred often in the Kyrgyz items which made disentangling potential sources of DIF challenging.

Analogy item 13 was another example of how multiple meanings might cause DIF. Item 13 was also a practical DIF item favoring the Russian group though only one evaluator initially predicted DIF. During the group analysis however, it became apparent that many believed this to be a DIF item (GA13). The analogy item stem was "television: watch." The answer key (б) was "automobile: go." The main problem was the second word in the Kyrgyz key - "жүрүү" –

which in Kyrgyz means “to go.” However, depending on the combination of words used with this word, it can mean to go by foot or by car. The Russian version, in contrast, employed the word “ездить” which means going *only* by some form of transportation: Russian has a different word for going *on foot*. Thus, the essential relationship between words that was crucial for the analogy to work was clearer in the Russian version. As the evaluators noted:

> **RM**: the problem is that the distractors are not good. **MD**: Yes, maybe the main problem is in answer key (б). In my comments, I wrote that “жүрүү” k. (go) is different from “ездить” r. (go by transport) because “жүрүү” can be walking by foot or going by car while “ездить” means going by transportation. In Kyrgyz perhaps “айдоо” (drive) would be a better choice for the pair because it has meaning like the Russian “ездить.” … If they used “айдоо,” (drive) they will get it quickly… I think this is an issue of translation – it is a good item but the direct translation is incorrect. Many of us thought for a very long time about what the correct answer here was to this item… (GA13).

Analogy item 19 was another DIF item favoring the Russian group. Initially, three evaluators marked it as DIF on their individual analyses. During the group analysis however, there was considerable discussion about a serious problem with the item stem. The Russian item stem would roughly be the English equivalent of “hot: recoil/jerk back quickly.” Apparently, the second word in the Kyrgyz stem pair “тартып алуу” has two meanings – “take away” and “pull” (IA19). The second word in the Russian stem, “отдёрнуть”r. (recoil/jerk back quickly), has only one meaning. In the Kyrgyz stem, in combination with the first word in the pair “ысык” (hot), the second word, could be understood as “attract or “pull warmth,” which implies attraction, not repulsion, the opposite of the what the Russian item stem implied with “recoil” (IA19, GA19).

> **RM:** there are multiple meanings of some words in the item stem; there needs to be a more careful selection of pairs of words – otherwise, the item misleads and it becomes impossible to find the correct answer. ... **NO**: I agree, depending on how they define the terms in the stem they could come to complete opposite meanings of the analogy... **MD**: Yes, the stem needs to be more clearly defined (contain no double meanings). **MK:** Absolutely, the stem and distractors should have only one meaningful interpretation (GA19).

While the above practical DIF items (3,13,19) all had issues with multiple meaning, it is important to highlight that in all three cases, the core problems were with either the answer key (3,13) or the item stem (19), not other distractors. Thus, the DIF in these cases seems to be as much about "location of the problem" or the extent to which the overall meaning of an item becomes confused as much as it is about "multiple meanings" in general. This finding is consistent with Ercikan (2002) and underscores how only through an examination of the minutia at the item level can DIF analysis be fruitful. I will return to this issue in the discussion chapter.

Discussion around sentence completion item 26 demonstrated that due to grammatical issues some Kyrgyz items were difficult to comprehend, even for the evaluators. While the DIF was negligible the effect size value was fairly large at .019 (favors Russian, two DIF marks). On how lack of clarity can complicate understanding, evaluators noted:

> **ZS**: "себептүү"k. (due to) in the item stem is not needed. It needs a different affix here. **MK**: I do not agree…. without this, the item loses the main idea. What will be the correct answer?
>
> **ZS**: This item is confusing, the translation is not clear in several places. **MD**: Hmmm… It seems that the Russian text allows a "double meaning," and Kyrgyz only one meaning. However, that meaning (for the Kyrgyz item) leads to a wrong answer. This is due to the way the item is structured.
>
> **MK**: What *is* the correct answer to the Kyrgyz item? **MD:** The complication is over the meaning of the word "production" which is quite unclear. **ZS**: If we can't find the correct answer, I don't think the children will either! (GA26).

Reading comprehension item 33 elicited discussion as several mistakes were noted. Item 33 was a negligible DIF item but had a relatively high effect size of .027. Five evaluators marked it as a DIF item. There was consensus about a translation mistake in distractor (A) of the Kyrgyz version, though distractor A was the least attractive distractor for both groups. Though several evaluators thought "the overall meaning of the text is similar in Russian and Kyrgyz" the

multiple meaning of some words was noted at the end of the Kyrgyz stem (IA33). It was also noted that the Kyrgyz stem contained a sentence that was too long. However, despite the mistakes in the Kyrgyz item, the item favored the Kyrgyz group. Item 33 also elicited some general commentary about the reading comprehension text:

> **ZS:** There are many problems with the translation of this difficult text; it is not well adapted. One resolution is to take an original Kyrgyz language text, related closely to this theme and then select the Russian text because it is difficult to completely pass on the entire meaning and deeply consider the question in a foreign language…I must say that the Russian text is quite good, as are most of the items in Russian. I can't find any difficult words, grammar mistakes, etc. But syntax issues might explain any differences.
>
> **MD**: It is easy to see a lack of connection due to translation issues. I think that the analytical thinking on the part of the Kyrgyz is different. **ZS**: I did not find any difficult words or issues with the item itself. Maybe some issues with the form of the sentences (constructions – syntax) in the reading text though. It is clear that the key is (B) but (G) is also an attractive answer (GA33).

An important issue that came up during group analysis was the alleged "Russification" of Kyrgyz syntax and linguistic expression in some test items (and the Kyrgyz language in general). This is interesting in light of the historical discussion presented in Chapter 2 about the Russification of Kyrgyz in the 1920s and 1930s. Analogy item 15 is an interesting case of how a Russian (source) item can allegedly influence the adaptation of a Kyrgyz (target) item. According to evaluators, the word employed "кубанычсыз" (*lit*. happiness + form for "without," the ending сыз) in distractor (Г) was "artificially created" or a made-up word. While "кубаныч"k. (happiness) is a word, the addition of the suffix in this case was inappropriate. In the Kyrgyz language "сыз" is often added to nouns to indicate "without." In theory, adding it here could have seemed like a creative way to convey the meaning of "unhappiness" which was easily conveyed in the Russian version.

Evaluators posited that it was likely created to help "fit" the Kyrgyz item to the Russian version (GA15). At the same time, one evaluator offered an improvement with a different Kyrgyz word choice - "көңүлсүз" (unhappy). While five evaluators marked this as a DIF item, the item had a relatively low chi-square and effect size measure (.008), making it a "negligible DIF" item. Further, this particular item actually favored the Kyrgyz group: Another example of an item with allegedly poor adaptation into Kyrgyz that nonetheless did not seem to be causing DIF in favor of the Russian group. Again, it seems plausible that obviously faulty distractors (a non-sense word in this case) could actually assist the Kyrgyz group by eliminating these particular distractors as viable answer choices.

There was considerable discussion about the linguistic "Russification" of the Kyrgyz versions of the sentence completion items. Sentence completion items introduced more complexity into the discussion as stems and distractors got longer and more complicated. In particular, evaluators argued that the "Russian origin" of item 21 was obvious. Evaluators maintained that they could tell that the original item was developed by a "Russian thinker" because the form of the Kyrgyz item had a Russian form. The result was that the Kyrgyz version was less authentic and even "artificial." The main issue was the inappropriate use of the Kyrgyz connector "жана" (and) which most evaluators argued could lead to considerable confusion (IA21). At the same time, several evaluators noted that, incorrect Kyrgyz or not, many Kyrgyz people use this expression incorrectly in their everyday speech. From the GA21 discussion:

> **MD**: I think this item needs to be completely changed as it will not be easy to simply adapt. The main problem is the incorrect use of the Kyrgyz term "жана" (and), which is obviously the result of a direct translation from the Russian sentence. **AA:** I agree, but the problem is that this is a common usage in Kyrgyz. It's on the radio, the national TV stations and other official media sources. **MD**: It is common but it is not correct. **AA**: I understand...

MD: I believe this usage is a one of those "Russianisms" that has crept into Kyrgyz through (ethnic) Kyrgyz, Russian language speakers. The main problem is that in Kyrgyz we don't use "and" as a connector when connecting two different verbs. Two verbs together often come together to convey a different meaning than when they are used singularly. The two verbs are simply put together, without the use of any connectors.

ZS: I agree with MD, villagers don't use "жана" (and) in this sense – they use Kyrgyz correctly… I think if Kyrgyz original texts had been used, there wouldn't be this problem… Our syntax is different and should be Kyrgyz – not Russian. **MD**: Well, theoretically, I agree of course. A big problem is that much of our literature in the sciences and the arts is translated in this way – translated directly from Russian. Much of the Russian influence is inevitable. Little is produced in Kyrgyz due to a lack of specialists and resources. **ZS**: OK, but what if we had several specialists work on developing the items at the same time and then decided whether they would work or not?

MD: To me, this item raises a bigger question from the perspective of the test translators. Should the items contain only language that is 100% correct or contain language that is incorrect but commonly used? Unfortunately, there is often a gap here. The situation and state of the Kyrgyz language is very sad. Further, we have to take into account language as it is used on a daily basis. Many people in the cities – and not only in the cities – speak Kyrgyz with lots of words and forms taken from Russian. Sometimes the language is simply all mixed up. This is a result of the language environment we live in. We combine Russian and Kyrgyz all the time in a sort of hybrid colloquial language. For example, "канча (how many - *kyr*) листов (lists (pieces) of paper - *rus*)?" or "сиз (you *kyr*) домой (home *rus*) барасызбы? (are going? *kyr) – lit. Are you going home?* There are hundreds of ways we do this. This item raises some big issues. (GA21).

Item 21 was predicted as a DIF item by six of the eight evaluators. While it was a negligible DIF item (favoring the Russian group), the effect size level of .024 was very close to "moderate" DIF at .035. Five evaluators correctly predicted that the item favored the Russian group. Along with item 3, these marks represented perhaps the strongest sense of agreement on the part of the evaluators throughout the analysis.

Discussion around item 21 also captured many of the issues that arose elsewhere in item and group analysis. One issue was that the Kyrgyz language was like a "moving target," constantly under pressure and evolving, lacking standardization, and thus too easily influenced by Russian speaking ethnic Kyrgyz who introduced Russian language forms into Kyrgyz and by

the arbitrariness and dispositions of individual translators. Again, historical context presented in Chapter 2 about the "unfinished business" of Kyrgyz language codification and standardization seems relevant to this discussion (Korth, 2005). Another issue was that evaluators began to question the methods of item adaptation from the source language (Russian) to the target language (Kyrgyz) for the sentence completion items.

The challenge of the adaptation of complex material also came up in sentence completion item GA23. Like item 21, this item was also negligible DIF (again favoring the Russian group) and the effect size measure was also large at .019. Three evaluators predicted DIF for this item and three evaluators correctly predicted that the item favored the Russian group. Evaluators noted that inherent differences between the two languages made the sentence completion items difficult to successfully adapt (GA23). The fact that Kyrgyz is an *agglutinative* language allegedly makes certain types of long Russian sentences too complex to be clear in Kyrgyz without significant adaptation.

In agglutinative languages, meaning is typically conveyed through the addition of affixes (typically suffixes in Kyrgyz) to nouns which can determine possession, number, location, direction, etc. For example, the noun "kiz" (girl), becomes "kiz*dar*" in the plural form. To indicate "to the girls" the form becomes "kizdar*ga*" (Oruzbaeva, 1997). In this way, words can become quite long but sentences usually remain relatively short. According to the evaluators, long Russian sentences can complicate understanding of complex material in Kyrgyz. Yet, in order to create the logical relations needed to make a sentence completion item work, sentences need to be relatively long, often complex enough to allow three to four "blank spaces" in the sentence where examinees must fill in the needed words to form coherent meaning. Some of the

conversation about item 23 led to further recommendations about how to modify the item development process for the sentence completion items. In the words of the evaluators:

> **CJ:** I had difficulty reading and understanding the Kyrgyz text; then, I read the Russian text and I understood. I think that with background knowledge (knowing Russian) they might be able to understand some of the meaning. However, Kyrgyz-only speakers will find it confusing. That is, if the Russian concepts are covered first, then one knows what the Kyrgyz authors meant to say.[68] However, the students do not get this advantage because they do not know the Russian version. **AA**: We (item analysts) have an advantage because we can read both items at the same time! … So, the text is not well adapted, and this makes it difficult to comprehend (GA23).

> **KK:** The desire to pass on the main idea of the task was too much, as there is the possibility of losing the literary nuances of Kyrgyz; important to consider the characteristics of the language – in Kyrgyz sentences are usually short – as words are "complex" (compounded – i.e. agglutinative), and the result of the direct translation is that translated texts (from Russian into Kyrgyz) are longer than usual for Kyrgyz speakers. As they become longer, they become more confused. And, the sentences eventually become even longer than the Russian versions.

> **KK**: … I believe that it is possible to find original texts in Kyrgyz and then translate them into Russian - then you will see the richness of differences of the languages. **ZS:** Yes, the problem is the translation and the adaptation due to stylistic differences… One adaptation suggestion would be to not have the Kyrgyz sentences copy the Russian style but to make them "more Kyrgyz." This means making the sentences shorter, even if it means more sentences... **MD**: ZS, I agree with your first point – in Kyrgyz we have "long words" but short sentences. This is important to remember in comparison with Russian (GA 23).

The issue of the length of Russian sentences (too long for concise adaptation) came up again in the discussion of item stem 24 (IA24). And, the recommendation to break up longer Russian sentences into shorter Kyrgyz ones in GA28:

> **MD**: The challenge for test writers is that for some Kyrgyz texts it becomes complicated when we try to repeat the Russian syntax and constructs. It becomes complicated when translation is literal. The best way to keep the Kyrgyz intact is to break the Russian sentences into more sentences rather than trying to capture the Russian structure. In

[68] With this conversation in mind, I conducted an experiment in which I split the Kyrgyz speakers into two groups and conducted more DIF analyses. I present the results of this experiment in the discussion chapter.

> Kyrgyz, ideas are built not through one complex sentence, but through a series (many sentences) with simpler ideas that when compounded, express the same idea (GA 28).

Agglutination was not the only alleged challenge to sentence completion item adaptation. Another issue raised was word order. In Kyrgyz, verbs, and hence essential meaning and ideas, come at the end of a sentence. In Russian, they can come before or after the noun.[69] They evaluators noted:

> **ZS**: … We often start to translate from the end of the sentence because the main idea comes last (the verb is at the end). Word order is different in Kyrgyz and Russian which can also cause complexities. Because of the word order, sometimes, I translate the literal sentences first – then rearrange them in order. The strategy here is to read sentences more times or to hear it in Russian first, and then piece together the puzzle… (GA28).

Another conversation about item 28 looked at the same issue from a different perspective. One of the test center employees, a Russian-only speaking individual who was observing asked, "In the Russian version of item 28, the syntax is difficult. Is it difficult in Kyrgyz?"

> **ZS**: In general, syntax is easier in Kyrgyz than Russian. We have a straightforward "cause – result." The structure of Russian is more difficult.
>
> **MD**: I think that there are several levels of structure in Russian. In Kyrgyz, it is "single level" – it is this *and* this *and* this. New ideas are "added" while in Russian there is a different structure. In Kyrgyz it is all part of the same syntactical level. **ZS**: My Kyrgyz students also tell me that the Russian constructions are difficult to learn at first. In Russian you have "Due to the fact … Because of the fact …" in Kyrgyz, more direct statements…
>
> **MK**: Yes, for example, in Russian you may have … "event/phenomena … which is/that is…" etc. In Kyrgyz we have "this happened" (stop) and "that happened" (stop) and then something else. It is all on "one level." **ZS**: In general, syntax is easier in Kyrgyz (GA28).

[69] For example, *Men kizdarga bara jatamin* in English is roughly "I am going to the girls." However, literally, the words are Men (I) kizdarga (kiz/dar/ga = girl/plural form/to) bara jatamin (go). Note both the agglutination - as the single word kizdarga indicates direct object, number, and direction, and the word order – compound verb at the end of the sentence.

From the data generated from the item rubrics and discussions it would appear that there are challenges to adapting the more complex ideas in the short sentence completion items. Interestingly, the evaluators correctly predicted DIF direction (favors Russian) on these items with a better than random estimate, 71% accuracy. These items were also concentrated at the upper end of effect size values, close to the moderate DIF cutoff. I will return to a discussion of these items in the final chapter.

Socio-Cultural Issues

There was also discussion about the potential for cultural, socio-economic and demographic differences to impact item results on some items. This concern was usually expressed in terms like "kids from villages won't know …" (GA23) or once, "urban kids won't know…" (IA3). Issues related to contextual knowledge, regional dialect, and interface with Russian speakers were all noted. On occasion, the discussion digressed from looking at item specifics to conversations about how rural kids might be disadvantaged on the test in general. No curricular or instructional issues were noted by evaluators as potentially problematic.

One Kyrgyz item that was identified as containing "dialect" was item 25. Of interest here were not the response patterns of Russian and Kyrgyz groups but rather whether different segments of the Kyrgyz population would be differentially impacted by regional differences in the Kyrgyz language. Two of the distractors contained forms of the word Kyrgyz "Пас" (low, down) that were consistently marked as "southern dialect" that "northern kids might not understand" (IA25). Five evaluators believed that this would be a DIF item (perhaps assuming that enough northern Kyrgyz would be penalized by the item to result in overall DIF between the Russian and Kyrgyz groups). The item was negligible DIF and favored Kyrgyz speakers overall. In order to understand how these dialect differences impact results among Kyrgyz speakers it

would be necessary to conduct a DIF study by regions (north vs. south for example), utilizing data from two different Kyrgyz groups. Item 19 also contained words in Kyrgyz that were allegedly dialect (IA19).

Other concerns were raised about familiarity with terms that certain demographic groups might not know. For example, for item 3, the concern was raised that the Kyrgyz word "күл" (ash) in distractor "г" would not be known by city kids. One evaluator noted that "city kids do not encounter "күл" (ash) as "they live in apartments … so this is a lack of vocabulary, nuance" (IA3). While item 3 was a DIF item and highly marked, GA3 indicated that evaluators believed a crude formatting mistake (presented below) was the most plausible cause of DIF.

There was also debate over item 22 and whether or not examinees would know the word "баобаб"г. (baobab tree) because it does not grow in Kyrgyzstan. Some felt that examinees would not know this word but others disagreed. Two evaluators marked this item as DIF and one as favoring the Russian group. As in sentence completion items 21 and 23, this item did in fact favor the Russian group though the DIF was negligible (.013 effect size). Evaluators noted that an incorrect pair of antonyms in distractor (А) made the item more difficult for the Kyrgyz group. They also noted an inappropriate word combination in item distractor (б) and that the distractors were much longer than the Russian distractors (IA22). In general, none the items with practical DIF were clearly associated with socio-demographic or cultural issues.

Format

There was minimal discussion about the moderate DIF on reading comprehension item 32 which favored the Kyrgyz group. One evaluator noted that the Kyrgyz item wasn't clear but another disagreed and believed there was nothing wrong with the item (GA32). There was a typographical mistake in the Kyrgyz version as one Russian letter was used (which is also a

word) - “и” (and) - but most evaluators did not believe that this would lead to problems (IA32). Two evaluators predicted that the item favored Russians due to the general quality of the reading text in Kyrgyz. Despite the high DIF value, only two evaluators marked item 32 as DIF and only one evaluator offered the specific feedback that Kyrgyz item distractor “г” needed to be more clearly worded (IA32). During the conversation about this item, one evaluator noted that there were structural differences in the way the reading comprehensions items 32, 33, and 36 were phrased in the Russian and Kyrgyz versions. In the Kyrgyz version of the item pairs, respondents were asked to answer a question while for the Russian version the respondents were required to “complete the sentence” (i.e. the stem was not phrased in question form). He noted:

> **MD:** When there is a “question – answer,” it might be easier than when you have to “build” a sentence. In some way this might make it (Kyrgyz) easier to solve than the Russian item but I am not sure about that, the distractors all seem pretty clear (GA32).

It turned out that items 32 and 33 did in fact favor the Kyrgyz group but item 36 was a non-DIF item. A larger study of item formats could test such a hypothesis that the format was affecting DIF levels. There was little evidence however, that evaluators had strong plausible hypotheses for why item 32 was a DIF item.

Three evaluators predicted DIF for item 38 due to format mistakes. Item 38 was correctly predicted by four evaluators to favor Russian examinees. Several evaluators noted that the form of the item stem in Kyrgyz made overall understanding difficult. Several evaluators also noted a format mistake where a nonsense word makes distractor (б) confusing - “токтуу”k. (*no meaning*) should be replaced with “токтотуу”k. (to stop). However, it was a negligible DIF item with an effect size of .011.

Many of the format issues noted appeared not to impact item responses. One exception was item 3. In item 3 a typographical error resulted in a non-sense word in the answer key and

thus no viable answer choice from among the distractors in the Kyrgyz group. Instead of "Чопо"k. (clay), the word "Чоно" (*no meaning*) was written, a one letter misprint which resulted in a total loss of meaning (IA3). Perhaps not surprisingly, this item had both the highest number of marks (6) from evaluators and the largest DIF level from all 38 items. For the Russian version the item had a .64 difficulty level and for the Kyrgyz version .21. Kyrgyz examinees selected from all the distractors in equal proportions. There was clear consensus from the evaluators that the item was highly problematic and that the format error was to blame (GA3). For the other format issues noted, most format or typographical errors such as a different arrangement of the order of distractors or missing letters in certain words did not seem to pose major problems for understanding. Items 35 and 40 both contained format errors but there were both few evaluator marks for DIF and no evidence of practical DIF on these items.

Grammar

Many items allegedly contained Kyrgyz grammar mistakes in individual words. These mistakes consisted of incorrect suffix use (items 10, 14, 15, 18, 26, 28, 29, 30, 31, 37), inappropriate use of compound words (item 17), incorrectly constructed word combinations (items 5, 18, 22, 33), incorrect use of connectors "and, but, because" (items 21, 27, 28), and word choice (items 28, 29, 30, 35, 38). Item 37 contained a grammar mistake in the Kyrgyz stem as "Эмнеде" was used instead of just "эмне" (what). However, most evaluators believed that simple grammar mistakes would not cause DIF on this item and the only mark was one mark for "somewhat similar." This item had negligible DIF with an effect size of .009. In most cases, the item response patterns did not seem influenced when the grammar issue was related to a single word. However, as presented in the above discussion, major syntax issues came up in regard to

sentence completion items which made producing equivalent sentences and ideas quite challenging for several items (GA21, GA22, GA23, and GA28).

Other Issues

A careful review of the data from the rubrics indicates that most analyses centered on discussion of the Kyrgyz (target language) items in the item pairs and issues with their adaptation. As the facilitator, I often asked "what about the source language (Russian) items?" In most cases the response was that the items were clear and correct. However, the Russian version of item 16 elicited some discussion. This item was classified as negligible DIF but had a very high effect size of .031 and favored the Kyrgyz examinees. Two evaluators marked it as a potential DIF item with one evaluator marking it as favoring the Russian group, one favoring the Kyrgyz group. Unlike most items, in which response patterns were similar across groups in terms of order of attractive responses, the Russian and Kyrgyz groups selected different distractors as their first choice on this item (neither of them the answer key). Interestingly, while negligible, this item also had the highest effect size level (.018) for non-uniform DIF.

The stem for item 16 was "author: writer." The same word "author" was used in both the Russian and Kyrgyz versions (Автор). The answer key was "furniture: table." The relationship was supposed to be "class of objects/member of that class." That is, a writer is part of the class or family of "authors" (including authors of scenarios, plays, etc.). However, 42% of the Russian group selected "numeral: digit" as their preferred choice. The Kyrgyz respondents were attracted to neither the answer key nor "numeral: digit," but instead 45% of them selected "journal: book" (журнал: китеп). One evaluator noted that the relationship in the stem could have been construed as a relationship of two synonyms instead of "part of class/family" as was the intent. This would explain the attractiveness of the other choices available.

Other than item 16, there were virtually no other in depth analyses of the Russian (source) items. As noted above, and clearly demonstrated in the rubrics, the overwhelming majority of discussion did not focus on issues within the Russian items themselves that might explain DIF. This item was one of many that favored the Kyrgyz group yet was perhaps the only item where the Russian version received focused attention. I will return to this issue of difference in focus of analysis in the discussion chapter.

Chapter 6: Discussion & Conclusions

Understanding Evaluators' DIF Predictions

In this study I sought an understanding of what bi-linguals were capable of accomplishing in a "blind review" of cross-lingual test items. This is important because a considerable amount of item adaptation and review throughout the world is conducted without the assistance of statistical DIF detection methods. As highlighted above, item evaluators in the KR have minimal (if any) formal training in psychometrics or experience as participants in DIF studies. The predictions of the selected evaluators served as proxies for "the best possible substantive estimates" in the KR due to their previous work experience (Chapter 4). Relative to other DIF prediction studies, a .45 correlation between their ratings of item difference levels and statistical DIF estimations can be considered a relatively high correlation. This indicates that evaluators were able to identify some differences in content, meaning and difficulty between the two item versions that threatened equivalence.

Evaluators were also able to identify problems with item pairs that were a function of the particular languages under study (e.g. agglutination in Kyrgyz led to complications in adapting sentence completion items). These insights into the unique, language-specific challenges of adapting Russian items into Kyrgyz items are also important. However, as presented in Chapter 5, the overall results of the study were somewhat ambiguous. This is because with the exception of the sentence completion items, evaluators were not able to predict which group was favored by DIF with more than chance accuracy. In fact, the overwhelming majority of their predictions were for differences to favor the Russian group. Thus, while the .45 correlation indicates a modest association between what they believed were "different items" with items with high chi-

square difference values, without accuracy in determining which group was favored by DIF it would be incorrect to infer that the evaluators were accurate in "predicting DIF" overall.

Evaluators' analyses focused almost exclusively on the quality of the Kyrgyz items and the challenge of adaptation from Russian into Kyrgyz, especially the sentence completion items (GA21, GA23, GA26, GA28). Virtually no hypotheses were generated as to problems that might lead to items favoring the Kyrgyz group even though the majority of the statistically significant (not practically significant) items favored that group. In this last chapter, I analyze these findings and interpret their implications for key stakeholders, highlight cautions to data interpretation, and provide recommendations for how to improve both item adaptation and DIF prediction accuracy based on lessons learned.

Accuracy in Substantive Item Review

The greatest challenge to evaluator accuracy was their inability to predict the direction of DIF. Of all 32 items classified as negligible, moderate, or large DIF, 18 items actually favored the Kyrgyz group while 14 favored the Russian group. It total, evaluators marked 26 items as favoring one of the groups. Of these 26 items, evaluators marked the Kyrgyz group as favored only twice, and this was done with a single mark in both instances; that is, there were only two individual votes for "favors Kyrgyz" in the entire study. This finding of one-sidedness in prediction of DIF direction is somewhat consistent with another study where DIF by racial group was analyzed in the USA (Engelhard et al., 1990*).* The researchers found that evaluators could not predict which test items would perform differently for black and white examinees when they had no empirical data. They proposed that one reason for the low agreement was the infrequent use of the category "favors blacks." They concluded that perhaps because some reviewers were asked to represent the interests of their race in a high stakes situation, this might have proved

stressful for some of them and influenced their marking. As in the Engelhard et al. study (1990), the category "favors Kyrgyz" was selected rarely in the substantive review. It seems plausible that in many contexts (not just in the KR) reviewers enter DIF analyses with the assumption that DIF and item bias most often penalizes minority or disadvantaged groups. Thus, one plausible explanation for the one-sided outcome is evaluator dispositions.

The overall context of the study plausibly explains these dispositions. Recall the dubious nature of the Soviets' "creation" of the Kyrgyz literary language in the 1920s as presented in Chapter 2 (Hu & Imart, 1989). This process entailed developing a new written language, multiple changes in orthography, imposition of foreign "Sovietisms," in addition to being a highly politicized endeavor in which the interests of the Soviet state were consistently prioritized over coherent or authentic Kyrgyz language development (Grenoble, 2003). Despite initial attention to native language education, the status of the Kyrgyz language was that of a second class language by as early as the end of the 1930s and certainly by the 1950s (Chapter 2). Contemporary attitudes towards Kyrgyz language use have perhaps remained relatively unchanged since independence, despite the improved symbolic status of the language (Korth, 2005).

Considering this historical context and the two troubled decades in Kyrgyz language development since 1991 (Chapter 2), perhaps the tendency to mark almost all the NST items as "favoring the Russian group" should not be so surprising. 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 education in the Kyrgyz medium of instruction in the KR (OSI 2002; Korth 2005; De Young et al., 2006). 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.

This finding underscores the need to conceptualize review of cross-lingual items as a context-bound, social and political process, not simply a technical endeavor. Languages in DIF studies are not simply neutral "variables" but are invested with symbolic social meaning and language politics can be the vehicle through which power relations between groups are mediated. Participants enter into the substantive review process with certain dispositions, prejudices and strongly held beliefs, all shaped by individual experience and social context. In an important sense, this result 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 important social dimensions to DIF studies which rely on substantive review. This social dimension appears manifest in the evaluators' consistent predictions of DIF direction to favor the Russian group.

Of course the one-sidedness of evaluators' predictions of DIF direction may not be solely attributable to evaluator dispositions. As indicated by item evaluators on the rubrics and noted in the historical overview in Chapter 2, one of the main differences between the Russian and Kyrgyz languages is the extent to which they are both coherent, "standardized" systems (Korth, 2005). Whatever the political dimensions of language - and regardless of how incomplete the endeavor to "standardize" Kyrgyz in the 1920s - the early Soviet language planners can not be held responsible for all of the inherent grammatical or syntaxical attributes of a language that make item adaptation challenging.

Indeed, the lack of Kyrgyz standardization and contested nature of what constitutes "correct literary Kyrgyz" kept the focus of most item analyses squarely on the Kyrgyz items. Almost all of the 82 distinct adaption, format and cultural issues raised by evaluators were related to alleged problems with the Kyrgyz language items. Discussions often focused not on the differences in how Russian and Kyrgyz examinees would respond to item differences, but rather on the correct style, grammar, meaning, and dialect 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." By contrast, the Russian language has long-standing, consistent rules and enjoys relative consensus about norms, syntax, grammar, and general use, at least within the context of the KR. 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 evaluators often noted, the Russian items tended to be "quite good" (GA33).

The lack of evaluator experience could also have contributed to the inaccuracy in prediction of DIF direction. The evaluators were not pyschometricians, had no experience with applied statistics in educational research, had no experience with probability models, or as participants in any form of DIF study. The evaluators had no information about the actual statistical DIF outcomes when they filled in the individual rubrics and participated in the group analyses. In only one case (item 32) were evaluators informed that an item was practical DIF, and only after the item had been discussed and characterized as not problematic by evaluators. While I informed the evaluators that the results of their evaluations would be compared to a statistical analysis, none of them had knowledge of how these analyses were typically conducted or what kind of results they could deliver.

It is plausible that their lack of experience contributed to the focus on such overt, Kyrgyz-related issues and distracted evaluators 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 evaluators could check their understandings of the Kyrgyz items. In addition to the high number of Kyrgyz-related item conversations, there was considerable digression away from item analysis into general discussions about the challenges posed by a lack of standardization of the Kyrgyz language in general (GA21, GA23, GA28). Perhaps many issues that could have led to Russian items being more challenging simply went unnoticed *in lieu of* "finding the mistakes" in the Kyrgyz versions.

It is conceivable that to novice evaluators, mistakes and contestation in one language version naturally leads to DIF that disadvantage that group. In other words, the "high quality" (and uncontested) items could perhaps become falsely associated with "advantage" while "lower quality" (contested, more mistake prone) items could become associated with "disadvantage" in the minds of evaluators. The fact the Russian items appeared to be of "high quality" might have led to the assumption that the Russians were favored in most instances where differences were evident. This line of thinking seems plausible when considering the inexperience of the evaluator group with DIF analyses.

Recommendations for Researchers and CEATM

Whether the reasons for inaccurate prediction of DIF direction were due to dispositions or lack of experience, I contend that there is nonetheless some room for optimism that evaluators in the KR can improve their estimations. First, the 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, as presented in the tables in Chapter 5,

evaluator marks on direction of DIF were often more tentative than the marks from section 2.1 (levels of difference). This indecision perhaps indicates that inexperience played as an important role in their estimations as dispositions. Below I propose several steps that could be taken to assess the hypothesis that the evaluators can improve prediction accuracy.

First, as Ercikan (2002) argues, DIF study outcomes differ depending on whether both versions of the items are reviewed simultaneously or individually by evaluators. She notes that when both item versions are presented in pairs, evaluators tend to focus on the comparability of overt issues like format, content, and language use. This seems like an accurate characterization of what transpired in this study. Ercikan (2002) proposes that when reviewers analyze a single item they focus more on the context and issues that might make the item biased for a particular group. In other words, the single item review approach leads to a more nuanced item analysis and facilitates the consideration of the possibility of different cognitive processes among comparison groups (Ercikan, 2002). This kind of approach could lead to a more considered estimation of DIF that favors the Kyrgyz group and with additional research this approach could be readily employed in the KR.

Second, exposure to statistical DIF detection methods by embedding them in some form of action research might also improve evaluators' accuracy. One way to do this would be to conduct several individual item analyses - stop - and then compare the evaluators' preliminary predictions with the actual statistical estimations and discuss the results together as a group. Such an approach would demonstrate the complex and tenuous nature of DIF prediction and interpretation to the novice evaluator. It would show that the language group with the lower average test score is not always the disadvantaged group at the item level. It would become more apparent that mistakes do not always lead to DIF; neither in the language where mistakes

occur nor in the other language involved. Finally, it would underscore the need to think deeply about the differences between item versions before predicting the direction of DIF. This kind of fine tuning and skills enhancement through the introduction of statistical methods holds promise for better analyses in the KR. With the increasing availability of on-line software and the option of relatively inexpensive statistical packages, the employment of statistical DIF detection methods is feasible in the KR in the near future.

In addition to employing statistical analysis as part of the item review process as highlighted above, there are other ways statistics can be used to improve DIF prediction processes. While predicting DIF is difficult on average, there is evidence that some reviewers are more accurate than others. Engelhard et al. (1990) discovered considerable variability across reviewers in the correlations between their individual marks and statistical DIF. Estimations of individual reviewer accuracy can be used both in training and as a quality control tool for CEATM when selecting reviewers to participate in item analyses. Individual estimations were not computed for this study but could be done in future work by CEATM with the consent of evaluators and CEATM employees.

Understanding the Causes of DIF

The second purpose of this study was to gather data about causes of DIF that could inform and improve the item adaptation process in the Kyrgyz Republic. One finding from this study was that in order to understand DIF causes, it was necessary to analyze the minutia of each item: Broad, categorical labels such as “translation differences” didn’t capture the nuances necessary to provide a real understanding of what was causing DIF in a particular item. For example, most of the *loan word*, new word, or socio-cultural issues projected as potential DIF

causes (items 2, 7, 9, 17, 22, 25) did not lead to DIF.[70] Some translation and adaptation problems did (items 13, 19) while others did not (items 10, 15, 18). Item 3 contained an obvious format problem that plausibly led to DIF while other items with format issues (items 35, 38, 40) remained unaffected. Incorrect Kyrgyz grammar at the word level did not seem to cause DIF in most cases but the adaptation of entire sentences in the sentence completion items was characterized as a dubious endeavor by all the evaluators (GA21, GA23, GA26, GA28).

Tenable hypotheses about causes of DIF were articulated when evaluators were able to breakdown the minutia of the item under review. The *location* of the difference or problem within the item, and the extent to which the difference impacted meaning or difficulty level across versions was paramount. Nuances that resulted in differences in key parts of words, phrases and sentences were important: "Key parts" meaning the place in the item where essential meaning is located (Ercikan, 2002). For example, essential meaning in analogies items is located in the item stem. If the stem is muddled in one version and the logical relationship between the pair of words in both versions not the same, the differences between the items are plausibly going to be problematic (IA19).

The same is true for the answer keys. If differences between item versions result in no viable answer key for one version, DIF is also highly possible (IA3, IA13). However, if there is a format or small translation mistake in a distractor that was obviously not plausible to begin with, this issue might be less likely to cause DIF. In this study, three of the four practical DIF items had serious issues with either the answer keys or item stems (GA3, GA13, and GA19). Thus, causes such as poor adaptation, translation and format problems, incorrect grammar, and questionable cultural comparability are best conceptualized as general constructs. They are

[70] With the possible exception of item 23 (negligible item but with a high r-squared delta), where the debate was over whether "insurance" was a known phenomenon (GA23).

useful primarily as organizing principles for data analysis, not as DIF causes *per se* as they are less meaningful terms outside the context of a particular item (Ercikan, 2002).

It is also clear from this study that mistakes in items do not necessarily result in DIF. Previous research has shown that while item evaluators are quite good at locating mistakes in test items, this is not the same thing as successfully predicting DIF (Engelhard, et al., 1999). Item evaluators in this study identified mistakes overwhelmingly with the Kyrgyz versions though not every mistake noted was widely agreed upon. However, the majority of these mistakes were not associated with statistical DIF and in many cases evaluators were divided as to whether they would or would not lead to DIF. For example, recall that half of the statistically significant reading comprehension items favored the Kyrgyz group despite the fact that the evaluators reported problems exclusively with the Kyrgyz text. As noted in Chapter 5, it is perhaps possible that mistakes in Kyrgyz items in non-essential locations (less plausible distractors for example) might have actually favored that group as such mistakes reduced the number of plausible answer choices (IA11, IA15, IA18, IA25, and IA33).

However, as no hypotheses were generated for why any of the items might favor the Kyrgyz group, it is difficult to suggest hypotheses about what caused some items to favor that group. While three of the four practical DIF items favored the Russian group, there were many items close to the moderate DIF cut-off that favored the Kyrgyz group. A tentative explanation for why many items favored the Kyrgyz group might be related to issues with word difficulty. For example, when words are adapted from the source language, they can become easier due to a lack of corresponding vocabulary at the same difficulty level in the target language (Schmidt & Belistein, 1987; Bejar, Chaffin & Embertson, 1991; Roccase & Moshinsky, 1997; Sireci & Allalouf, 2003). Recall that for some items in this study, there were single Kyrgyz words that

are differentiated by several different words or concepts in the Russian language: That is, the Russian language might have finer degrees of distinction for some concepts and some of these distinctions might have an impact on item difficulty.

For example, the word for orchard and trees is the same word in Kyrgyz while in Russian there are different words for these concepts (IA3). It could be that such distinctions make the use of some words equivalent in meaning but divergent in difficulty level due to differences in commonality of use. Such differences are not overt and they are not easy to identify without deep probing and analysis. Of course as there were no hypotheses generated about why any item might favor the Kyrgyz group in this study, this is conjecture at best. It is interesting to note however, that nine of the eleven negligible DIF analogy items (not practically significant) did favor the Kyrgyz group. It is possible that word difficulty could be an issue for these particular language groups on analogy type items. Unfortunately, the data allow no more than tentative hypotheses about the issue at this time.

The evaluators recognized problems in three of the four practical DIF items and predicted their DIF direction correctly. However, for the one practical DIF item that favored the Kyrgyz group, item 32, a conclusive determination of the cause of DIF remained elusive as no widely agreed upon hypothesis was offered to explain the DIF. Only the sentence completion items tended to favor the Russian group on a consistent basis. Most of the evaluators' specific hypotheses about key differences between Russian and Kyrgyz items were generated in discussions about these items despite the fact that none of these items were practical DIF (GA21, GA23, GA26). They were successful in their predictions of DIF direction however, 71% of the time for sentence completion items. In short, as in previous studies, it is apparent that DIF causes overall are not always easy to identify but as this study indicates there may be variation in

success rates by item type (Plake, 1980; Engelhard et al., 1990; Rutledge, 1990; Gierl et al., 1999; Jodoin & Gierl, 2001; Ercikan & McCrieth, 2002).

Despite the above qualifications about our limited ability to generalize about DIF causes beyond the *location* of cause within each item, the data do point to a few recurring patterns in regard to item type. First, the reading comprehension items demonstrated the lowest negligible DIF levels and generated the least amount of critical commentary. This result is consistent with several other studies of verbal reasoning items noted in Chapters 3 and 5. Second, one actionable finding from the study was the issue of the lack of "linguistic fit" of Russian and Kyrgyz sentence completion items.

Specific hypotheses about the Kyrgyz versions of the sentence completion items were clearly articulated and widely supported by evaluators (GA21, GA23). These items elicited both the most commentary and the most accurate DIF direction predictions on the part of evaluators. Evaluators even found a few of these items difficult to answer themselves without being able to reference the original Russian item (GA26, GA28). Though none of the practical DIF items were sentence completion items, these items were clustered around the highest chi-squared and effect size values. Items 21, 22, 23, and 26, were some of the most problematic items from the entire 38 items according to the evaluators. All four of them statistically favored the Russian group. Sentence completion items 25 and 28 were also negligible DIF with high effect size values.

The problem with the sentence completion items was related to the fact that Russian and Kyrgyz syntax was not easily reconcilable within the context of these items. Recall that syntax is the body of rules in a given language that determine how words and phrases come together to form grammatically correct sentences. The syntaxical clash between Russian and Kyrgyz

manifested itself in Kyrgyz items as incorrect use of compound words, incorrect word combinations, artificial "Russified" sentence structure, and general confusion of Kyrgyz items. Evaluators consistently noted that the items failed in Kyrgyz because they were being "forced" into a Russian syntaxical style that did not work (GA21, GA23, GA26, GA28). Unlike the reading comprehension items, that allow for a more natural flow of language due to the absence of constraints on text size (the Kyrgyz reading comprehension text is more than twenty lines longer than the Russian text), the sentence completion items must be short and concise by definition. They require examinees to make logical connections by filling in the missing words that makes the sentence(s) most logically complete. They consist of one or two sentences at the most but the sentences must be relatively long.

Recall from Chapter 5 that evaluators believe that *agglutination* keeps Kyrgyz sentences short and not conducive to the longer kind of sentences necessary for the sentence completion items (GA21, GA23, GA26). It is hard to imagine sentence completion items with sentences of three to four total words. Yet, such short sentences are common in Kyrgyz (GA21, GA23). If it takes more (shorter) sentences to convey the same meaning and level of complexity in one language than another, it makes intuitive sense that item types that allow only one or two sentences become problematic for adaptation. This finding is consistent with other studies that found that DIF can be caused by differences in sentence structure that are inherent to the language under study (Gierl et al., 1999).

Recommendations for Researchers and CEATM

While the claim made by some evaluators that Russian syntax is inherently "more difficult" is perhaps questionable, it is certainly plausible that specific syntax differences can create specific challenges for certain item types. Evaluators made specific proposals for

rectifying this state of affairs with sentence completion items. They proposed that the long Russian sentences be broken into shorter (but more) sentences in the Kyrgyz versions (GA21, GA22, GA23, GA25, GA28). Evaluators also proposed that if breaking up items into smaller parts was not feasible, instead of adapting these items from Russian, in the future the test center should create them either separately (Kyrgyz and Russian items) or create them in Kyrgyz first and then adapt them into Russian; or, perhaps not use this type of item at all.

CEATM notes that bi-linguals play an important role in item development and adaptation and that procedures are in place to guarantee item equivalence throughout the test item development cycle (see Chapter 4). The overall low amount of DIF detected in this analysis supports the contention that these processes are working well. However, in their item reviews, evaluators noted on several occasions that it seemed obvious that Kyrgyz language items were developed in Russian, or by "Russian thinkers," without enough consideration for the authenticity of the Kyrgyz version. For example, in regard to item 21, one evaluator stated, "It seems that this item was obviously adapted from Russian. I think if Kyrgyz original texts had been used, there wouldn't be this problem. We could avoid syntax problems like this" (GA21). There was wide agreement among evaluators about this contention. Similar comments can be found in GA23 and GA28. In GA33, one evaluator noted, "There are many problems with the translation of this difficult text… one resolution is to take an original Kyrgyz language text, related closely to this theme and the Russian text…" In the IA (reading comprehension text) evaluators also requested more Kyrgyz original texts.

At a minimum, the evaluators' comments merit reconsideration of current adaptation procedures, especially considering the issues raised in regard to the sentence completion items. Hambleton and Kanjee (1995) recommend "de-centering" the item development process in

challenging cross-lingual situations. In the context of item adaptation, "de-centering" means providing opportunities to return to the source item version (Russian in this case) and making changes to that source item if necessary due to challenges in adaptation to the target language. In GA15 one of the evaluators proposes: "Perhaps it would be possible to compare the translated Kyrgyz text with the original Russian text? That is, adjust the Russian text again if the translation into Kyrgyz does not seem to work?"

Solano-Flores (2006) recommends what he calls "concurrent development" of test items. While his work focuses on English and Spanish speakers in the United States, several of his ideas are relevant to other cross-lingual contexts. He proposes that all test items be developed exclusively by bi-linguals. This forces test developers to seriously consider how culture and context are inextricably related to language. In some of his work, he has utilized two groups of bi-linguals to concurrently develop the two versions of a given test item. Through this process, modification of items becomes an iterative, negotiated endeavor that does not proceed without consensus. All recommendations for changes to one version of an item are only considered after the proposed changes have been analyzed in relation to how they will impact the other language group.

He also recommends the use of "blueprints" or general item guides (like mini-specifications) to mediate discussion around each item. Through the process of "localization" bi-lingual test developers work from these blueprints but have considerable freedom in adaptation in order to facilitate linguistic alignment between the two versions. The result is that some items will inevitably be slightly different versions but ultimately serve the same aims. Solano-Flores (2006) insists that "localization" in item development is essential as research has demonstrated that even bi-linguals do not always have consistent or accurate perceptions of all the linguistic

aspects of items that are critical to properly understand their functioning. Further, simply being a native speaker does not necessarily enable evaluators to identify those key aspects.

Hambleton and Kanjee (1995) also offer a useful method for determining comparability of items that could be employed in the Kyrgyz Republic during item development and analysis. They propose utilizing examinee interviews to determine the cognitive processes items elicit as examinees engage with items. For example, Kyrgyz language examinees could explain their reasoning for answering certain ways on the sentence completion items. Judges would follow along with both Russian and Kyrgyz versions on hand and compare how well the items capture similar meaning and constructs. If the responses of the examinees correspond to the intent of the source item (Russian), then the items can arguably be considered equivalent. While labor intensive, this type of analysis could be performed as a follow-up for items that seem to be problematic according to DIF statistics or other forms of analysis, not necessarily for all items. In the case of the NST 2010 items, such follow-up analyses for the sentence completion items could be fruitful. While not done formally for this study, this kind of individual interview could also be conducted with item reviewers on problematic items.

While the above suggestions for possible modifications to item development procedures seem reasonable, they of course must be realistic in terms of resources available to invest in item development. The government of the KR currently does not provide CEATM with financial support. CEATM resources are generated through student fees for test services (approximately 4-5 US dollars per test per student on the NST). The use of item development groups that employ multiple levels of review and other elaborate iterative processes is a labor intensive enterprise that demands a significant time and resource commitment. Thus, CEATM must

carefully consider both what can learned from “best practices” in cross-lingual item development as well as resource realities.

Another important issue is how to distinguish between an “adapted” item and a “new” item altogether. For example, would breaking the Russian sentence completion items into smaller sentences constitute appropriate adaptation or the creation of completely different items with different meaning and difficulty levels? If new items are utilized, replacing one item type with another does not absolve CEATM of the need to employ items of similar aim, meaning and difficulty if they intend to make comparative inferences across groups. Testing practitioners and policymakers in the KR should be sensitive to this challenge and conduct further research, invest in training of reviewers, and experiment with different test item types to the greatest extent possible. The above findings in regard to the challenge of sentence completion items are also relevant for test developers in neighboring countries that also develop standardized tests in the Russian and other Turkic (agglutinative) languages such as Uzbek and Kazakh.

Statistical DIF and the NST Verbal Items

Not all multi-lingual countries provide opportunities for education through multiple language media or cross-lingual testing in high stakes situations. In many Asian countries, pupils and students are schooled in and sit for examinations in their second language. Hambleton and Kanjee (1995) argue that one of the main benefits of cross-lingual testing is the elimination of bias that potentially exists when students must sit for examinations in a language that is not their native tongue. In multi-lingual societies therefore, cross-lingual testing is potentially a good policy option for a variety of uses if inferences on cross-lingual tests can be validated.

The identification of only 4 of the 38 total items as practical DIF on the NST items is a relatively low number for a cross-lingual assessment: Other studies with the same approximate number of items have revealed that up to half the items are often flagged as DIF (Chapter 3). This is a positive result for both CEATM and higher education admissions policy makers in the KR. While the analyses included only a portion of the total number of NST items, policy makers and stakeholders concerned about the feasibility of employing cross-lingual testing in HEI scholarship selection now have empirical evidence that supports the inference that CEATM administers a test with a very high number (proportionally) of equivalent items. The low number of practical DIF items indicates that CEATM has done a reasonably good job utilizing the available linguistic and cultural resources and suggests that the bi-linguals employed are reasonably effective at developing equivalent cross-lingual test items. If the test center can incorporate statistical methods to assist with DIF detection and improve the item adaptation process, it can feasibly further improve the reliability of the NST and enhance the validity of selection inferences based on the NST.

The overall low number of practical DIF items, the modest correlation between substantive review and statistical DIF (in terms of difference levels and chi-squared values), and the relative ease with which evaluators identified some causes of DIF are all reasons for cautious optimism. The overall low number of DIF items is perhaps best explained by the nature of the NST as a *within country* cross-lingual test. The large number of bi-lingual and bi-cultural scholars, teachers and adapters, readily available to the test center means that the cultural distance between groups is relatively small: Differences in schooling, curricula, instruction, and other intervening cultural and linguistic variables that can impact DIF levels across groups can "be known" quite easily in the KR (Ercikan, 2002).

Yet, despite low levels of statistical DIF, the results of this study for the item evaluators are not straightforward. In the previous section I noted that dispositions, characteristics of the Kyrgyz language, and evaluator inexperience all plausibly impacted evaluators' inaccurate prediction of DIF direction. If the original item developers for the NST 2010 come from the same general population of item reviewers employed in this study in terms of experience and training - and CEATM believed that they did - one might expect if not accuracy, at least some element of "randomness" to their predictions of DIF direction overall, not the one-sided estimations - "favors Russian" - across almost all the items.[71] The paradox seems to be that while the cultural intimacy of the *within country* study in some ways makes cross-lingual testing more feasible than in broader cross-nation comparisons, there appears to be an added dimension of sensitive language politics (and subjectivity) when the research touches on sensitive questions such as "who benefits from item differences?" While this was not an anticipated result of this study, it was not too surprising considering the context of the DIF study and the history of Russian and Kyrgyz language politics in the KR.

Cautions to Statistical DIF Interpretation

The logistic regression method proposed by Swaminathan and Rogers (1990) with the effect size measure proposed by Jodoin and Gierl (2001) yielded clear and interpretable results. As the purpose of this study was not to evaluate the accuracy of various statistical DIF detection methods, the actual number of DIF items detected by the logistic regression method was not of primary importance. However, there are some important qualifications to the interpretation of

[71] Recall that many of these evaluators did have experience working with CEATM on NST test adaptation in previous years (see chapter four for a breakdown of professional background and experience in testing).

the statistical findings related to important contextual factors that could impact statistical outcomes. In the next section I elaborate on these qualifications in detail.

First, statistical methods in DIF studies are not 100% accurate in detecting DIF (Hambleton, 1995). The logistic regression (LR) method - while comparable to other DIF detection methods in accuracy - has had power rates of between 70-80% in experimental studies with various combinations of ability levels, item types, item characteristics and sample sizes (Jodoin & Gierl, 2001). Thus, in any given study relying on the LR method, it is feasible that some DIF items could remain unidentified, though a large sample size like the one employed in this study should lead to a relatively high success rate.

There are other factors however, that could threaten the accuracy of the statistical estimations. The lower reliability of the Kyrgyz NST items and the large difference in ability distributions between the Russian and Kyrgyz populations could introduce statistical error (Narayanan & Swaminathan, 1996). Differences in item characteristics could also pose a challenge to accurate estimation. Hambleton et al.'s (1993) study indicated that items with lower discrimination were associated with items likely to be missed in some DIF detection methods. They also found that very difficult items were more likely to be missed, regardless of ability level. The researchers indicated that this is especially true for DIF studies in which comparison groups have dissimilar ability distributions. Upon request, CEATM provided the test item characteristics for the 2010 items. The average discrimination value for the Russian items was .45, while for the Kyrgyz items it was .32. The average difficulty level for Russian items was .54 while for the Kyrgyz items it was .33.

Another important issue is knowledge of the Russian language on the part of some Kyrgyz language examinees. Variation in the Kyrgyz population in terms of how much Russian

they know could be influencing statistical results in hidden, unpredictable ways. Recall that evaluators noted on several occasions that they were only able to solve some Kyrgyz items because they knew Russian or had the Russian item available (IA21, GA21). They also noted that Kyrgyz examinees with Russian knowledge might be advantaged. This raises perhaps one of the more viable threats to DIF studies in the KR in general. Despite the fact that schooling is not bi-lingual by design, as shown in Chapter 2, bi-lingualism is common in Kyrgyzstan.

Knowledge of Russian can be acquired through study as a second language at school or through the news media, social or cultural engagement with Russian speakers in everyday activities. In 1989, 83% of all urban Kyrgyz, 23% of the urban population, reported fluency in Russian (Fierman, 1991). Among ethnic Kyrgyz who are schooled in the Kyrgyz language, there is tremendous diversity in terms of Russian knowledge. It is possible to be educated in a Kyrgyz language school but also live in a community where Russian is widely spoken. Northern Chui Valley communities like Sokuluk, Kant, Tokmok, Kemin and Kara-Balta as well as many towns in the Issyk-Kul *Oblast* have large numbers of both Russian and Kyrgyz speakers (Census, 2010).

Knowledge of Russian might favorably impact some Kyrgyz language examinees as they struggle to decode incoherent Kyrgyz items that were initially developed in the Russian language (Ackerman, 1992). Recall that at times item evaluators were at times able to identify the "Russian thinking" behind some items - some examinees might be able to do the same. Another way of conceptualizing the problem is to consider the difficulty of defining the "typical Kyrgyz language examinee." When evaluators offered, "Kyrgyz kids won't know this…", perhaps they were envisioning mono-linguals in the farthest outlying regions of the country; those with almost no exposure to the Russian language or the "Russianisms" in the Kyrgyz language used by many

urbanites. However, there are other "typical Kyrgyz examinees" that do have at least some knowledge of Russian, if not very good functional command (Korth, 2005). Thus, this "mixing" of the constitution of the Kyrgyz language sample in terms of background knowledge means that the overall statistical outcomes could be hiding some of the real impact of particular item issues for certain subgroups of the Kyrgyz population. Higher statistical DIF levels might be evident if all subjects in the study knew one and only one language. This kind of hypothesis could be tested with additional research if reliable data on knowledge of Russian as a second language could be attained.

In order to probe for how background Russian knowledge might have impacted the DIF statistics, I conducted one additional statistical analysis. Because student test identification numbers were tied to their region of registration, it was possible to determine from which region each examinee came from. I conducted an additional DIF analysis by breaking the Kyrgyz sample into two groups of about 750 examinees each, Kyrgyz 1 and Kyrgyz 2. In one group I put all of the Bishkek (capital city) examinees and in the other, primarily rural examinees. In essence, I used residence as a proxy for language knowledge under the assumption that Kyrgyz language examinees from Bishkek would be more likely to have Russian language knowledge.[72]

In theory, all 38 items should have shown "no statistical DIF" when analyzed as groups Kyrgyz 1 and Kyrgyz 2. The results however, were interesting: Thirty-three of the thirty-eight items were indeed "no statistical DIF" (full statistics in Appendix V). Four of these items were just barely significant at the .05 level, test statistic 3.814: item 5, chi-square value of 3.92 (sig., .048), effect size .003; item 3, chi-square value of 4.98 (sig., .026), effect size .004; item 18, chi-

[72] Though it is not possible to know which kids who sat for the NST have this language background, Census (2010) data support the contention that Kyrgyz urbanites are more likely to know Russian as a second language than Kyrgyz rural residents in general.

square 5.03 (sig., .025), effect size .003; and item 30, chi-square 5.72 (sig., .017), effect size .005. Item 21 however, the one Kyrgyz item that was unanimously and repeatedly claimed to have been created "by Russian thinkers" had the highest level of negligible DIF from all thirty-eight items. The chi-square difference value for this item was 12.68 (sig., .000), effect size .010. This analysis was of course only an investigative probe. In order to further investigate the possible impact of Russian knowledge on Kyrgyz item responses, experimental studies with more accurate data on language background of the participants needs to be conducted.

Finally, while Jodoin and Gierl (2001) and Zheng, Gierl, and Cui (2004) have empirically analyzed the r-squared delta effect size measure employed in this study in both experimental studies and with actual assessment data, it has not been widely tested.[73] In light of both the potential threats to statistical estimation highlighted above, and the relatively untested state of the effect size measure, the statistical outcomes in this study need to be viewed somewhat cautiously. I recommend a more conservative interpretation that acknowledges that some items not identified as practical DIF by the LR method might nonetheless be problematic.

Recall that the positive rank order correlation indicates that negligible DIF items with higher chi-square values were more closely associated with higher evaluator marks than the lower value chi-square items. Further, substantive data collected from evaluators indicates that several "borderline" negligible DIF items might be problematic. One purpose of relying on substantive evaluations (despite their modest reliability) is to provide additional confirmatory evidence of differences between items in the studied pairs (Gierl et al., 1999). Four of the negligible DIF items had effect size values above the median and at the same time received five

[73] In these two studies the r-squared delta effect size measure correlates highly with two other commonly accepted effect size measures used for the SIBTEST and Mantel-Haenzsel DIF detection methods.

or more marks as likely DIF from the evaluators (i.e. they were predicted as DIF by the evaluators).

Recall also that the initial cut-score for evaluator DIF prediction was four votes for DIF. However, this scale was developed with the assumption that the scores from ten evaluators would be utilized. After dropping two evaluators, I nonetheless maintained the four mark cut-score because that was the original scale. However, one could argue that three marks from eight evaluators for DIF is also a reasonably strong vote for DIF. Further, several items receiving three marks for DIF also had negligible DIF well above the median effect size value. With this in mind, I recommend that CEATM consider several other items near the practical DIF cut-off as potentially problematic.

Table 6-1 below presents additional items for CEATM to consider. The criteria for their inclusion was that each item had a score of three or more evaluator marks for DIF and the negligible DIF value for each item was over the over the median value of .009. In addition, there were four other negligible DIF items with very high effect size values that received minimal marks from evaluators that are not in table 6-1 below. These items are item 22 (.013 effect size, rank order 27), item 26 (effect size .019, rank order 28), item 28 (effect size, .029, rank order 34), and item 16 (.031, rank order 31). These four items could also be considered "borderline" DIF items worthy of investigation. Six of these additional twelve items are sentence completion items, the only item type in which the evaluators' accuracy in prediction of DIF direction was better than random.[74]

[74] I emphasize that I am not trying to "find practical DIF" where it doesn't exist. I do believe that the reliance on bilingual test adapters could make within country, cross-lingual DIF levels lower than typically seen on across country studies like PISA and TIMSS (despite their greater methodological sophistication). CEATM does have item review procedures in place and their specialists make great efforts to produce quality test items. However, the sum of the accumulated

Table 6-1: Items Above Median Effect Size with Three or More DIF Marks

Item	Evaluators' Marks	χ^2 Difference	χ^2 Rank Order	Effect Size
10	xxx	15.510	19	.009
38	xxx	20.210	23	.011
5	xxx	22.576	25	.015
25	xxxxx	23.006	26	.016
23	xxx	38.703	23	.019
21	xxxxxx	42.413	30	.024
33	xxxxx	43.427	32	.027
11	xxxxx	49.326	33	.028

Of course the identification of additional "borderline DIF" items is somewhat arbitrary, especially in terms of where to draw the line. For example, the test center might want to also re-examine item 15 that received many marks from evaluators but had an effect size value lower than the items noted above. In general, the best way to further test the accuracy of the statistical estimations in this study would be employ multiple methods of DIF detection on the items and then compare the results across methods (Rogers, 1989; Hambleton, 1995; Jodoin & Gierl, 2001).

Recommendations for Improving Studies of Substantive Methods

In this last section I provide recommendations for future substantive DIF prediction studies based on lessons learned from the administration of the evaluation rubrics. Understanding the limitations of the data collection process will help to put the results in perspective. The first issue is related to the quality of the data collection tools themselves. The original item analysis rubric contained an element of unnecessary complexity. I developed a

---

evidence, including the threats to accurate measurement noted above, indicate that the underestimation of practical DIF in this study is possible.

coding scheme that asked each evaluator to code "the nature of the difference/issue" as definitively as possible (section 2.2 of the rubrics). In general, the *a priori* categories of potential item adaptation problems were useful as a guide to help evaluators think about the items. However, asking them to "over think" in this area proved counter-productive. After conducting a pre-test, it was apparent that the coding categories were somewhat problematic. Though the issue was addressed before administering the rubrics, it is worth highlighting so future researchers can avoid adding unnecessary complexity to their data collection tools.

The first problem was that the descriptive typologies in 2.2 were not always mutually exclusive or easy to disentangle. The result was that during the pre-test the evaluator lost time trying to distinguish between adaptation issues and translation issues when she could have been describing a problem with a specific test item in greater detail. Second, the purpose of having the substantive review with ten members was to collect a broad spectrum of opinions as to the nature of the problems with the item pairs. As individuals, the evaluators see only their own work in isolation. It is the researcher who should collect and categorize their work and present it in summative format, drawing on the opinions and conclusions of all ten evaluators. The larger purpose of the individual evaluations was not to determine how consistently each evaluator precisely defined each problem, but rather to get a basic understanding of what the differences were between the language versions. In other words, it is not so much the individual's marks that matter as much as the totality of the collective whole which better represents their professional guidance.

The utility of the data from section 2.5, suggestions for improving the item pairs, was also questionable for a limited study like this one. While interesting, it demanded that the evaluators take significant time away from item analysis and description and devote that precious

time to what in essence became "item writing." Perhaps due to time constraints, this was not filled in diligently in most cases anyway. It seemed redundant as evaluators often "corrected" the item when they wrote their descriptive comments under section 2.3. Many of the comments in this section amounted to not much more than "next time, do a better job with translation." Overall however, as can be seen from the data in Appendix W, considerable data was collected from the individual analyses from sections 2.1, 2.3 and 2.4, as well as the group analyses. In the future I would recommend that evaluation protocols require each evaluator to only (1) assign a mark as to the level of difference (section 2.1), (2) describe the differences in detail (section 2.3), and (3) predict which group was advantaged (section 2.4).

Finally, the statistical analyses of inter-rater reliability and the rank order correlation were both calculated with data taken from the individual analyses: That is, the initial marks evaluators made on their individual item rubrics. While not a critical mistake, the problem was that the benefits of the group analysis were not reflected in two of the key data analyses in the study. Evaluators did not have the opportunity to revise their original marks after learning more about the items through the group analysis (though the discussion transcripts reflect some newly generated hypotheses). The failure to do this was related to time and resource limitations.

If I assume that the group analysis actually assisted in generating a more accurate understanding of the items - and I believe it did - it is possible that (1) inter-rater reliability would have been higher, (2) the rank order correlation between their predictions and the chi-squared values would have been higher, and (3) perhaps their estimations of which group was favored would have been more accurate. Anecdotally, I am confident that the marks of several individual analyses do not reflect the evaluators' post-discussion predictions of DIF for moderate and high DIF items 13 and 19, neither of which had been predicted as DIF by the majority of

evaluators initially.[75] Evaluators predicted DIF for item 3 but not for item 32 – neither before group discussion, nor after.

This recommendation assumes of course that the impact of the group analysis would be to improve the accuracy of their estimations. It is theoretically possible that as a group, the accuracy of their estimations could get worse rather than better after group analysis. After all, there were persuasive personalities in the evaluator group who were not always personally correct in their predictions of how differences would impact DIF levels. So, peer pressure can certainly push group results in different directions. Nonetheless, in future studies I recommend adding the additional step of evaluators individually rescoring each item (section 2.1) after the group analysis has been carried out. In general, in order to conduct a more informed DIF prediction study, I would recommend twice the time that was allocated for this study as 10.5 hours was not enough.

Challenges to Collecting and Interpreting Data from the Substantive Review

Evaluation of cross-lingual test items as an individual process is influenced by the knowledge, experience, skills and dispositions of individual evaluators. Item evaluation as a collective process is influenced by the above factors plus the social dynamics of setting and context. Time and resource constraints also limited the amount of discussion that occurred for every test item. Group analysis was complicated by the fact that unlike in individual interviews, not all side-bar conversations, comments, and issues could be fully captured. Sometimes, many participants talked at the same time. The item discussion process, while recorded, was a

[75] In a private conversation with one evaluator, I learned that the reason why these two items (13, 19) received so few marks initially was because evaluators couldn't answer the items with confidence themselves. Thus, paradoxically, low marks (or "absence" of marks) can also indicate item trouble when evaluators struggle to make sense of the item and don't know how to respond! More evidence for why it is imperative to rescore after group analysis.

vociferous and at times, a muddled affair. However, I contend that the nature of this study - and its focus on language - is enhanced, not limited by the inevitable "negotiation" that accompanied data collection. In matters of language, the collective view, even if contested by some, is more accurate than the unopposed view of one, on average, most of the time (Hambleton, 2005).

Selecting what item data to highlight from both the individual rubrics and group analyses was an interpretive process. The researcher always influences data collation by selecting what data to present or not present, by proposing what is representative, informative, relevant or irrelevant, and in general by making claims as to what is worthy of attention. Capturing and faithfully representing the tone, focus, agreement and disagreement in conversations about test items was challenging: When an evaluator highlighted a certain issue that came up during discussion, how does the reader know the extent to which the rest of the group concurred, disagreed, or was divided on that issue? This challenge was there for both the interpretation of the individual analyses as well as the group discussion. In part, the coding system on the individual rubrics was there to "empiricize" the process to the greatest degree possible; it sets boundaries to interpretation and serves as the voice of participants to some extent. Nonetheless, the subjective stance of the researcher inevitably came into play in the presentation and interpretation of the raw data.

Much DIF research is also complicated by the fact that the subjects themselves (item evaluators) are also highly subjective respondents with their own biases and proclivities. While their may be "strength in numbers" when ascribing validity to claims and hypotheses, in DIF studies, the majority view can also be the wrong view: Determining "what was correct" turned out to be a highly contested endeavor. What for some were highly problematic items, for others, were not. In the results chapter I tried to faithfully distinguish between opinions that seemed to

be representative from those that might be outliers. When generalizing I have tried to note exceptions, limitations, or any problems that might have challenged my interpretations and inferences.

Interpretation of data from the group analysis was especially challenging. At that point in the study, I moved from the role of data collector to that of participant as the discussion facilitator. I took on the roles that facilitation typically requires: Time keeper, task manager, referee, in addition to observer and recorder of the proceedings. My facilitation of the process inevitably impacted the data collection process and hence the data itself through my choices of what merited discussion and how much time to allow per item. Without such facilitation however, the collection of this kind of data is not possible. In some ways, the "outsider" is perhaps in a good position to serve as a facilitator participant when the stakes to evaluator participants are connected intimately to language and identity. As an American whose native language is English, not Russian or Kyrgyz, I was able to maintain some distance from sensitive questions in the data collection process.

Finally, what was perhaps most attractive to collect and report was data from items that elicited much conversation, items that were heavily critiqued, presented contradictions, or items that seemed to represent a systemic problem, issue, or challenge. Some items elicited few written comments and little discussion. While it was also important to understand why these items did not elicit commentary - or perhaps understand *why they worked* - for the most part, the focus centered on *what didn't work.*

Conclusion

Of paramount importance in the cross-lingual test adaptation process is the proven ability of test developers to successfully adapt test items across languages in meaningful ways. In

situations where sophisticated statistical DIF detection methods are not utilized, the accuracy of item adapters and reviewers in discerning differences between items is especially important. In some ways, the results of this study are ambiguous. Evaluators' marks were positively correlated with statistical DIF outcomes in terms of which item pairs had differences that made them problematic. At the same, based on evaluators' inaccuracy in estimating which group was favored by group differences, it is difficult to discern with certainty just how well they actually understood the differences in item pairs.

An interesting finding of the study was the consistency with which all but two of their predictions were for item differences to favor the Russian group. As has been pointed out, the evaluators focused on Kyrgyz language items but generated no hypotheses for causes of DIF that favored the Kyrgyz group. Thus, in a sense, the prediction of DIF direction was not random at all, but could be perhaps more accurately characterized as "one-sided." I offered three explanations for this result. First, many Kyrgyz language issues are highly contested; therefore, it is natural that much attention would be paid to these items. Second, evaluators had no experience with DIF studies and the complex task of disentangling DIF causality. Third, social and historical contextual factors likely shaped the evaluators' dispositions to the extent that it was almost taken for granted that "of course the Russians are favored." As I argued above, I believe that two out of three of these issues, experience and dispositions, can be addressed by employing statistical methods in further studies. My recommendation that additional items (beyond the four identified as DIF) be carefully reconsidered is based on the uncertainty of the statistical estimation in the context of large differences in ability distribution, relatively lower Kyrgyz test reliability and the possibility of the nuisance factor of language knowledge on the part of some Kyrgyz language examinees.

From the perspective of the test center, three of the four items with practical DIF can be addressed because they appear to be related to overt adaptation issues. However, many problems with the Kyrgyz language items would be out of their control, even if some of these issues caused DIF (though most did not). The privileged status of the Russian language, the lack of a "standardized" Kyrgyz language, the lack of investment of resources for its study in general (or poor use of those resources), frame the contextual setting for test development and Kyrgyz item quality in the republic. Everyday uses of hybrid Kyrgyz, dialect and regional differences in vocabulary and knowledge of loan words are all socio-linguistic issues that can not be easily controlled by political volition.[76] These phenomena are the product history, culture, and demographics. The unique geography of Kyrgyzstan with its mountain barriers, isolated communities, and variation in the extent of engagement with other language groups has resulted in the evolution of many unique language system(s) which will continue to present challenges for those developing cross-lingual testing, whatever the resources allocated.

Nonetheless, in addition to the actionable findings in regard to the three DIF items noted above, there were findings in regard to specific Russian to Kyrgyz adaptation issues that are within the power of the testing center to control. Both substantive evaluations and relatively high effect size values on most sentence completion items support the notion that syntax differences between the two languages make sentence completion items somewhat problematic to adapt from Russian into Kyrgyz. At a minimum, the center should evaluate these items more closely. Or, perhaps reconsider need to keep this item type on the NST. This finding answers the question raised at the outset about whether or not DIF issues related to the specific languages

[76] I am not arguing that Kyrgyz "needs to become standardized," but rather emphasizing that contested languages poses challenges to standardized testing. Standardization always has winners and losers. For an interesting discussion on the "ideology of standardization" see Milroy (2001).

under study could be identified. It would appear that while generalizing is difficult for many items and item types, for at least for one set of items the answer is yes. At the same time, the fact that the reading comprehension items appeared to be the least problematic of the item types supports the idea that there are some "general challenges" to item adaptation, regardless of languages employed (Agnoff & Cook, 1988).

While clearly not infallible, if used properly, statistical methods can highlight inefficiencies, shed light on misconceptions and false beliefs about DIF and item bias, and demonstrate the strengths and weaknesses of a given testing program in regard to their development of instruments and specific item types. Specifically, statistical approaches can be employed to demonstrate that item response is complex and that item flaws will not always favor the Russian group. In general, statistical and substantive analyses are both needed to confirm hypotheses generated about the quality and nuances of cross-lingual test item adaptation (Hambleton, 2005).

There is now empirical evidence that DIF studies can be used to identify specific challenges in cross-lingual test item adaptation from Russian into Kyrgyz in the KR. In regard to the quantity of DIF, the results are heartening: The low number of practically significant DIF items indicates that cross-lingual adaptation in Kyrgyzstan is feasible. Data from such studies such as this one can be used to improve the NST. To my knowledge, at the time of this study, there have been no DIF studies conducted on cross-lingual tests in any of the former Soviet Republics. Therefore, the results of this study will be of special interest to researchers not only in the Kyrgyz Republic but in other countries where Russian and Turkic language(s) are the primary languages of instruction and assessment.

APPENDICES

APPENDIX A: SCHOOLS BY LANGUAGE(S) OF INSTRUCTION IN THE KR

Table A-1: Schools by Language(s) of Instruction in the KR

Buildings with medium	No. Schools	% Schools	% Students
Kyrgyz	1261	66.0	54.9
Russian	221	11.6	13.1
Uzbek	151	7.9	8.8
Kyrgyz/Russian	234	12.2	19.9
Kyrgyz/Uzbek	31	1.6	1.8
Russian/Uzbek	8	0.4	0.9
Kyrgyz/Russian/Uzbek	5	0.3	0.5
Total:	**1911**	**100.0**	**100.0**

Ministry of Education Data (2003).

APPENDIX B: STUDENTS (%) IN MAIN LANGUAGE TRACKS BY *OBLAST*

Table A-2: Students (%) in Main Language Tracks by *Oblast*

	Kyrgyz	Russian	Uzbek	Tajik
Republic	63.3	22.7	13.4	.30
Northern Oblasts				
Bishkek	34.8	65.2	-	-
Chui	39.9	60.0	.14	-
Talas	88.2	11.8	-	-
Issyk-Kul	72.7	27.3	-	-
Naryn	88.2	11.8	-	-
Southern Oblasts				
Batken	74.5	7.2	15.2	3.1
Djalal-Abad	71.4	8.4	20.2	-
Osh	63.8	7.4	28.7	.06

Year 2000, Herczynski (2003)

APPENDIX C: NST PARTICIPATION RATES IN THE KR

Table A-3: NST Participation Rates by *Oblast* & Language

	2009	**2010**	**2009**	**2010**	**2009**	**2010**	**2009**	**2010**
Region	N	N	Kyrgyz	Kyrgyz	Russian	Russian	Uzbek	Uzbek
All Republic	33,579	30,264	63%	60%	33%	36%	04%	03%
Bishkek (capital)	6,526	6,427	28%	25%	71%	75%	(n = 4)	(n=2)
Chui (northern)	4,405	3,848	41%	39%	59%	61%	(n = 1)	(n=1)
Issyk-Kul (northeastern)	3,881	3,561	69%	66%	31%	34%	(n = 2)	00%
Naryn (south central)	2,703	2,481	78%	81%	22%	19%	00%	00%
Talas (western)	1,724	1,533	80%	80%	20%	20%	00%	00%
Djalal-Abad (southern)	4,903	4,203	79%	78%	15%	17%	06%	05%
Osh City (southern)	1,398	1,186	49%	45%	40%	43%	10%	12%
Osh (southern)	5,011	4,534	84%	82%	07%	08%	09%	10%
Batken (southwestern)	3,028	2,491	80%	81%	11%	11%	09%	08%

2009 Annual NST Report, 2010 Annual NST Report, (www.testing.kg)

APPENDIX D: DEMOGRAPHICS AND TEST SCORES (2010)

Table A-4: Demographics and Test Scores

	% Poor*	% Higher Ed**	Avg. NST Scores***	% NST Russian***
All Republic	56.2	12%	113.5	36%
Bishkek (capital)	6.0	26%	135.4	75%
Osh City (southern)	n/a	17%	120.1	43%
Issyk-Kul (northeastern)	30.6	13%	111.1	34%
Chui (northern)	26.6	11%	116.5	61%
Naryn (south central)	90.5	11%	104.2	19%
Talas (western)	67.0	10%	103.9	20%
Djalal-Abad (southern)	73.0	8%	106.2	17%
Osh (southern)	65.7	7%	100.3	11%
Batken (southwestern)	65.7	7%	103.5	08%

Herczynski (2003)* Census (2009)** CEATM[a] (2010)***

APPENDIX E: DEMOGRAPHICS OF SCHOLARSHIP WINNERS

Table A-5: NST Winners by Language, *Oblast* (2010)

NST 2010	Kyrgyz			Russian			Uzbek		
	Part. %	Win %	Avg.	Part. %	Win %	Avg.	Part. %	Win %	Avg.
Region									
All Republic	60%	66%	125.6	36%	33%	153.9	03%	01%	130.2
Bishkek (capital)	25%	22%	134.0	75%	78%	164.7	(n=2)	00%	-
Chui (northern)	39%	34%	125.3	61%	66%	150.2	(n=1)	00%	-
Issyk-Kul (northeastern)	66%	60%	127.4	34%	40%	148.9	00%	00%	-
Naryn (south central)	81%	82%	123.4	19%	18%	136.4	00%	00%	-
Talas (western)	80%	76%	124.7	20%	24%	147.1	00%	00%	-
Djalal- Abad (southern)	78%	85%	122.5	17%	12%	142.1	05%	03%	132.6
Osh City (southern)	45%	52%	126.8	43%	45%	158.1	12%	03%	152.8
Osh (southern)	82%	93%	125.9	08%	03%	133.0	10%	03%	134.8
Batken (southwestern)	81%	89%	128.9	11%	05%	146.0	08%	06%	117.5

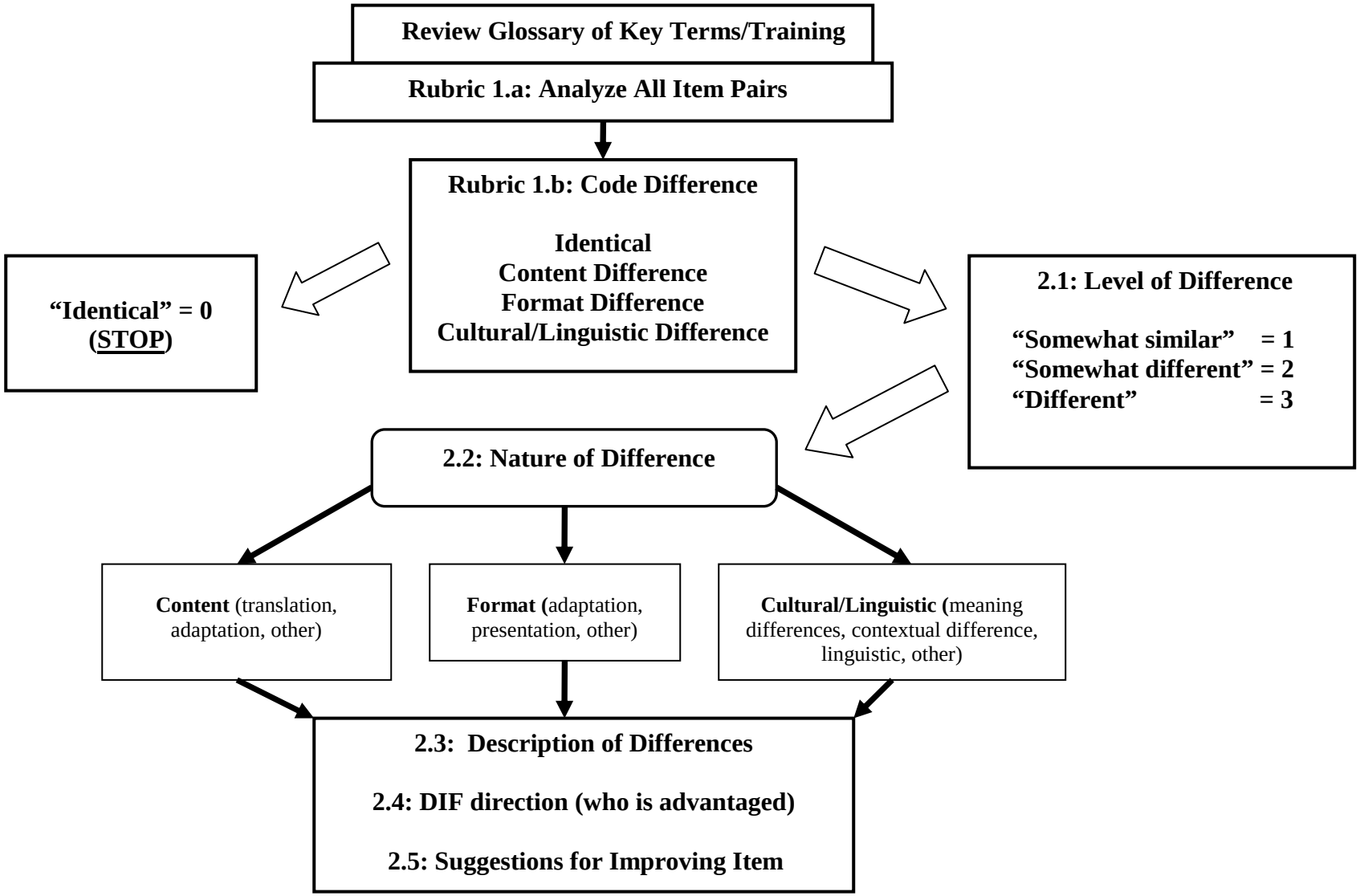
Data constructed from CEATM[a] (2010)

APPENDIX F: SELECTIVITY OF HEIs IN THE KR

Table A-6: Average NST Scores of Scholarship Winners

Institution		2009		2010	
	Location	**Average**	**Scholarships**	**Average**	**Scholarships**
Kyrgyz-Turkish Manas University	Bishkek	185.0	105	182.1	99
Kyrgyz-Russian Slavonic University	Bishkek	184.6	150	182.2	113
Kyrgyz State Medical Academy	Bishkek	176.2	199	177.7	217
International University	Bishkek	- -	- -	173.6	30
Kyrgyz Economic University	Bishkek	172.2	85	165.7	37
Kyrgyz State Technical University	Bishkek	158.5	526	145.6	574
Kyrgyz National University	Bishkek	143.9	474	143.2	436
Osh State University	Osh	143.1	473	135.8	431
Bishkek Humanities University	Bishkek	141.8	200	140.4	173
Building and Transport University	Bishkek	138.8	281	133.5	254
Issyk-Kul State University	Kara-Kol	131.3	196	133.6	147
Osh Technical University	Osh	130.3	285	118.5	277
Arabeava Pedagogical University	Bishkek	129.4	406	125.0	379
Institute of Mountain Technology	Bishkek	--	--	124.8	8
Kyrgyz-Uzbek University	Osh	127.8	245	- -	- -
Kyrgyz Agricultural University	Bishkek	126.6	186	120.9	255
Djalal-Abad State University	Djalal-Abad	126.4	344	120.2	343
Batken State University	Batken	122.6	165	122.7	160
Academy of Internal Affairs	Bishkek	121.7	230	107.1	230
Talas State University	Talas	121.2	84	116.4	60
Naryn State University	Naryn	117.0	145	115.4	150
Military Institute	Bishkek	104.1	150	107.1	99
	All:	**139.6**	**4,929**	**134.9**	**4,472**

APPENDIX G: COMPLETING THE ITEM ANALYSIS RUBRICS



APPENDIX H: GLOSSARY OF KEY RUBRIC TERMS

(*English Version*)

Differences between two language versions of the test items can potentially invalidate the inferences based on test results. It is generally understood that *different* means *not the same*. Differences can be caused by variation in wording due to translation or adaptation mistakes, content differences, format or item presentation differences or the way that different cultural groups interpret the test items. In the context of this study, there are four key aspects of difference that merit attention – differences in the meaning of individual words, differences in overall item meaning, differences in relative difficulty and differences in cultural interpretation of the two versions of the item. Ultimately, *equivalence* of items is achieved when the item in both language versions has the same content, meaning, same relative difficulty level, and can be interpreted similarly in the different linguistic and cultural groups.

Relative Difficulty

Individual words as well as phrases, concepts and ideas can have similar overall meaning in both versions but still be problematic for one group of examinees. That is because certain topics or concepts can differ in their conceptual difficulty in the two groups. An obvious example is when one language has five synonyms for the same word 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 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 understand for one group. Complex metaphors are sometimes adapted to a more literal meaning in the target language which can lead to the target language examinees having greater changes of success on an item.

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 the content rubric).

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 the format rubric).

Cultural/Linguistic Differences
Meaning of items to both Russian and Kyrgyz examinees is different, relevance to different schooling contexts and cultures is different, lack of similarity of dispositions of two groups, lack of similarity of norms, psychological construct not present in both groups, lack of equivalence of linguistic expression, lack of similarity of linguistic structure and grammar which makes equivalence challenging, differences in symbolism, metaphorical meaning, level of explicitness different, etc. (see prompts on rubric).

Terms from All Rubrics (2.1. Level of Differences)

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 certain that these differences will impact item response patterns.

Different
These items clearly have 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 the Content Rubric (Section 2.2. Nature of Differences)

The incorrect translation of individual words, the addition or omission of a word can cause differences in item meaning or content. This problem can sometimes be resolved relatively easily by improvements in translation. The word ***translation*** 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 acknowledges that direct, literal, translation is often not possible (nor 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. For this rubric, the term

adaptation is utilized to denote the process of conveying similar overall meaning, regardless of how individual words may or may not correspond.

Terms from Cultural/Linguistic Rubric (Section 2.2. Nature of Differences)

Meaning Differences
Under the meaning differences for the cultural category, I refer not to meaning differences caused by translation mistakes, but meaning differences that might occur even when the translation/adaptation 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.

Context Differences
Contextual differences can impact response when examinees have different levels of exposure to ideas, knowledge or situations due to demographic or social differences. In Kyrgyzstan, Russian speaking examinees are (on average) concentrated in urban areas while Kyrgyz speakers (on average) are concentrated in rural areas. Urban examinees might have less knowledge about horsemanship or animal husbandry and the vocabulary, knowledge and norms that are common to rural youth. Geographic concepts and terminology about mountains might also advantage those who live in high mountain regions. Or, the opposite, urbanites might be more familiar with issues connected to the ways of life of urban dwellers. Success on an item should also not depend on exposure to similar curriculum and schooling practices that might vary by group.

Cultural/Linguistic Differences
Cultural understandings may differ between the groups enough to make the intended meaning of some items unclear, irrelevant, or have a different difficulty level for one group. Due to cultural understandings some words, concepts, or ideas might be more familiar to one group than the other, even after controlling for residence. For example, a focus on the cultural heroes, myths, legends might also be problematic across cultures. Like meaning and contextual differences, cultural differences might not be apparent in the quality of translation/adaptation (which may be accurate) but must be considered nonetheless.

The most obvious form of "linguistic difference" becomes evident when items are poorly translated or adapted. However, there are also may be inherent differences in the way languages form, express and convey meaning that are irrelevant to the quality of adaptation. 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. The way two languages express or articulate ideas and concepts could make meaning more "difficult to locate" in some languages than others. Some languages might more efficiently convey meaning than others in some situations. 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 Russian and Kyrgyz below. 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. Are there differences in meaning and/or difficulty of the two paragraphs below? Are these differences related to inherent language differences? 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.”

APPENDIX I: ITEM RUBRICS 1.a & 1.b

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 the content, format and cultural/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 answering the test items in 1.a. and circle the descriptor that best characterizes each pair of items. Please circle differences (I, II, or III) 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
Item 3:	0. No differences	I. Content Differences	II. Format Differences	III. Cultural/Linguistic Differences
Item 4:	0. No differences	I. Content Differences	II. Format Differences	III. Cultural/Linguistic Differences
Item 5:	0. No differences	I. Content Differences	II. Format Differences	III. Cultural/Linguistic Differences
Item 6:	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

APPENDIX J: ITEM RUBRIC 2

Directions

Fill in item rubric 2 for each item not identified as "identical" on rubric 1.b (above). 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 reason 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.

Purple Color **Rubric 2: Content** **Item Number: _____**			
Consider the Equivalence of: Skills or knowledge demanded; vocabulary, ideas, situations, topics; words, expressions, sentences and phrases; 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.			
2.1. The content of these items is (*circle one*):	Somewhat Similar (1)	Somewhat Different (2)	Different (3)
2.2. The difference is best characterized as (*circle one*):	(a) Translation (individual word issues)	(b) Adaptation (general meaning)	(c) Other
2.3. Describe the difference(s) in detail:			
2.4. Advantage:	If the item content is different, do you think that it favors one of the groups? Which one? *(circle one*): Russian or Kyrgyz		
2.5. Improving Equivalence:	Can the equivalence problem(s) with this item be resolved? How?		

Green Color **Rubric 2: Format** **Item Number _____**			
Consider the Equivalence of: Overall item presentation, item length, clarity of directions, order of words and ideas, number of words, punctuation, capitalization, typeface, typographical errors, missing letters or words, editing and general formatting, etc.			
2.1. The format of these items is: ***(circle one)***	Somewhat Similar (1)	Somewhat Different (2)	Different (3)
2.2. The problem is best characterized as: ***(circle one)***	(a) Adaptation	(b) Presentation	(c) Other
2.3. Describe the difference(s) in detail:			
2.4. Advantage:	If the item format is different, do you think that it favors one of the groups? Which one? *(circle one*) Russian or Kyrgyz		
2.5. Improving Equivalence	Can the problem(s) with this item be resolved? How?		

Pink Color **<u>Rubric 2: Cultural/Linguistic</u>** **Item Number _____**				
Consider the equivalence of: Russian and Kyrgyz schooling context, curriculum 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.				
2.1. Cultural equivalence between the two items is *(circle one):*	Somewhat Similar (1)	Somewhat Different (2)	Different (3)	
2.2. The problem is best characterized as: *(circle one)*	(a) Meaning differences	(b) Contextual differences	(c) Linguistic differences	(d) Other
2.3. Describe the difference(s) in detail:				
2.4. Advantage:	If the items are not equivalent for cultural reasons, do you think that it favors one of the groups? Which one? (*circle one*) <u>Russian</u> or <u>Kyrgyz</u>			
2.5. Improving Equivalence	Can the problem(s) with this item be resolved? How?			

Опросник 2: Содержание

Номер задания: _____

Рассмотрите задания на эквивалентность по следующим вопросам:

Навыки и знания, словарный запас, идеи, ситуации, предмет, слова, выражения, предложения, сложность фразировки, пропуск слова или добавление слова, грамматика; частота слов, степень нюансов, степень очевидности, буквальное или переносное значение, использования метафор, идиом и т.п.

2.1. Содержание заданий (обведите одно):	небольшие различия (1)	средние различия (2)	значительные различия (3)
2.2. Причина различий (обведите одно):	(a) Перевод (суть отдельных слов)	(b) Адаптация (общее понятие)	(c) Другое
2.3. Подробно опишите различия:			
2.4. Преимущества:	Если содержание задания отличается, у какой группы больше шансов на правильный ответ: кыргызской или русской? (обведите одно)		
2.5. Улучшение эквивалентности:	Можно ли решить проблему эквивалентности? Как?		

Опросник 2: Формат **Номер задания: _____**			
Рассмотрите задания на эквивалентность по следующим вопросам: Представление задания в целом, длина вопроса, четкость инструкций, порядок слов и идей, количество слов, пунктуация, использование заглавных букв, шрифт, редактирование и общее форматирование и т.д.			
2.1. Формат заданий (обведите одно):	небольшие различия (1)	средние различия (2)	значительные различия (3)
2.2. Причина различий (обведите одно):	a) Адаптация	b) Вид	c) Другое
2.3. Подробно опишите различия:			
2.4. Преимущества:	Если формат задания отличается, у какой группы больше шансов на правильный ответ: кыргызской или русской? (обведите одно)		
2.5. Улучшение эквивалентности	Можно ли решить проблему эквивалентности? Как?		

Опросник 2. Культурные/лингвистические различия

Номер задания: _____

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.

2.1. Степень различия по культурному признаку (обведите одно):	небольшие различия (1)	средние различия (2)	значительные различия (3)	
2.2. Причина различий (обведите одно):	(a) Различия в значении	(b) Контекстуальные различия	(c) Лингвистические различия	(d) Другое
2.3. Подробно опишите различия:				
2.4. Преимущества	Если задания не эквивалентны по культурным признакам, у какой группы больше шансов на правильный ответ: кыргызской или русской? (обведите одно)			
2.5. Улучшение эквивалентности	Можно ли решить проблему эквивалентности? Как?			

APPENDIX K: UNIFORM DIF STATISTICS

Table A-7: Uniform DIF Statistics for 38 Verbal Items

	Model 1 (compact)		Model 2 (w/group)							
	χ^2	R^2	χ^2	R^2	χ^2 Difference	R^2 Δ (effect size)	Group β_2	sig.	Odds ratio Exp(β)	FAVORS
Item										
9	*609.384*	*0.365*	*609.878*	*0.365*	*0.494*	*0.000*	*-0.085*	*0.482*	*0.919*	
2	*620.330*	*0.372*	*621.063*	*0.372*	*0.733*	*0.000*	*0.112*	*0.393*	*1.118*	
24	*426.294*	*0.256*	*427.046*	*0.257*	*0.752*	*0.001*	*-0.097*	*0.385*	*0.908*	
7	*159.771*	*0.103*	*161.089*	*0.103*	*1.318*	*0.000*	*0.122*	*0.252*	*1.13*	
17	*446.021*	*0.297*	*448.098*	*0.298*	*2.077*	*0.001*	*-0.202*	*0.149*	*0.817*	
29	*91.075*	*0.061*	*93.444*	*0.062*	*2.369*	*0.001*	*0.17*	*0.125*	*1.185*	
39	381.248	0.245	385.981	0.248	4.733	0.003	0.278	0.031	1.32	Kyrgyz
35	256.366	0.162	261.162	0.165	4.796	0.003	-0.242	0.028	0.785	Russian
36	340.026	0.228	345.016	0.231	4.99	0.003	0.298	0.026	1.347	Kyrgyz
27	292.44	0.186	297.733	0.189	5.293	0.003	0.268	0.022	1.307	Kyrgyz
30	78.421	0.056	84.629	0.06	6.208	0.004	0.307	0.013	1.359	Kyrgyz
14	44.124	0.030	50.523	0.034	6.399	0.004	-0.275	0.011	0.759	Russian
31	513.364	0.302	521.002	0.306	7.638	0.004	-0.308	0.006	0.735	Russian
34	189.212	0.12	198.916	0.126	9.704	0.006	-0.331	0.002	0.718	Russian
12	377.694	0.238	387.473	0.244	9.779	0.006	0.385	0.002	1.469	Kyrgyz
40	401.746	0.243	412.05	0.249	10.304	0.006	-0.351	0.001	0.704	Russian
15	350.341	0.226	365.231	0.234	14.890	0.008	0.451	0.000	1.57	Kyrgyz
18	554.431	0.324	569.895	0.332	15.464	0.008	0.456	0.000	1.578	Kyrgyz
10	350.534	0.214	366.044	0.223	15.510	0.009	-0.428	0.000	0.652	Russian
37	213.746	0.136	229.341	0.145	15.595	0.009	0.429	0.000	1.536	Kyrgyz
8	298.915	0.192	317.089	0.202	18.174	0.010	0.515	0.000	1.673	Kyrgyz
4	609.253	0.379	628.754	0.390	19.501	0.011	0.574	0.000	1.776	Kyrgyz
38	464.279	0.278	484.489	0.289	20.21	0.011	-0.507	0.000	0.602	Russian
20	262.557	0.203	283.736	0.218	20.749	0.015	0.741	0.000	2.098	Kyrgyz
5	6.929	0.005	29.505	0.020	22.576	0.015	0.497	0.000	1.644	Kyrgyz

Table A-7 (cont'd)

25	35.889	0.026	58.895	0.042	23.006	0.016	0.583	0.000	1.792	Kyrgyz
22	441.505	0.264	465.075	0.277	23.57	0.013	-0.532	0.000	0.587	Russian
26	328.791	0.202	362.884	0.221	34.093	0.019	-0.634	0.000	0.531	Russian
23	530.86	0.311	569.563	0.33	38.703	0.019	-0.694	0.000	0.5	Russian
21	425.032	0.262	467.445	0.286	42.413	0.024	-0.738	0.000	0.478	Russian
16	161.157	0.123	204.570	0.154	43.413	0.031	0.98	0.000	2.663	Kyrgyz
33	188.954	0.122	232.381	0.149	43.427	0.027	0.76	0.000	2.138	Kyrgyz
11	314.134	0.195	363.460	0.223	49.326	0.028	0.791	0.000	2.205	Kyrgyz
28	295.451	0.183	345.596	0.212	50.145	0.029	0.796	0.000	2.127	Kyrgyz
19	558.080	0.333	652.350	0.381	94.270	0.048	-1.171	0.000	0.31	Russian
32	201.991	0.128	298.325	0.185	96.334	0.057	1.101	0.000	3.007	Kyrgyz
3	736.971	0.414	848.057	0.464	111.086	0.05	-1.247	0.000	0.287	Russian
13	264.243	0.166	392.577	0.238	128.334	0.072	-1.218	0.000	0.296	Russian

Items are arranged by order of chi-squared difference values in ascending order.
At the .05 level, the test statistic for 1 degree of freedom is 3.841.
When $\beta_2>0$, uniform DIF favors the reference group (Kyrgyz language). When $\beta_2<0$, uniform DIF favors the focal group (Russian language).

APPENDIX L: ITEMS WITH MODERATE OR LARGE DIF

Table A-8: Verbal Items with Moderate or Large DIF

Item	χ^2 difference	Effect Size		β_2	sig.	Odds Ratio Exp (β)	Favors	Item Type
		moderate	large					
19	94.270	0.048		-1.171	0.000	0.31	Russian	Analogy
32	96.334	0.057		1.101	0.000	3.007	Kyrgyz	Reading
3	111.086	0.05		-1.247	0.000	0.287	Russian	Analogy
13	128.334		0.072	-1.218	0.000	0.296	Russian	Analogy

Items are arranged by ascending chi-square difference values.

APPENDIX M: ITEMS WITH NO DIF

Table A-9: Non-Significant Verbal Items

Item	χ^2 difference	β_2	sig.	Odds Ratio Exp (β)	Item Type
9	0.494	-0.085	0.482	0.919	Analogy
2	0.733	0.112	0.393	1.118	Analogy
24	0.752	-0.097	0.385	0.908	Sentence Completion
7	1.318	0.122	0.252	1.13	Analogy
17	2.077	-0.202	0.149	0.817	Analogy
29	2.369	0.17	0.125	1.185	Sentence Completion

Items are arranged by ascending chi-square difference values.

APPENDIX N: NON-UNIFORM DIF STATISTICS

Table A-10: Non-Uniform Verbal DIF Statistics

	Model 2 (w/language)		Model 3 (interaction)						
	χ^2	R^2	χ^2	R^2	χ^2 Difference	$R^2 \Delta$ (effect size)	β_3	sig.	odds ratio Exp(β)
Item									
31	*521.002*	*0.306*	*521.004*	*0.306*	*0.002*	*0.000*	*0.000*	*0.969*	*1*
33	*232.381*	*0.149*	*232.397*	*0.149*	*0.016*	*0.000*	*0.000*	*0.899*	*0.999*
5	*29.505*	*0.020*	*29.554*	*0.020*	*0.049*	*0.000*	*0.001*	*0.826*	*1.001*
38	*484.489*	*0.289*	*484.626*	*0.289*	*0.137*	*0.000*	*-0.003*	*0.711*	*0.997*
40	*412.05*	*0.249*	*412.297*	*0.249*	*0.247*	*0.000*	*-0.004*	*0.619*	*0.996*
2	*621.063*	*0.372*	*621.345*	*0.373*	*0.282*	*0.001*	*0.005*	*0.596*	*1.005*
14	*50.523*	*0.034*	*50.913*	*0.034*	*0.390*	*0.000*	*-0.004*	*0.532*	*0.996*
22	*465.075*	*0.277*	*465.513*	*0.277*	*0.438*	*0.000*	*-0.005*	*0.508*	*0.995*
13	*392.577*	*0.238*	*393.526*	*0.239*	*0.949*	*0.001*	*-0.007*	*0.330*	*0.993*
11	*363.460*	*0.223*	*364.417*	*0.223*	*0.957*	*0.000*	*0.008*	*0.330*	*1.008*
8	*317.089*	*0.202*	*318.l64*	*0.203*	*1.075*	*0.001*	*0.008*	*0.302*	*1.008*
32	*298.325*	*0.185*	*299.544*	*0.186*	*1.219*	*0.001*	*0.008*	*0.272*	*1.008*
39	*385.981*	*0.248*	*387.206*	*0.249*	*1.225*	*0.001*	*0.009*	*0.271*	*1.009*
19	*652.350*	*0.381*	*653.597*	*0.381*	*1.247*	*0.000*	*-0.009*	*0.263*	*0.991*
3	*848.057*	*0.464*	*849.495*	*0.465*	*1.438*	*0.001*	*0.010*	*0.233*	*1.011*
35	*261.162*	*0.165*	*262.809*	*0.166*	*1.647*	*0.001*	*0.009*	*0.201*	*1.009*
34	*198.916*	*0.126*	*200.596*	*0.127*	*1.68*	*0.001*	*0.009*	*0.196*	*1.009*
12	*387.473*	*0.244*	*389.866*	*0.245*	*2.393*	*0.001*	*-0.012*	*0.121*	*0.988*
28	*345.586*	*0.212*	*348.014*	*0.213*	*2.428*	*0.001*	*-0.012*	*0.118*	*0.988*
26	*362.884*	*0.221*	*365.855*	*0.223*	*2.971*	*0.002*	*-0.012*	*0.084*	*0.988*
4	*628.754*	*0.390*	*632.316*	*0.391*	*3.562*	*0.001*	*0.023*	*0.059*	*1.024*
23	569.563	0.33	573.618	0.332	4.055	0.002	-0.015	0.044	0.985
37	229.341	0.145	234.861	0.149	5.52	0.004	0.017	0.020	1.017
10	366.044	0.223	372.257	0.226	6.213	0.003	-0.018	0.012	0.983
30	84.629	0.06	91.957	0.065	7.329	0.005	-0.019	0.007	0.981

Table A-10 (cont'd)

21	467.445	0.286	474.789	0.29	7.344	0.004	-0.022	0.007	0.978
9	609.878	0.365	617.288	0.369	7.410	0.004	0.028	0.007	1.029
24	427.046	0.257	437.017	0.262	9.971	0.005	-0.023	0.002	0.977
17	448.098	0.298	459.319	0.305	11.221	0.007	-0.029	0.001	0.972
20	283.736	0.218	295.348	0.226	11.612	0.008	0.031	0.001	1.032
18	569.895	0.332	581.663	0.338	11.768	0.006	0.031	0.001	1.031
25	58.895	0.042	71.278	0.051	12.383	0.009	-0.024	0.000	0.976
27	297.733	0.189	312.692	0.198	14.959	0.009	-0.027	0.000	0.973
29	93.444	0.062	109.723	0.073	16.279	0.011	-0.026	0.000	0.974
36	345.016	0.231	362.639	0.242	17.623	0.011	-0.033	0.000	0.967
15	365.231	0.234	384.341	0.245	19.110	0.011	0.040	0.000	1.041
7	161.089	0.103	181.691	0.116	20.602	0.013	0.032	0.000	1.032
16	204.570	0.154	229.358	0.172	24.788	0.018	0.042	0.000	1.043

Items are arranged in ascending order by chi-squared difference values.
At the .05 level, the test statistic for 1 degree of freedom is 3.841.
Non-uniform DIF = $\beta_3 \neq 0$, regardless of the value of β_2

APPENDIX O: ITEM LOCATION ACROSS EFFECT SIZE VALUES

Table A-11: Continuum of Effect Size Values by Item Type

Non-DIF Items (6 items total)

	.000	.000	.000	.001	.001	.001	% total*
Type							
AN	9	2	7		17		22%
SC				24		29	20%
RC							0%

* indicates the total % of all this item type found in this classification. I.e. 22% of all analogies fall under the "non-DIF" classification.

Negligible DIF Items (Below effect size median, 14 items total)

	.003	.003	.003	.003	.004	.004	.004	.006	.006	.006	.008	.008	.009	.009	% total
Type															
AN						14			12		15	18	10		28%
SC				27	30										20%
RC	39	35	36				31	34		40				37	70%

Negligible DIF Items (Above effect size median, 14 items total items)

	.010	.011	.011	.013	.015	.015	.016	.019	.019	.024	.027	.028	.029	.031	% total
Type															
AN	8	4			20	5						11		16	33%
SC				22			25	26	23	21			28		60%
RC			38								33				20%

Practical DIF Items (4 items total)

	.048	.050	.057	.072	% total
Type					
AN	19	3		13	17%
SC					0%
RC			32		10%

APPENDIX P: EVALUATOR SCORING MATRIX

Table A-12: Evaluator Item Scoring Matrix

Evaluator	1	2	3	4	5	6	7	8	Total
Item									
2	0	1	0	1	0	0.57	1	1	4.57
3	2	3	0	3	3	0	3	2	16
4	0	2	0	2	1	0	0	1	6
5	3	2	0	3	0	0	1	0	9
7	0	2	0	2	0	0	2	2	8
8	1	2	0	1	0	1	1	0	6
9	0	0	0	1	0.57	1	1	1	4.57
10	0	0	0	1	2	1	2	2	8
11	3	0	3	3	1	0	2	2	14
12	0	0	0	1	0	0	2	2	5
13	0	0	0	3	0	1	1	0	5
14	0	0	0	0	0	0	1	1	2
15	0	3	3	2	3	2	1	0	14
16	0	0	2	3	0	0.85	1	0	6.85
17	0	0	0	1	0	0	1	1	3
18	2	0	2	3	0	0	1	3	11
19	3	1	1.4	2	2	0	1	1	11.4
20	0	0	0	0	0	0	0	0	0
21	0	3	2	3	3	2	2	1	16
22	1	0	2	0	0	2	0	0	5
23	0	3	2	1	2	1.1	0	0	9.1
24	0	2	2	0	0	0.57	0	0	4.57
25	0	2	2	2	1.42	0	2	2	11.42
26	0	0	0	0	2	2	0	1	5
27	0	0	0	3	0	0	0	0	3
28	0	0	0	0	1	0	0	0	1
29	0	0	0	0	0.28	0	2	0	2.28
30	0	0.28	0	0	2	0	0	0	2.28
31	1	3	2	1	0	1.28	1	1	10.28
32	0	0	0	0	3	3	0	0	6
33	3	0	2	3	3	0	2	1.85	14.85

Table A-12 (cont'd)

34	0	2	0	0	0	0	0	0	2
35	1	0	0	1	1	0	1	3	7
36	3	0	0	0	2	0	0	0	5
37	0	0	0	0	1	0	0	0	1
38	0	0	0	3	2	0	2	1	8
39	0	0	0	0	0	0	0	0	0
40	1	0	0	1	0.42	0	1	0	3.42

APPENDIX Q: INTER-RATER RELIABILITY

Table A-13: Reliability Statistics

Case Processing Summary

		N	%
Cases	Valid	38	100.0
	Excluded[a]	0	.0
	Total	38	100.0

a. Listwise deletion based on all variables in the procedure.

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.663	.657	8

Item Statistics

	Mean	Std. Deviation	N
V1	.6316	1.07606	38
V2	.8232	1.15466	38
V3	.6684	1.02644	38
V4	1.3158	1.21043	38
V5	.9655	1.10491	38
V6	.5097	.79436	38
V7	.9474	.83658	38
V8	.7855	.92961	38

Table A-13 (cont'd)

Inter-Item Correlation Matrix

	V1	V2	V3	V4	V5	V6	V7	V8
V1	1.000	.005	.273	.361	.187	-.251	.218	.262
V2	.005	1.000	.353	.229	.216	.203	.150	-.014
V3	.273	.353	1.000	.356	.217	.287	.137	.146
V4	.361	.229	.356	1.000	.173	-.112	.551	.391
V5	.187	.216	.217	.173	1.000	.341	.282	.135
V6	-.251	.203	.287	-.112	.341	1.000	-.067	-.233
V7	.218	.150	.137	.551	.282	-.067	1.000	.605
V8	.262	-.014	.146	.391	.135	-.233	.605	1.000

Intraclass Correlation Coefficient

		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.198[b]	.101	.338	2.971	37	259	.000
Average Measures	.663	.473	.804	2.971	37	259	.000

Two-way random effects model where both people effects and measures effects are random.

a. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

b. The estimator is the same, whether the interaction effect is present or not.

APPENDIX R: RAW DATA FOR RANK ORDER ESTIMATION

Table A-14: Chi-Square Values & Evaluators' Scores

Item	Chi-Square Difference*	Evaluators' Score
9	0.494	4.57
2	0.733	4.57
24	0.752	4.57
7	1.318	8
17	2.077	3
29	2.369	2.28
39	4.733	0
35	4.796	7
36	4.99	5
27	5.293	3
30	6.208	2.28
14	6.399	2
31	7.638	10.28
34	9.704	2
12	9.779	5
40	10.304	3.42
15	14.890	14
18	15.464	11
10	15.510	8
37	15.595	1
8	18.174	6
4	19.501	6
38	20.21	8
20	20.749	0
5	22.576	9
25	23.006	11.42
22	23.57	5
26	34.093	5
23	38.703	9.1
21	42.413	16
16	43.413	6.85
33	43.427	14.85
11	49.326	14
28	50.145	1
19	94.270	11.4
32	96.334	6
3	111.086	16
13	128.334	5

***Data presented by ascending chi-square difference values.**

APPENDIX S: RANK ORDER CORRELATION

Table A-15: 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).

APPENDIX T: EVALUATOR MARKS AND DIF STATISTICS

Table A-16: Evaluator Marks and DIF Statistics

Item	Marks	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
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

APPENDIX U: NUMBER, NATURE OF DIFFERENCES BY ITEM

Table A-17: Number and Nature of Differences by Individual Item

Item	# Distinct Issues	Nature of Difference	# Marking DIF	Effect Size
Analogies				
2	2	1. ADAPTATION (1 word, multiple meanings) 2. ADAPTATION (loan word used)	0	.000
3	3	1. SOCIO-DEMOGRAPHIC (city kids will not know a word) 2. FORMAT (misprint resulted in unknown word in answer key) 3. ADAPTATION (multiple meanings)	6	.05
4	2	1. ADAPTATION (needed direct translation, not adaptation) 2. TRANSLATION (needed literary word)	2	.011
5	3	1. GRAMMAR (incorrect word combination) 2. TRANSLATION (single word translated incorrectly) 3. ADAPTATION (Different words, same relationship)	3	.015
7	1	1. SOCIO-DEMOGRAPHIC (rural examinees lack knowledge)	4	.000
8	1	1. TRANSLATION (incorrect translation of a single word)	1	.010
9	2	1. ADAPTATION: (incorrect word combination) 2. TRANSLATION (incorrect translation which makes the pairs different)	0	.000
10	2	1. TRANSLATION (incorrect direct translation of one word) 2. TRANSLATION: (incorrect translation of one word)	3	.009
11	3	1. TRANSLATION (mistake in the translation produces "opposite" of what was intended) 2. TRANSLATION (incorrect direct translation in distractor) 3. TRANSLATION (incorrect translation)	5	.028
12	1	1. GRAMMAR (word can only be used in combination with other words)	2	.006
13	3	1. ADAPTATION (answer key has multiple meanings) 2. TRANSLATION (direct translation is incorrect) 3. TRANSLATION (incorrect translation)	1	.072

Table A-17 (cont'd)

14	1	1. ADAPTATION (style: singular vs. plural use)	0	.004
15	2	1. ADAPTATION (commonality of word used) 2. ADAPTATION (artificially created word)	5	.008
16	1	1. TRANSLATION Incorrect translation of a single word)	2	.031
17	2	1. ADAPTATION (two word combination makes the answer obvious) 2. SOCIO – DEMOGRAPHIC (unknown word for some regions)	0	.001
18	3	1. TRANSLATION (incorrect translation) 2. ADAPTATION (a word is used in simple speech, not literary) 3. TRANSLATION (incorrect translation)	4	.008
19	4	1. FORMAT (missing letter) 2. ADAPTATION (incorrect, makes the stem have the opposite of intended meaning) 3. TRANSLATION (incorrect literal translation) 4. TRANSLATION (incorrect nuance in meaning)	3	.048
20	0	NO ISSUES	0	.015
Sentence Completion				
21	5	1. TRANSLATION (too literal from Russian) 2. TRANSLATION (single word) 3. TRANSLATION (single word) 4. TRANSLATION (single word) 5. FORMAT (spacing differences)	6	.024
22	4	1. SOCIO-DEMOGRAPHIC (regional differences in vocabulary) 2. ADAPTATION (stylistic differences in distractors (negatives) 3. ADAPTATION (stylistic differences in distractors (antonyms) make sentences longer/hard to solve in Kyrgyz) 4. ADAPTION (equivalence of two pairs of words not good).	2	.013
23	3	1. GRAMMAR (Kyrgyz sentence difficult to understand) 2. ADAPTATION (lack of equivalence in word concept) 3. ADAPTATION (one word has no meaning in Kyrgyz)	3	.019
24	3	1. ADAPTATION (sentences too long in Kyrgyz) 2. TRANSLATION (do not need to adapt, use loan words) 3. TRANSLATION (incorrect single word)	2	.001

Table A-17 (cont'd)

25	1	1. TRANSLATION (use of dialect)	5	.016
26	4	1. TRANSATION (incorrect single word) 2. TRANSLATION (text size too big/ causes loss of meaning) 3. GRAMMAR (mistake in distractor a) 4. GRAMMAR (mistake in distractor b)	2	.019
27	2	1. FORMAT (typo in one word) 2. GRAMMAR (mistake)	1	.003
28	3	1. GRAMMAR (incorrect connector used) 2. GRAMMAR (incorrect ending) 3. GRAMMAR (incorrect form of word)	0	.029
29	3	1. GRAMMAR (incorrect ending) 2. GRAMMAR (incorrect form of word) 3. TRANSLATION (poor word choice)	1	.001
30	3	1. GRAMMAR (incorrect word choice) 2. GRAMMAR (incorrect word choice) 3. GRAMMAR (spelling mistake)	1	.004
Reading Comprehension				
31	3	1. GRAMMAR (incorrect word combination) 2. FORMAT (distractor order different) 3. GRAMMAR (different ending needed)	2	.004
32	2	1. FORMAT (typo error) 2. ADAPTATION (content of questions different)	2	.057
33	2	1. TRANSLATION (incorrect direct translation) 2. ADAPTATION (question form incorrect in Kyrgyz)	5	.027
34	1	1. TRANSLATION (direct translation incorrect)	1	.006
35	2	1. FORMAT (distractors in different places in two versions) 2. TRANSLATION (single word incorrect)	1	.003
36	1	1. ADAPTATION (item is long and complex in Kyrgyz)	2	.003
37	1	1. GRAMMAR (possessive ending incorrect in Kyrgyz)	0	.009
38	2	1. ADAPTATION (incorrect form of sentence) 2. FORMAT (typo, missing letter)	3	.011
39	0	NO COMMENTS	0	.003
40	1	1. FORMAT (distractors in difference places)	0	.006

APPENDIX V: KYRGYZ ONLY DIF ANALYSIS

Table A-18: DIF Statistics for Kyrgyz Rural and Urban Students

	Model 2 χ^2	Model 1 χ^2	χ^2 Difference	Model 2 R^2	Model 1 R^2	R^2 Δ	Sig.
Reading Comp 40	126.02	124.80	**1.22**	.105	.104	0.001	.270
Reading Comp 39	125.02	124.64	**0.38**	.117	.117	0.000	.539
Reading Comp 38	101.65	101.43	**0.22**	.092	.092	0.000	.644
Reading Comp 37	130.11	130.05	**0.06**	.107	.107	0.000	.788
Reading Comp 36	41.56	40.49	**1.07**	.042	.041	0.001	.300
Reading Comp 35	87.81	87.46	**0.35**	.077	.076	0.001	.552
Reading Comp 34	62.10	60.31	**1.79**	.054	.052	0.002	.181
Reading Comp 33	99.07	98.97	**0.1**	.084	.084	0.000	.749
Reading Comp 32	161.75	151.74	**0.01**	.132	.132	0.000	.911
Reading Comp 31	176.00	175.41	**0.59**	.146	.146	0.000	.441
Sentence Comp 30	*12.67*	*6.95*	***5.72***	*.012*	*.007*	*0.005*	*.017*
Sentence Comp 29	8.76	7.93	**0.83**	.008	.007	0.001	.364
Sentence Comp 28	144.21	143.95	**0.26**	.118	.118	0.000	.609
Sentence Comp 27	34.04	32.76	**1.28**	.030	.029	0.001	.258
Sentence Comp 26	43.62	42.76	**0.86**	.039	.038	0.001	.353
Sentence Comp 25	.97	.49	**0.48**	.001	.000	0.001	.490
Sentence Comp 24	65.71	65.08	**0.63**	.057	.056	0.001	.427
Sentence Comp 23	94.43	94.35	**0.08**	.084	.084	0.000	.780
Sentence Comp 22	106.48	106.43	**0.05**	.091	.091	0.000	.823
Sentence Comp 21	*115.39*	*102.704*	***12.686***	*.095*	*.085*	*0.010*	*.000*
Analogy 20	130.29	130.21	**0.08**	.151	.151	0.000	.768
Analogy 19	53.84	53.82	**0.02**	.058	.058	0.000	.886
Analogy 18	*349.16*	*344.13*	***5.03***	*.268*	*.265*	*0.003*	*.025*
Analogy 17	30.42	30.41	**0.01**	.036	.036	0.000	.975
Analogy 16	168.50	167.03	**1.47**	.170	.168	0.002	.226
Analogy 15	304.94	304.32	**0.62**	.243	.243	0.000	.432

Table A-18 (cont'd)

Analogy 14	3.48	1.91	**1.57**	.003	.002	0.001	.215
Analogy 13	30.94	304.32	**0.27**	.026	.026	0.000	.601
Analogy 12	77.24	77.15	**0.09**	.072	.072	0.000	.774
Analogy 11	207.56	206.79	**0.77**	.167	.166	0.001	.383
Analogy 10	25.38	23.22	**2.16**	.022	.020	0.002	.142
Analogy 9	365.94	365.65	**0.29**	.280	.280	0.000	.591
Analogy 8	130.80	130.46	**0.34**	.115	.115	0.000	.562
Analogy 7	144.50	144.05	**0.45**	.118	.118	0.000	.500
Analogy 5	*15.69*	*11.77*	***3.92***	*.013*	*.010*	*0.003*	*.048*
Analogy 4	384.10	383.06	**1.04**	.301	.301	0.000	.855
Analogy 3	*201.13*	*196.15*	***4.98***	*.190*	*.186*	*0.004*	*.026*
Analogy 2	173.30	170.01	**3.29**	.167	.164	0.003	.070

APPENDIX W: SUMMARY ITEM ANALYSIS RUBRICS

Evaluator Rubric (coded summary data)						
Item 2 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑ ☑ ☑					
Content		☑				
Format		☑				
Cult/Ling.		☑ ☑				
Other						
						0

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. <u>Advantage:</u> **Russian:** ☑☑ **Kyrgyz:**

2.5. <u>Can these items be reconciled?</u>

<u>Discussion:</u>

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.

Evaluator Rubric (coded summary data)						
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 (B), which is the answer key; wrote "Чоно" (*no meaning*) – should be "Чопо" (clay)
- Orthographical mistake in distractor (B) – 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 (B) – 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 (B).
- 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 (a) 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) 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: 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 can not 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).

Evaluator Rubric (coded summary data)						
Item 4 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content						
Format						
Cult/Ling.		☑☑		☑☑		
Other						
						☑☑

2.3. Describe Differences in Detail:

Format:

- Incorrect adaptation of the Russian word “шарф” (scarf) to the Kyrgyz “моюн жоолук” (*lit*. neck wrap)- should have used original Russian loan word, not a translation.

Culture/Language:

- The translation of “шарф”r. (scarf) into Kyrgyz is problematic – do not use the direct translation.
- In distractor (г) should have left “шарф”r. (scarf) in both the Russian and Kyrgyz versions.
- The literal translation of the word “шарф”r. (scarf) – “моюн жоолук” (*lit*. neck wrap) is not a widely used word and can impact understanding of the main idea.
- In answer (г), the word “шарф”r. (scarf) is translated literally as ““моюн жоолук”k. (*lit*. neck wrap).” It is necessary to use a word more appropriate to the original meaning.
- Incorrectly adapted “шарф”r. (scarf) to “моюн жоолук”k. (*lit.* neck wrap)”
- In distractor (В) the word “жабыштыруу”k. (to glue) is more “literary” than “чаптоо”k. (to glue) and is a better fit for this situation.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5 Can these items be reconciled?

- Yes, leave the original Russian word in the Kyrgyz item as well – "шарф" (scarf)
- It is not necessary to translate some words literally because kids can translate in their own way.
- Use the original Russian.
- Use a more literary term in distractor (B).

Discussion:

ZS: I think foreign words (cognates) should stay in their original form. **AA**: I agree, if there are no commonly used equivalents for foreign words, use the commonly used version. **MK**: It is best to use active, commonly used words.

MD: the problem here seems to be a too literal translation; sometimes there is no reason to translate. I recommend leaving the original if the foreign words are used widely. If not, then it should be translated. **NO**: Actually I think there is a Kyrgyz equivalent to "шарф" (scarf) but it is not used very often. **MD**: Well… how can we say what "often" is – how do we know this?

Evaluator Rubric (fully coded data)						
Item 5 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content					☑	
Format						
Cult/Ling.		☑		☑	☑	
Other						☑☑☑

2.3. Describe Differences in Detail:

Content:

- The pair of words, “палец: отпечаток” (finger: fingerprint) in the Russian version (distractor A) is not the same pair of words in the Kyrgyz version, perhaps they relied on a full adaptation, not translation?

Culture/Language:

- Incorrect translation of single words in distractor (A), but the adaptation seems correct because the relationship of the words seems to be the same.
- There is a difference with the translation in distractor (A). “Бут”k. (leg) means “нога”r. (leg) in Russian but “палец” r. (finger) is used in the Russian version. A direct translation of finger into Kyrgyz would be “манжа**”**k. (finger). However, this should not impact the test results if the Russian and Kyrgyz versions of the test will be used separate from each other.
- In answer (A) the pair of words are “бут: из”k. (leg: track, footprint) while the Russian version is “палец: отпечаток” (finger: fingerprint): The translation contains the same relationship but through completely different words.
- The word “бут”k. is incorrectly translated. The correct translation is “бармак”k. (finger), (A) difference in meaning.
- In distractor (б), “нерсе”k. (subject, thing) should be used in conjunction with other words like “бир нерсе.”
- For answer (б) the word “нерсе”k. is usually used in combination with other words.
- The word “нерсе” can be understood as “что-то”r. (something) or “что-либо”r. (anything).
- In distractor (г), the Kyrgyz “сургуч” has two meanings in Russian; it can mean both “тряпка” (cloth) and “щётка” (brush); “пыль” (dust) usually is wiped with “тряпка”r. (cloth) in Kyrgyz it is adapted well.
- Also, “сургуч”k. has two Russian equivalents – “тряпка”r. (cloth) and “тёрка”r. (grater)

- The word “сургуч”k. means “тряпка” (cloth) but in Russian version the word “щётка” (brush) is used.
- In Kyrgyz conversation, the Russian word “щётка”r. (brush) is often used.
- The word “щётка”r. (brush) is translated as “сургуч”k. which may or may not be correct.

2.4. <u>Advantage:</u> **Russian:** **Kyrgyz:** ☑

2.5. <u>Can these items be reconciled?</u>

- Be more attentive to the translation.
- Use different words.

<u>Discussion:</u>

MK, ZS: Word choice is a bit incorrect. It would have been better to translate according to the context. The problem (not significant) is with the translation. **MD**: It seems that this item should not be difficult to solve however because the differences in translation do not seem to always impact the relationships in the pairs of words. That is, if the core relationships in an analogy item are maintained, there might not be any differences in response patterns.

Evaluator Rubric (fully coded data)						
Item 7 **1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content				☑☑☑☑		
Format						
Cult/Ling.						
Other						
						☑☑☑☑

3. Describe Differences in Detail:

Content:

- In rural schools where testing is in Kyrgyz, they might not be familiar with the terms “Терапевт” r. (therapist) in the item stem and “слесарь" r. (metalworker) in distractor (г).
- Kids from rural schools will not know the word “Терапевт” r. (therapist).
- “Адвокат”r. (advocate) might be an unknown loan word. Also, “слесарь" r - in rural areas Kyrgyz use the word “темир уста” k., which literally translates as ““мастер по железо”r. (master of iron), which is not the same as “Кузнец” r. (blacksmith).
- The word “слесарь" r. (metalworker) could have been changed to a different word, not a loan word, a word familiar to Kyrgyz rural kids. This is a contextual difference.
- In this item, Russian loan words are used for professional specialties like “Терапевт” r. (therapist) and “Адвокат” r. (advocate).
- For some village kids these words won’t be understood.

4. Advantage: Russian**:** ☑☑☑ Kyrgyz:

5. <u>Can these items be reconciled?</u>

- Use words that are used and understand by representatives of both groups.
- It is possible to have used a different medical term.
- Maybe use other words, also understood by the masses.
- The word “слесарь" r. (metalworker) can be translated as “темирчи”k.

<u>Discussion:</u>

All: There may be problems of contextual differences and word use. The stem and distractors contain many Russian loan words (five words total are the same in both items) that some Kyrgyz students may not understand.

Evaluator Rubric (fully coded data)						
Item 8 **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 distractor (A) the Kyrgyz pair is “чака: куу” which means “Ведро: Лить” (pail: pour) in Russian; however, in the Russian version it is given “Лейка: Лить” r. (watering can: pour).
- A “чака” in Kyrgyz is “Ведро”r. (pail) which is not the same thing as “Лейка”r. (watering can) in the Russian version.
- Incorrect translation of the word “чака”k. (pail) from “Лейка”r. (watering can). It would be better to translate “Лейка”r. as “суу күйгүз.” But this shouldn’t impact the correct answer if the Russian and Kyrgyz versions will be used separately.
- “Лейка”r. (watering can) is translated incorrectly as “чака”k. (pail)

Culture/Language:

- In distractor (A) the word “чака”k. is not “Лейка” but “Ведро” (pail). Need to use the word “суу күйгүз”if the Russian version is to remain unchanged.
- The word “Лейка” can be translated as “суу күйгүз.”

2.4. Advantage: Russian: ☑ Kyrgyz: No Advantage: ☑

2.5. Can these items be reconciled?

- Adapted normal in both languages.
- Need to use a different pair of words in distractor (A).

Discussion (None)

Evaluator Rubric (fully coded data)						
Item 9 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑					
Content		☑☑				
Format						
Cult/Ling.		☑☑				
Other						
						0

2.3. Describe Differences in Detail:

Content:

- In distractor (б), for the Kyrgyz version, instead of the word “кыйын” (difficult) it would be better to use the synonym “татаал”k. (difficult), because the combination “түшүнүү” with “татаал” is better; for example “түшүнүүгү татаал”k. (it is difficult to understand). The synonyms “кыйын – оор”k. (difficult) are more common (but not literary), in my opinion.
- One of the distractors is translated incorrectly. In (B) the Kyrgyz version is “суу: кургатуу”k. (water: dry) but the corresponding Russian version is “мокрый: сушить”r. (*wet*: to dry). If these words were used in context (in a sentence) then it would be OK. For example – “I was in the rain and got wet.” However, when no context is given, this is a problem and literal translation is necessary.

Culture/Language:

- “суу” k. (water) is “Вода” r. (water) – not “мокрый”r. (wet). If it is to be translated there needs to be a correction, perhaps “суу болуп калды”k. (it gets wet) or “суу болуу”k. (to become wet). However, kids might understand from the context.
- Answer (B) “суу-кургатуу”k. (water: dry) while in the Russian variant it is “мокрый: сушить” r. (wet: to dry), i.e. there is a different of the meaning of the words.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Consider the specifics of the language.
- It is necessary to use to the word “cyy”k. (water) in a pair with a different word.
- Sometimes it is necessary to use a literal translation, especially with analogy items.

Discussion:

Evaluator Rubric (fully coded data)						
Item 10 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑ ☑ ☑					
Content				☑ ☑		
Format						
Cult/Ling.		☑		☑		
Other						
						☑ ☑ ☑

2.3. Describe Differences in Detail:

Content:

- The word “Ярче” r. (brighter) in the Russian version, distractor (B) was translated incorrectly into Kyrgyz as “карарaак” (darker) – which in principal changes the relations between the words.
- In distractor (B) there is an incorrect translation of the word “Ярче” r. (brighter) into “карарaак” k. (blacker, darker). “Ярче” r. (brighter) is actually “ачыгыраак” in Kyrgyz.

Culture/Language:

- In distractor (B), “карарaак”k. is actually “чернее”r. (darker), not “Ярче”r. (brighter) which is given. In order to make the pair equivalent you need the Kyrgyz word “ачыгыраак”k. (brighter).
- In Kyrgyz, in the pair of words there is a contradiction: “даана: карарaак” (clear: darker) while the Russian version is “чёткий: Ярче” (clear: brighter).
- The Russian pair in distractor (B) чёткий: Ярче” (clear: brighter), can be translated like “так, ачык” in Kyrgyz.
- The Kyrgyz pair (B) “даана: карарaак”k. is not the same as the Russian pair - “чёткий: Ярче”r.: “Ярче” (brighter) in Kyrgyz would be “ачык.”
- (r) incorrect translation: “ылдам”k. is not “быстрый”r.; “ылдам*ду* = быстрый” – but they used the first version.

2.4. <u>Advantage:</u> Russian: ☑☑☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

- Check the translation.
- “карарааk”k. should have been “ачыгырааk”k.
- Possibly change the distractor: However, I am not sure because maybe it was done to maintain the differences between the correct answer and the distractors.

<u>Discussion:</u>

All: The problem is poor translation and problematic adaptation.

Evaluator Rubric (fully coded data)						
Item 11 **2.1. <u>Difference Levels:</u>**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑					
Content		☑			☑	
Format						
Cult/Ling.				☑☑	☑☑	
Other						
						☑☑☑☑☑

2.3. <u>Describe Differences in Detail:</u>

Culture/Language:

- In distractor (б), the pair of Russian words “Включить: выключить” (turn on: turn off) are translated incorrectly into Kyrgyz. The given Kyrgyz pair “үзүү: кошуу,” (break: connection) are equivalent to a different Russian pair, “отрывание: соединение” r. (break, tear off: connection)
- Incorrect translation of the Russian word “выключить” r. (turn off) into Kyrgyz in distractor (б). The correct translation is “өчүрүү” k. (turn off)
- Incorrect translation of the word “выключить” r. (turn off)” – the correct version would be “өчүрүү” k. (turn off)”
- “үзүү” k. (break) is “оборвать” r. (break), not “выключить” r. (turn off). The word needed is “өчүрүү” k. (turn off). “үзүү” k. (break *eng*) is used when we speak about breaking a thread or a rope. This mistake makes the item difficult for Kyrgyz examinees. Also, in the item stem, the word “күйгүзүү”k. (to light a lamp, or to burn your hand) needs to be replaced with “жак”k. (to light) or “жандуруу”k. (to light *a lamp*)
- (б) “үзүү”k. (break) = “оторвать”r. (break) - “кошуу” k. (add) = “дабавлять” r. (add)
- (б) “үзүү”k. is equivalent to “оторвать, рвать”(break); “кошуу”k. is equivalent to “дабавлять”r. (add)
- “үзүү”k. means “оторвать, рвать” – выключить = өчүрүү. The incorrect word is used, usually “өчүрүү”k. is used.
- In item distractor (г), “угуу: айтуу“k. (listen: speak), the translation from Russian is incorrect. In Russian the pair is “ответить: спросить” (answer: ask)
- The distractor (г) also has translation problems “угуу: айтуу“k. (listen: speak) is not what the Russian pair is. “спросить”r. = “суроо”k., “ответить”r. = “жооп берүү”k.

2.4. <u>Advantage:</u> Russian: ☑☑☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

- Check the translation
- Needed to use different words.

<u>Discussion:</u>

MK: There are two distractors with translation problems here. Distractor (<u>г</u>) has an obvious incorrect translation; distractor (б) is also not an exact translation. I think this is important because the Russian distractor (б) is attractive, but in Kyrgyz (б) is less attractive due to translation mistake. **MD**: Maybe they are looking for associations of "like words."

Evaluator Rubric (fully coded data)						
Item 12 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Similar**	**Different**	**Total Diff.**
	☑☑☑☑☑					
Content						
Format						
Cult/Ling.		☑		☑☑		
Other						
						☑☑

2.3. Describe Differences in Detail:

Culture/Language:

- “жакыныраак”k. (closer) should be used in combination with different words. For example, “жакыныраак карoo”k. (look closer) – depending on the context.
- “пристальнее”r. (fixedly, intently) is given as “жакыныраак”k. (closer) which results in an incorrect relationship between the pair of words. “пристально смотреть”r. (stare) is more accurately - “тигилип карoo”r. (*lit* stare at).
- “пристальнее”r.(fixedly) is “тигилип карoo”k. (stare at). For some reason, the incorrect word “жакыныраак”k. (closer) “ближий” is used instead, which means to look closely at something.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Use two words together.
- Use the correct words.

Discussion: (None)

Evaluator Rubric (fully coded data)						
Item 13						
2.1. Difference Levels:	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑					
Content		☑				
Format						
Cult/Ling.		☑			☑	
Other						
						☑

2.3. Describe Differences in Detail:

Culture/Language:

- The item pairs contain mistakes in the distractors. For example, in the answer key (б) the word “жүрүү” in Kyrgyz means “to go.” However, depending on the combination of words used with this word, it can mean either going by foot or by car. In contrast, the word “ездить” in Russian means “going *only with* transportation – by bus, by car, by taxi, etc. This can change the relationship of the analogy pairs.
- In distractor (B), “көчө”k. (street) = “улица”r. (street); “дорога” (road) = “жол” (road) - so there is an incorrect translation here.
- In distractor (B), actually “көчө”k. (street) is “улица” r. (street) but the Russian word “дорога” (road)” is used instead.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- The main idea needs to be considered.
- The word “жол” (road) needed to be used.

Discussion:

MD: There are several problems with the distractors and answer key in these analogies. The problem is many words have many meanings, and the choice needs to be determined by the context. **AA**: the translation is literal. **CJ**: I see the problem as multiple meaning of words and associations. **AA**: First, in distractor B the translation of the Kyrgyz word "көчө" is incorrect because it means "street" not "road." The Russian version uses the word "дорога" (road) instead. I think that "дорога" (road) means something that has asphalt. **MD**: Well, I don't necessarily agree with that and don't believe everyone defines it that way… many villages have "roads" which are not asphalted.

RM: The problem is that the distractors are not good. **MD**: Yes, maybe the main problem is in the answer key. In my comments I wrote that "жүрүү" k. is different from "ездить" r. because "жүрүү"k. can be walking by foot or going by car while "ездить"r. means going by transportation. In Kyrgyz, perhaps "айдоо" (drive) would be a better choice for the pair because it has meaning like the Russian "ездить." In this case, it is like equating the English "walking" and "go" - in one term the meaning is wider while in the either the meaning is narrow … If they used "айдоо," (drive) they will get it quickly… I think this is an issue of translation – it is a good item but the direct translation is incorrect. Many of us thought for a very long time about what the correct answer here is to this item… Distractor "Г" appears to be a very good choice here…

Evaluator Rubric (fully coded data)						
Item 14						
2.1. Difference Levels:	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑					
Content						
Format						
Cult/Ling.		☑☑				
Other						
						0

2.3. Describe Differences in Detail:

Culture/Language:

- In all the Kyrgyz pairs of words, the words should be used in singular form. It would be correct in Kyrgyz to use "кол: манжа" (arm: hand), "адам: бут" (person: arm), "тил: тиш" (tongue: tooth) "бет: көз" (face: eye); in the Russian version all the words are used in the plural form.
- In distractor (A), "манжалар"k. is "кистьи"r. (wrist) in Russian but in the Russian version the word "пальцы" (fingers) is used. The word "пальцы"r. (fingers) should be translated as "бармактар"k.

2.4. Advantage: **Russian:** **Kyrgyz:**

2.5. Can these items be reconciled?

- Use the words "бармактар."

Discussion:

ZS: In the Kyrgyz language there are certain nuances and it is important to maintain certain norms in translation. In this example, pairs of Kyrgyz words need to be used in their singular form. Instead, the Russian way (plural) is utilized which does not follow the norms and rules of Kyrgyz during the translation. **NO**: "манжа" is the hand and fingers and includes the wrist, correct? **CJ**: No, it does not include the wrist, it is only the hand. This is not correct. **MD**: I think манжа is OK to use in an analogy item it if it is used in the singular form.

Evaluator Rubric (fully coded data)						
Item 15 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑					
Content				☑		
Format						
Cult/Ling.		☑		☑	☑☑☑	
Other						
						☑☑☑☑☑

2.3. Describe Differences in Detail:

Content:

- The word “бейкаруулук”k. (weakness) in distractor (B) is a word used rarely in Kyrgyz.
- The word “кубанычсыз”k. (*lit*. happiness + form for “without”- сыз) in distractor (Г) is not used in Kyrgyz. A better choice would have been “көңгүлсүз”k. (unhappy).
- The word “бейкаруулук”k. (weakness) is not often used.
- The word “кубанычсыз” k. (*lit*. happiness + form for “without”- сыз) is a created word (artificially created by test writers?).

Culture/Language:

- The word “бейкаруулук”k. (weakness) is not widely used in Kyrgyz and its use could result in a lack of understanding.
- It is possible that Kyrgyz kids in the city will not understand the word “бейкаруулук” (weakness).
- In city schools it is possible that the word “бейкаруулук” (weakness *eng*) will not be understood as it is not widely used in conversation; it is important to use words that are common in normal speech.
- Considering the correct answer, we need to make an accent on the distractor (B). The incorrect words “бейкаруулук – кубанычсыз” contradict logic and the grammar rules of the Kyrgyz language.
- The word “слабость”r. (weakness) can be translated as “али жоктук”k. (weakness)

2.4. <u>Advantage:</u> Russian: ☑☑☑☑☑ Kyrgyz:

2.5. <u>Can these items be</u> reconciled?

- Change the words “кубанычсыз” to “көңгүлсүз”k. (unhappy).
- Need to use more common words. Use “алсыздык”k. (weakness) instead of “бейкаруулук”k. (*no meaning*)
- Use the word “али жок”k. (weakness)
- Use a synonym.

<u>Discussion:</u>

MD: The problem here is poor translation and adaptation in two of the distractors (В, Г). **NO**: Yes, it is necessary to use “active” words which are used in everyday conversation. I finished a Kyrgyz school in Bishkek and “бейкаруулук”k. (weakness) is unfamiliar to me.

CJ: I agree with my colleagues that it is important to use commonly used words.

MD: Perhaps it would be possible to compare the translated Kyrgyz text with the original Russian text? That is, adjust the Russian text again if the translation into Kyrgyz does not seem to work? **KK**: Need to use commonly used words instead of literary terminology. Plus, I think there are some outright mistakes here, it is not just a problem of perspective. For example, “кубанычсыз” is just not said. If it is said, then this is a dialect problem.

Evaluator Rubric (fully coded data)						
Item 16 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content				☑		
Format						
Cult/Ling.		☑			☑	
Other						
						☑☑

2.3. Describe Differences in Detail:

Content:

- Was it really difficult to find a synonym in Kyrgyz for the Russian word “стул” (chair) in distractor “G”?

Culture/Language:

- “стул”r. (stool, chair) should not be translated into Kyrgyz as “Диван”k. (divan or couch)
- “Диван”k. (divan or couch) is incorrectly translated from the Russian “стул”r. (stool, chair).
- The word “стул” (stool or chair) is translated incorrectly into Kyrgyz. There is a Kyrgyz equivalent but it is not used.

2.4. Advantage: Russian: ☑ Kyrgyz: ☑

2.5. Can these items be reconciled?

- Look at the meaning.
- Use the word “отургуч”k. (stool) instead of “Диван”k. (divan or couch)

Discussion:

ZS: I had difficulty answering this item correctly. Not sure what the relationship is between “Автор” (author) - “писатель” (writer) - the item seems difficult.

MD: I think the relationship is one of general categories to a more specific part – that is, the second word in the pair is a part of the first category. However, I can see how the pair of words in the stem could be considered synonyms and thus make it hard to resolve. That is, I think the answer is “furniture: table” but I can see how they might have selected “journal: book” in Kyrgyz. Also, the Russian distractor “digit: number” is also somewhat attractive.

Evaluator Rubric (fully coded data)						
Item 17						
2.1. Difference Levels:	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑					
Content						
Format						
Cult/Ling.		☑☑☑				
Other						
						0

2.3. Describe Differences in Detail:

Format:

- “санда жок”k. (be in the minority) is the only compound word (two words) in the pair, which makes it obvious.

Culture/Language:

- City school (kids) might not understand the word “арзыбаган”k. (insignificant)
- The word “санда жок”k. (be in the minority) should have been translated as “кышындай”k.
- “арзыбаган”k. (insignificant) is unclear. Perhaps kids will translate it as “не достойный внимания”r. (not worthy of attention) which is slightly misleading.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Test the translation if he uses difficult synonyms.
- Use different words.
- Use different words with related relationships.

Discussion:

Evaluator Rubric (fully coded data)						
Item 18 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑					
Content				☑☑		
Format						
Cult/Ling.		☑			☑☑	
Other						
						☑☑☑☑

2.3. Describe Differences in Detail:

Content:

- Distractors (б) and (г) both have problems. In (б) “шашма”k. (hurried) does not correspond to the Russian version “откровенный” (open) and “шамдагай”k. (quick, nimble) is not the same as “болтливый”r. (talkative). In other words, neither word in this pair corresponds well to the pair in the other language. This distractor does not work. (г) also has an incorrect adaptation as “колдойгон”k. (attack) is used in simple speech, not as a literary term. Further, the meaning of the pair of words in Kyrgyz does not correspondence well to the meaning of the words in Russian.

Culture/Language:

- Translation is incorrect- (б). “шашма”k. = “торопливый”r. (hurried eng); “шамдагай”k. = “шустрый”r. (quick, nimble) not “talkative" as is given.
- “шашма:шамдагай”k. should be translated into Russian like “торопливый: ловкий” (hurried: dexterious), but not like “откровенный: болтливый” (frank: talkative) as is given.
- In distractor (б) “шашма”k. is “торопливый”r. (hurried) not “откровенный” (frank). “шамдагай” should be translated as “ловкий” (dexterious), or “шустрый” (quick, nimble) into Russian because “болтливый” (talkative) is “көп сүлүгөн.”
- “откровенный” (frank) is “ачык”k. and “болтливый”r. (talkative) - “көп сүлүгөн”k. or “сайранан”k.
- The word “шашма”k. (hurried) is translated incorrectly from the Russian as “откровенный” (frank): the correct translation of “откровенный” (frank) would be “ачык”k. (open)
- In distractor (В), “этият”k. is not the same as “осторожный”r. (careful). The translation is not accurate.
- Answer (г) in the Kyrgyz and Russian versions are completely different.
- (г) “колдойгон”k. is “большой”r. (big)”; “жүжүрөгөн”k. – “маленький”r. (small)” does not correspond to the given Russian pair.

- The words in (г) Russian are not translated correctly – the correct translation is “бысып келген”(to have come on foot, attacked) and “коргоочу” (defender).

2.4. <u>Advantage:</u> Russian: ☑☑☑☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

- Test the items first, consider the main idea.
- Use “этиятуу” (B).

<u>Discussion:</u>

ZS: There is incorrect, inaccurate translation in several of the distractors in this item and the use of the incorrect meaning of some words. **NO**: the problem is related to the specifics and nuances of the Kyrgyz language. The thing is, some words can be used only in combination with each other; in certain contexts they can’t be used individually. Therefore, this issue is poor adaptation.

RM: Yes, the problem is that some words must be used in combination.

KK: Yes, there is the incorrect use of some words in Kyrgyz. **MD**: the use of some words out of context makes them impossible to understand, they can not be used individually; the problem is adaptation.

RM: But, there are also simply grammar mistakes. The endings of some words are incorrect which means the students will not know what it all means - “тилеген”k. (desired) should not be used! Due to grammar mistakes the endings are not correct and they won’t know what it will mean. On the other hand, the item was not difficult to answer.

NO: In distractor (б) the translation is simply incorrect as “шашма”k. is not “откровенный”r. There are many such moments. Also, “мүмкүн”k. (possible, may) in distractor (A) without context is one meaning, used with other words it has different meanings.

Evaluator Rubric (fully coded data)					
Item 19 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**	**Somewhat Different**	**Different**	**Total Diff.**
	☑				
Content				☑	
Format		☑☑	☑		
Cult/Ling.		☑	☑		
Other					
					☑☑☑

2.3. Describe Differences in Detail:

Content:

- “тартып алуу”k. in the item stem has two Russian equivalents – “притянуть”r. (pull) and “отбирать”r. (take away). The Russian word used however, is “отдёрнуть”r. (draw back quickly), which is neither one.

Format:

- Number of letters. Instead of “укпай калуу”k. (go deaf) (correct) one letter is absent from the word.
- “укпай калу_”k. (go deaf) – one letter is missing. In Kyrgyz, the infinitive form is created with affixes (“уу” in this word).
- There is a typing mistake in distractor (г) - “укпай калуу”k. (go deaf) is missing a letter; And (б) the correct form is “жабышчак”k. (sticky) instead of “жабышкак”k.
- In the Kyrgyz language, “липкий”r. (sticky) is correctly translated like “жабышчак”k.

Culture/Language:

- The word in the item stem, “тартып алуу” has two meanings in Kyrgyz. These meanings are “отбирать”r. (take away) and “притягивать”r. (pull). However, the Russian stem given is actually “отдёрнуть”r. (draw back quickly). In the Kyrgyz stem, in combination with the first word in the pair “ысык”k. (hot), the analogy can be understood as “ысык тартуу”- that is “притягивать тепло” (attract warmth). In this case, the correct answer will be distractor (б).
- “жабышчак”k. (sticky) – from “липкий”r. (sticky) is not the best translation. A more accurate one will be “батамашкан, батталуу”k.
- Incorrectly translated equivalent of “зажмуриться”r. (tightly closing the eyes) the correct translation is “көздү бекем жүмүү”k. (tightly closing the eyes).

2.4. <u>Advantage:</u> Russian: ☑☑☑☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

- More accurate translation.
- Add a letter.
- Check and recheck the translation.
- It can be adapted to the understanding of students.

<u>Discussion:</u>

RM: the multiple meaning of some words in the item stem means that there needs to be a more careful selection of pairs of words – otherwise, the item misleads and it becomes impossible to find the correct answer. There seem to be some misprints as well. **NO**: I agree, depending on how they define the terms in the stem they could come to complete opposite meanings of the analogy. The given Russian word is "отдёрнуть"r. (draw back quickly). which implies a "pushing away from" while the Kyrgyz equivalent, "тартып алуу", can mean "притягивать"r. (pull) and might even be interpreted as "pulling towards." So, depending on how they interpret the meanings of the words in the stem pair, their answers might be different. **MD**: Yes, the stem needs to be more clearly defined (contain no double meanings). **MK**: Absolutely, the stem and distractors should have only one meaningful interpretation.

NO: Also, the word used here in the Kyrgyz distractor (б) "жабышкак"k. (sticky) is unknown to me. Perhaps this is some form of dialect? **MD**: Yes, this is a word but it is not so widely used, and one letter is incorrect. **CJ**: Yes, but even with the correct spelling this is a word but the problem here is related to dialect use. According to the context though, the Kyrgyz will understand "something sticky."

AA: There is a difference in nuance with the word "зажмуриться"r. (squint) and the Kyrgyz equivalent. **MK**: Does the Kyrgyz term mean squinting or blinking? **AA**: In Kyrgyz it reads as "sneaky look" i.e. a dangerous person. **MK**: I think there is also a difference between blinking vs. squinting and the connotations of these words (negative connotation of evil for the Kyrgyz item). **MD**: The translation makes it an expressive, stylistic colorization, but in the pair of words – it works – as the relationships are maintained. All: There are two problems with this item– uncommon words as well as words with multiple meanings. There is a difference here between these two analogy pairs and that will probably impact the results.

Evaluator Rubric (fully coded data)						
Item 20	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
2.1. Difference Levels:						
	☑☑☑☑☑☑☑☑					
Content						
Format						
Cult/Ling.						
Other						
						0
2.3. Describe Differences in Detail:						
2.4. Advantage: Russian: Kyrgyz:						
2.5. Can these items be reconciled?						
Discussion:						

Evaluator Rubric (fully coded data)						
Item 21 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑					
Content					☑☑☑	
Format				☑☑		
Cult/Ling.		☑		☑		
Other						
						☑☑☑☑☑☑

2.3. Describe Differences in Detail:

Content:

- The translation was done literally and this resulted in a problem with the Kyrgyz variant. Therefore, for a clearer understanding – the word “жана”k. (and) needs to be changed to “ошону менен”k. (with this, together with this) or “ошол эле убакта”k.(at the same time) or “бирок”k. (but).
- The translation was done correctly but the main idea was lost.
- It would be better to change the word “болбойт”k. (will not) to “жөнөтөт”k. (send) as before.
- “жумушка орношкондон кийин”k. implies he had “возможность” (possibility). It follows that the more correct form of the last phrase in the stem will be “мүмкүнчүлүгү жөнөтөт”k. (possibility to send).
- The phrase from Russian “определенное количество**”**r. (certain quantity) is translated incorrectly into Kyrgyz. The word “канчадыр”k.(some) makes the question unclear and could mean lost time for the student.
- The word “определенный”r. (determined) is not the same as “канчадыр”k.(some)

Format:

- The spacing of the blanks in the two different versions is different to the advantage of the Russian version. Differences appeared due to the adaptation. The instructions in Kyrgyz are not clear. In the translation, the punctuation is incorrect.

Culture/Language:

- Need to use “же”k. (or) or “бирок”k. (but) instead of “жана”k. (and).
- Literal translation. Instead of the connection “жана”k. (and) it is necessary to use “ошону менен бирге”k. (together with this) because the connector doesn’t sound Kyrgyz, but sounds Russian.

- It is obvious that this sentence was translated completely from Russian in Kyrgyz.

2.4. <u>Advantage:</u> Russian: ☑☑☑☑☑ Kyrgyz: ☑

2.5. <u>Can these items be reconciled?</u>

- Need to use “же”k. (or) or “бирок”k. (but) instead of “жана”k. (and)
- Need to use “бирок”k. (but) instead of “жана”k. (and)
- Don’t translate literally but make an adaption - use “Өз убактысын бир бөлүгүн / белгилүү бир бөлүгүн”k. (part of your personal time/part of some time)
- “Канчадыр бир убакытты”k. (for some time) - is a very scientific style and could be changed.
- Of course in order to solve the problems with this item, experienced (the best) translators are needed. In selecting the text many factors need to be considered if it is important how well the task is completed.
- Instructions need to be clear.

<u>Discussion:</u>

MD: I think this item needs to be completely changed as it will not be easy to simply adapt. The main problem is the incorrect use of the term “жана”k. (and) which is obviously the result of a direct translation from the Russian sentence. **AA**: I agree, but the problem is that this is a common usage in Kyrgyz. It’s on the radio, the national TV stations and other official media sources. **MD:** It is common but it is not correct. **AA**: I understand.

MD: I believe this usage is a one of those “Russianisms” that has crept into Kyrgyz through (ethnic) Kyrgyz, Russian language speakers. The main problem is that in Kyrgyz we don’t use “and” as a connector when connecting two different verbs. Two verbs together often come together to convey a different meaning than when they are used singularly. The two verbs are simply put together, without the use of any connectors.

ZS: I agree with **MD**, villagers don’t use “жана”k. (and) in this sense – i.e. they use Kyrgyz correctly. To me, this raises the issue of adaptation. It seems that this item was clearly adapted from Russian. I think if Kyrgyz original texts had been used, there wouldn’t be

this problem. We could avoid syntax problems like this. Our syntax is different and should be Kyrgyz – not Russian. **MD**: Well, theoretically, I agree of course. A big problem is that much of our literature in the sciences and the arts is translated in this way – translated directly from Russian. Much of the Russian influence is inevitable. Little is produced in Kyrgyz due to a lack of specialists and resources. **ZS**: OK, but what if we had several specialists work on developing the items at the same time and then decide whether they will work or not?

MD: To me, this item raises a bigger question from the perspective of the test translators. Should the items can contain only language that is 100% correct or contain language that is incorrect but commonly used? Unfortunately, there is often a gap here. The situation and state of the Kyrgyz language is very sad. Further, we have to take into account language as it is used on a daily basis. Many people in the cities – and not only in the cities – speak Kyrgyz with lots of words and forms taken from Russian.

Sometimes the language is simply all mixed up. This is a result of the language environment we live in. We combine Russian and Kyrgyz all the time in a sort of hybrid colloquial language. For example, "канча (how many - *kyr*) листов (lists (pieces) or paper - *rus*)?" or "сиз (you *kyr*) домой (home *rus*) барасызбы (are going *kyr)? – lit. Are you going home?* There are hundreds of ways we do this. This item raises some big issues.

Evaluator Rubric (fully coded data)						
Item 22 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑					
Content						
Format		☑				
Cult/Ling.				☑☑		
Other						
						☑☑

2.3. Describe Differences in Detail:

Content:

- Perhaps the *abiturients* did not learn at school that the tree “баобаб” (baobab) grows in Kyrgyzstan (especially kids that study in rural areas) and therefore finding the answer might be difficult.
- The antonyms used in the distractors make the item difficult.
- Regional differences should be taken into account.
- These Kyrgyz texts need to be developed very carefully. This is because the negative form given is located in the word itself. The distractors are very different as well – in Kyrgyz they are very long. There is no reason to use this example.

Format:

- Typographical error.
- Poor equivalence of Russian and Kyrgyz distractors. In distractors (A) & (г) “плотные”r. (dense) is not the same as “өтө ык”k. (very strong). In distractors (B) & (б) “пористый” (porous) is not a good match for “майда тешиктүү”k. (with small holes). The word “келген”k. in the stem creates a correct combination with “өтө ык”k. (very strong) in (A) & (г) but is worse fit with “майда тешиктүү”k.(with small holes) in (B) & (б).

Culture/Language:

- The tree “баобаб” (baobab) in the item stem does not grow in Kyrgyzstan and is unfamiliar to many students. The translation is correct however.
- The Kyrgyz version takes more attention to solve than the Russian version. The antonyms in the distractors, (б) for example, are longer, more complex in Kyrgyz than in Russian. Please do not use this task.

2.4. <u>Advantage:</u> Russian: ☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

<u>Discussion:</u>

RM: I found this item to be poorly adapted. In many ways, the Kyrgyz and Russian versions are not equivalent in meaning. Several of the distractors contain words which convey different meanings. **KK**: I agree, there is a problem of psychological understanding the text and some words in Kyrgyz, it does not seem natural. There are stylistic problems here. **MD**: There is a problem of a lack of knowledge of some specific words on the part of students and some awkward terminology. **AA**: In my opinion, the problem is not a lack of knowledge as they cover this in school, but instead a lack of attention to language detail on the part of the item writers.

MK I think the content is also at issue. The word “баобаб” (baobab) will not be understood by many students – I studied the item for several minutes. This seems like a word that only biologists will know – it is too technical.

AA: Actually, I disagree, because the word does not make the kids miss the logic here. They don’t need to know that word to resolve the logic of the sentence completion. Plus, I think they learn this term in biology – it is covered in Kyrgyz schools.

Evaluator Rubric (fully coded data)						
Item 23 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑ ☑ ☑					
Content		☑		☑ ☑		
Format						
Cult/Ling.					☑	
Other						
						☑ ☑ ☑

2.3. Describe Differences in Detail:

Content:

- The grammatical mistakes in this sentence demand time to understand the main idea of the sentence. The literal translation of the word “возмещение” (compensation) in the item stem does not fit the given situation (in Kyrgyz) and creates a false association.
- In the Kyrgyz item stem, the words “айдап кетишкен учурда” is not the equivalent of “угон”r. (hijacking, car theft) in Russian. This is a mistake.
- The word “камсыздандыруу” (provision, guarantee, insurance) in the item stem has no real meaning in Kyrgyz.

Culture/Language:

- “камсыздандырылган”k. (be provided for, insured) in all the item distractors is not an accurate translation but kids can understand from the context.
- In Kyrgyz, the word “камсыздандырылган” has a wider connotation than “обеспечение” (provision) than “застрахование”r. (insured) does. The word “страхование”r. - камсыздандыруу”k. is typically used amongst specialists.
- The word “застрахованный”r. (insured) – and “незастрахованный”r. (uninsured) are incorrectly translated from Russian into Kyrgyz (differences in translation).

2.4. <u>Advantage:</u> Russian: ☑☑☑ Kyrgyz:

2.5. <u>Can these items be reconciled?</u>

- Adapt or translate correctly. Change the word “тургузулбайт”k. (will not be provided) to the word “тойлойбойт”k. ()
- Find a synonym for the word “страховать”r. (to insure)
- Yes, for “угон”r. (hijacking, car theft) use the words “урдап кетуу”k. (stolen) in the Kyrgyz version.

<u>Discussion:</u>

CJ: I had difficulty reading and understanding the Kyrgyz text; then, I read the Russian text and I understood. I think that with background knowledge (knowing Russian) they might be able to understand some of the meaning. However, Kyrgyz-only speakers will find it confusing. That is, if the Russian concepts are covered first, then one knows what the Kyrgyz authors meant to say. However, the students do not get this advantage because they do not know the Russian version. **AA**: We (item analysts) have an advantage because we can read both items at the same time! And, “insurance” is rarely used in Kyrgyz terminology. So, the text is not well adapted, and this makes it difficult to comprehend.

KK: Some words (used) in this item can only be understood in specific contexts. There are both translation and adaptation problems.

RM: the problem is adaptation, not translation. **AA**: During task creation, it is necessary to consider many nuances and the commonality of certain words. And also the weak vocabulary of Kyrgyz speakers.

KK: The desire to pass on the main idea of the task was too much, as there is the possibility of loosing the literary nuances of Kyrgyz; important to consider the characteristics of the language – in Kyrgyz sentences are usually short – as words are “complex” (compounded – i.e. agglutinative), and the result of the direct translation is that translated texts (from Russian into Kyrgyz) are longer than usual for Kyrgyz speakers. As they become longer, they become more confused. And, the sentences eventually become even longer than the Russian versions. I believe that it is possible to find original texts in Kyrgyz and then translate them into Russian - then you will see the richness of differences of the languages. **ZS:** Yes, the problem is the translation and the adaptation due to stylistic differences.

MK: The problem is the terminological meaning of the words and the difficulty of the sentences in Kyrgyz. **AA:** Yes, there are not enough words in Kyrgyz for some of these concepts. Kyrgyz in general, has fewer words. On the other hand, if items are limited to only what everybody knows…

MD: This is not a question of commonality of the concept of "insurance" - which is also new to Russian speakers, but rather use… In Russian the item is simply easier… To be honest, many of the evaluators could fully understand only after they read the Russian version of the item. **RM**: To me, this indicates that the items are not being prepared with enough consideration for the Kyrgyz language. I think the items should be written in Kyrgyz first, then have the Russian adapt to the Kyrgyz version. Why can't this be done?

ZS: I agree with **RM**. One adaptation suggestion would be to not have the Kyrgyz sentences copy the Russian style but to make them "more Kyrgyz." This means making the sentences shorter, even if it means more sentences. That is the first point. Also, the issue of unknown concepts is also important here. Many new terms are constantly being formed all the time in Kyrgyz while in Russian the concepts are well known. For example, in Kyrgyz there are four or five completely different ways to say "entertainment center." People do not know which are correct at this point. Therefore, Kyrgyz often use Russian loan words. Some use Kyrgyz words that are not known. As teachers we see this on a regular basis. Many words are "created" but not yet well known. In this item not everyone knows how to say "uninsured" especially in rural areas where there is no such thing as insurance.

MD: **ZS** eje, I agree with your first point – in Kyrgyz we have "long words" but short sentences. This is important to remember in comparison with Russian. Also, in regard to the point about word use, as I said in regard to other items we often mix languages in such cases. For example, it is common to say "Страховка"r. (insurance) "барбы"k.? (Do you have insurance?). These Russian-Kyrgyz hybridizations are OK in everyday speech but obviously become problematic in testing.

AA: Also, "угон" (hijacking, car theft) is not understood unless the whole context is known. Otherwise in Kyrgyz it is understood as the "the car was left open" not *stolen* as is clear from the Russian version. **NO**: I disagree, because some other words provide hints. The problem is that it needed to be adapted fully, not simply translated – it is too literal here.

Evaluator Rubric (fully coded data)						
Item 24 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑					
Content				☑		
Format						
Cult/Ling.				☑		
Other						
						☑☑

2.3. Describe Differences in Detail:

Content:

- The stem is too long in the Kyrgyz version. Due to agglutination, sentences in Kyrgyz are usually kept shorter than Russian sentences. When direct translation comes from Russian the result is sometimes high complexity and difficulty in understanding.
- Kyrgyz understand the word “кругозор” (outlook) (a Russian word), but will not understand the attempt to create a new Kyrgyz equivalent.

Culture/Language:

- “ар тараптан”k. – “с разниых сторон”r. (from a different sides); “ар тараптуу”k. – ““разносторонний”r. (multi-sided).” It seems to me that the word “өнүт”k. might be misunderstood by the general mass of examinees.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Shorten the stem in Kyrgyz and refrain from making up new Kyrgyz words.
- Change the word “өнүт” to the word “тарап”k. (side)

Discussion:

AA: the main problem is that uncommon words are used.

Evaluator Rubric (fully coded data)						
Item 25 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑					
Content				☑		
Format						
Cult/Ling.				☑☑☑		
Other				☑		
						☑☑☑☑☑

2.3. Describe Differences in Detail:

Content:

- The word “Пас”k. (low, down) is dialect and may not be understood.
- The word “Пас” k. (low, down) is dialect and may not be understood by many.
- During the process of translation a syntaxical mistake was made.

Culture/Language:

- The word “Пас” k. (low, down) is dialect.
- In city schools and in the northern regions, the word “Пас” k. (low, down) is not used, in the south they may understand this word.
- The word “Пас” k. (low, down) is dialect which is not known to kids from the north.
- The word “Пас” k. (low, down) is dialect, northern kids may not understand.

Other:

- This question demands specific knowledge. Who knows, maybe scholars have demonstrated that high voices are more difficult to understand or the perhaps the opposite? That’s why in both the Russian and Kyrgyz versions it could be difficult to answer correctly.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Yes, use: эркектердүн үнүн кабыл алуу жеңил болуш керек эле.
  Натыйжада аялдардын үнүн кабыл алуу жеңил болуп чыкты.
- Yes, use: Окумутруулар үндүн бийигирээк тембрин кабыл алуусун кыйындатат деген тяанакка келишти. Аялдардын үндөрү демейде эркектердикинен бийигирээк келет, ошондуктан алдардын кебин укканда, анны түшүнүүгө кыйынраак болуп калат.
- Change “Пас” to “төмөн”k. (low, down)
- Maybe use the word “төмөнүраak”k. (lower)
- Instead of “Пас” use the word “төмөн.”k. (low, down)

Discussion:

NO: The problem here is the use of dialect instead of general use, literary language. In order to understand the task, it is necessary to understand the meaning of the text. It is possible that kids will understand the opposite of what is meant. **MK**: the dialect used here is common in the south of the country.

Evaluator Rubric (fully coded data)						
Item 26 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑ ☑ ☑ ☑ ☑					
Content				☑		
Format				☑		
Cult/Ling.		☑				
Other						
						☑ ☑

2.3. Describe Differences in Detail:

Content:

- “производство”r. (production) is translated as “өнүмдөрдү чыгаруудан”k. which actually changes the situation and leads to a lack of clarity in the Kyrgyz item stem.
- In distractor (A), the words “чыгашалар көбөйгөн” are incorrect; should be “… көбөйгөндүктөн”k.(because of the increase in something).
- In the process of translation there is a mistake in formulation (б) too.
- It is not possible to translate literally texts of this size and content. No good result will come from this. It is very important to carefully select the texts. The word “зыяны”k. (harmful) in the stem makes it difficult to understand the text and forces one to read it again.

Format:

- In my opinion, this is an example of when an item was obviously translated from Russian into Kyrgyz.

Culture/Language:

- The words “баш тартуу”k. (refuse, avoid) in the stem has different meanings from the Russian version.
- The suffix on “себептүү”k. (due to) needs to be changed to a different word (use suffix to form possessive). This will result in fewer words, difficult sentence, but also will not change the meaning (it will improve it).

4. <u>Advantage:</u> Russian: ☑☑☑ Kyrgyz:

5. <u>Can these items be reconciled?</u>

- Yes, translate the word “производство”r. (production) as “өндүрүш”k. (production)
- When the texts are large, they need to be adapted not translated directly. An offered solution “ишканалар үчүн экологияга зыяны азыраак өнүмдөрдү чыгаруудан баш тартуу максатка ылайык эмес, анткени кирешелерди көбөйтүү үчүн бул аларга пайда көрсөтөт”k. (It is not reasonable for the enterprises to deviate from producing products that would be less harmful for environment, because they can be useful for increasing their profit).
- An offered solution: Кирешелер көбөйгөн себептүү (because of the increased profit) – кирешелердин көбөйтүүсү (increase of profit).
- An offered solution: 1. “Экологияга зыяны азыырак”k. (less harmless for environment) – “экологиялуу таза азык-түлүк же…”k. (environmentally safe ("clean") foods or...). The word harm hinders comprehension of the sentence right away, so one needs to read it again. 2. “чыгашалар көбөйгөн себептүү”k. (because of increased expenditures) should be – “чыгышалар көбөйгөндуктан”k. (in connection with increased expenditures).

<u>Discussion:</u>

ZS: “себептүү”k. (due to) in item stem is not needed. It needs a different affix here. **MK**: I do not agree… without this, the item loses the main idea. What will be the correct answer? **ZS**: this item is confusing, the translation is not clear in several places.

MD: Hmmm… It seems that the Russian text allows a “double meaning,” and Kyrgyz only one meaning. However, that meaning (for the Kyrgyz item) leads to a wrong answer. This is due to the way the item is structured. **MK**: What *is* the correct answer to the Kyrgyz item? **MD**: the complication is over the meaning of the word “production” which is quite unclear. **ZS**: If we can’t find the correct answer I don’t think the children will either!

Evaluator Rubric (fully coded data)						
Item 27 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑☑					
Content						
Format					☑	
Cult/Ling.						
Other						
						☑

2.3. Describe Differences in Detail:

Content:

- Grammar mistakes in sentence formulation. In distractor (A), replace “анткени”k. (because) with “себеби”k. (because).

Format:

- Typographical error – instead of “арзан”k. (inexpensive) it is written “арзар”k. (*no meaning*)

2.4. Advantage: Russian: ☑ Kyrgyz:

2.5. Can these items be reconciled?

- Check the spelling.
- Yes, I propose: «Японияда квалификациялуу эмгек кымбат турат, себеби анын ашыкчылыгы аныk, ошондуктан Япония квалификациялуу эмгекти талап кылган арзан товарларды өндүрүп чыгарууга жөндөмдүү эмес». (In Japan, highly qualified jobs are paid appropriately (well), because they are limited; therefore, Japan doesn't need to develop cheap products necessary for highly qualified jobs).

Discussion:

All: There are a few typographical errors but from the context you can understand the meaning of the words in both items.

Evaluator Rubric (fully coded data)						
Item 28 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑☑					
Content						
Format						
Cult/Ling.		☑				
Other						
						0

2.3. Describe Differences in Detail:

Content:

- Incorrect content in distractor (A) because of the incorrect use of the words “байланыштуу”k. (to connect) and “жана”k. (and). The translation should like look this (see below):
- There is a grammar mistake in item stem of the Kyrgyz version; instead of “прогноздого,” should be “прогноздорго”k. (prognosis)

Culture/Language:

- In distractor (A) of the Kyrgyz version, “жол бербеген”k. (to hinder) should have been “тоскоол болгон” k. (to hinder)

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- The correct version of the Kyrgyz stem is “прогноздорго”k. (prognosis)
- I recommend “Атмояферада жүрүүчү процесстер өтө татаал болгондуктан, атмосферада аба-ырайына так прогноздорго жол бербеген кубулуштар болушу мүмкүн” (Perhaps because processes taking place in our environment (atmosphere) are very complex, therefore, it is hard to make prognoses about the weather).

Discussion:

ZS: In the Russian school classes there are many Kyrgyz language students. Many of them speak in Kyrgyz at home. They don't know Russian well as for them it is not their native language. I finished Russian school but I intuitively understand the Kyrgyz test better. That is why native language is easier. **MD**: **ZS** eje, Well… we are speaking about something else here because the students answer only one version of the items, they aren't asked to answer both versions…

ZS: In general, syntax is easier in Kyrgyz than Russian. We have a straightforward "cause – result." The structure of Russian is more difficult.

MD: Textbooks are mostly translated. We see the same issues in our textbooks at schools… I think that there are several levels of structure in Russian. In Kyrgyz, it is "single level" – it is this "and" this "and" this. New ideas are "added" while in Russian there is a different structure. In Kyrgyz it is all part of the same syntactical level. **ZS**: My Kyrgyz students also tell me that the Russian constructions are difficult too learn at first. In Russian you have "Due to the fact …. Because of the fact …" in Kyrgyz, more direct statements.

MK: Yes, for example, in Russian you may have … "event/phenomena … which is/that is…" etc. In Kyrgyz we have "this happened" (stop) and "that happened" (stop) and then something else. It is all on "one level." **ZS**: In general, it is easier in Kyrgyz.

MD: The challenge for test writers is that for some Kyrgyz texts it becomes complicated when we try to repeat the Russian syntax and constructs. It becomes complicated when translation is literal. The best way to keep the Kyrgyz intact is to break the Russian sentences into more sentences rather than trying to capture the Russian structure. In Kyrgyz, ideas are built not through one complex sentence, but through a series (many sentences) with simpler ideas that when compounded, express the same idea.

ZS: Related to this, we often start to translate from the end of the sentence because the main idea comes last (the verb is at the end). Word order is different in Kyrgyz than Russian which can also cause complexities. Because of the word order, sometimes, I translate the literal sentences first – then rearrange them in order. (*The*) Strategy here is to read sentences more times or to hear it in Russian first, and then piece together the puzzle.

Evaluator Rubric (fully coded data)						
Item 29 **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 my opinion, it is necessary to pay attention to the form of the verb “деген экен”k. (turns out).
- In the stem, the pronoun “аларга”k. (to them) is used incorrectly. The correct form would be “аларды”k. (their). There is a similar grammar mistake in the word “адистерди”k. (at the specialists). Here, the more correct form is “адистер менен”k. (with specialists).

Culture/Language:

- In the Kyrgyz version the words “деген экен”k. (turns out) are used while in the Russian version the words “говорил”r. (he said) are used. “деген экен”k. (turns out) actually means “оказываться”r. (turns out)

2.4. Advantage: Russian: ☑ Kyrgyz:

2.5. Can these items be reconciled?

- Don’t use the word “экен,” or use the word “айткан,” or “айтыртып.”
- Yes, make the following corrections: change “Аларга”k. (to them) to “аларды”k. (their), and “Адистерди”k. (at the specialists) to “адистер менен”k. (with specialists).
- An improved version: Генрид Форд мындай деп айтыптыр - “Эгерде мен өз атаандаштарымдан озуп өтүүнү кааласам,

анда аларды жакшы адистер менен камсыз кылмакмын, себеби эң бир мыкты идеядан кемсилик табу жана ошонун аркасы менен анын иш жүзүнө ашырылышына жолтоо болккшкна алардын колунан келет.”k. (Henry Ford said: “if I want to be ahead from my rivals, then I wish them good workers, because they will be able to find weaknesses of the best idea and by this they will interfere with its realization).
Discussion:

Evaluator Rubric (fully coded data)						
Item 30 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑					
Content						
Format				☑		
Cult/Ling.						
Other						
						☑

2.3. Describe Differences in Detail:

Content:

- There are differences in the use of grammar in the word “болуусу”k. - The correct form is “болууш”k. (official).
- Another grammar mistake is the incorrect form of “Жайлатат” – a better form is “жай кылат.”

Format:

- Orthographical mistake (spelling) “болууш”

2.4. Advantage: Russian: Kyrgyz: ☑

2.5. Can these items be reconciled?

- “тоскоол”k. (hindrance) instead of “тоскол.”
- Yes, in the Kyrgyz stem, do not use **“**Болуусу” but болушу. Also change “Жайлатат”k. (slow down) – to жай кылат (do something slowly).
- Another recommended change: “Туруксуз табигый түрлөрдүн пайда болушу эволюциялык процесстерди тай кылат. Түрдүн туруксуздугу анын өнүгүсүүнө тоскол болбойт деген ой-пикир бул фактыга карама-каршы келбейт.”k. (The existence of the unstable natural types will create an evolutionary process. The unstableness of the type will not be an obstacle to its development, such opinion will not contradict this fact).

Discussion:

Evaluator Rubric (fully coded data)						
Item 31 **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 distractor (г) there is a mistake. The Kyrgyz word combination “жээк аймактарында”k. (at the coastal territories/areas) is not used at all.

Format:

- (б) and (В) are not in the same place for the items.
- (б) and (В) are not in the same order.
- (б) and (В) are in different places in the items.
- In Kyrgyz distractor (г), “жашагандарды”k. (people living) should be “жашагандарга”k. (to the people living)

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- In order to correct distractor (г), you need to use “жээкде”k.(at the coast) instead of the above.
- Look at the order

Discussion:

Evaluator Rubric (fully coded data)						
Item 32 **2.1. Difference Levels:**	**No Diff.**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑					
Content					☑	
Format					☑	
Cult/Ling.						
Other						
						☑☑

2.3. Describe Differences in Detail:

Content:

- In the Russian version a "concrete" task is given – that is, something that can be related to. In the Kyrgyz version, it asks for the how the ideas are connected (between the sections).

Format:

- Not complete, and incorrect word combination. Need to remove the Russian letter "и"r. (and) from the Kyrgyz version.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2,5. Can these items be reconciled?

- For distractor (Г) in the Kyrgyz version: «Натыйжаларды жана алардын келип чыгуу шарттарын сыпатто"k. (analyze the results and their origins)

Discussion:

KK: The question is not that clear.

ZS: Actually I don't agree.

MD: there is a small typo - “и” is there in the Kyrgyz version instead of with “жана” but I don’t think that should impact results. Besides the Russians answered worse and this is a problem with the Kyrgyz item.

ZS: I would have projected that the item favors Russians because the text in Kyrgyz has some problems that we already discussed.

MD: There is a difference in the way the stems are worded, structurally speaking. I mean, the Russian version requires students to “complete the sentence.” The sentence goes on and then stops, the four distractors each represent a possible continuation of the idea. For the Kyrgyz item, it is actually a complete question with a question mark at the end. I think this issue exists. When there is a question- answer – it might be easier than when you have to “build” a sentence. In some way this might make it easier to solve than the Russian item but I am not sure about that, the distractors all seem pretty clear. I didn’t notice this the first time we analyzed the items, I only now- with more careful inspection - see these differences.

ZS: I think “b” is correct. I am not sure if I agree with **RM**’s proposed change for one distractor but… **MD**: As far as **KK**’s recommendation – I think we should be sensitive to the fact that when you change grammar you sometimes you loose the main point of the item… Actually, I can’t seem to find any explanation for why this item might be harder for Russians except what I mentioned before about differences in the structure of the stem (question) itself – complete question vs. fill in the blank.

Evaluator Rubric (fully coded data)						
Item 33 **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 distractor (A) “Подробно”r. (in detail) is translated as “жеке”k. (special, separate) which is incorrect.

Format:

- The form of the question (stem) in Kyrgyz is incorrect.

Culture/Language:

- In distractor (A), “жеке”k. (special, separate) is translated as “частный”r. (private) which is incorrect.
- Distractor (A) uses the word “жеке”k. instead of the word “тагыраак"k.” (more precisely). “жеке”k. is translated incorrectly as “частный”r. (private)
- The word “подробно”r. (in detail) is translated as “жеке”k., which could mean “отдельный”r. (separate)

2.4. Advantage: Russian: ☑☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Examine the translation and change the stem. Remove “Кандай болгонун”k. (about its state) and replace with “төмөнкүдөй билдирет”k. (states as below)
- Use the word “тагыраак"k.” (more precisely) in distractor (A) instead of “жеке”k. (special, separate)
- Use “толук”k. (complete/holistic) in distractor (A).

Discussion:

MK: There is some incorrect translation (like distractor (A)) with this item, though the overall meaning of the text is similar in Russian and Kyrgyz. The main difficulty arises from the multiple meaning of some words, and the difficult word combinations of some long sentences.

ZS: There are many problems with the translation of this difficult text; it is not well adapted. One resolution is to take an original Kyrgyz language text, related closely to this theme and the Russian text, because it is difficult to completely pass on the entire meaning and deeply consider the question in a foreign language. **NI**: It is very difficult to find original texts in Kyrgyz.

MD: It is easy to see the lack of connection due to translation issues. I think that the analytical thinking on the part of the Kyrgyz is different. **ZS**: I did not find any difficult words or issues with the item itself. Maybe some issues with the form of the sentences (constructions – syntax) in the reading text though. It is clear that the key is (B) but (G) is also an attractive answer.

ZS: There are many translation problems in the reading text. I want to reiterate that test writers should select Kyrgyz tests first and then adapt them. I must say that the Russian text is quite good, as are most of the items in Russian. I can't find any difficult words, grammar mistakes, etc. but syntax issues might explain any differences.

Evaluator Rubric (fully coded data)						
Item 34 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑☑					
Content						
Format						
Cult/Ling.				☑		
Other						
						☑

2.3. Describe Differences in Detail:

Culture/Language:

- The word “каржылардыk. (finances) in the Kyrgyz distractor (A) would be better translated as “Каражаттардыk. (means/resources)

2.4. Advantage: Russian: Kyrgyz:

2.5. Can these items be reconciled?

Discussion:

Evaluator Rubric (fully coded data)						
Item 35 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑					
Content						
Format		☑☑☑☑			☑	
Cult/Ling.						
Other						
						☑

2.3. Describe Differences in Detail:

Format:

- (б) and (B) have changed places.
- (б) and (B) are in different places.
- (б) and (B) are in different places in the Kyrgyz version.
- The word “төмөн жакта”k. (below) in the Kyrgyz stem is incorrect. The correct word is “төмөндө”k. (below).
- The distractors (б) and (B) are in different places in the two versions.

2.4. Advantage: Russian: ☑ Kyrgyz: No Advantage: ☑

2.5. Can these items be reconciled?

- Examine the items.
- Change “төмөн жакта”k. (below) to “төмөндө”k. (below).”

Discussion:

Evaluator Rubric (fully coded data)						
Item 36 **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 task the sentence is not correctly formed in Kyrgyz. It is long and difficult to understand.
- The translation of the question is poor. The word order is bad, structure is too complex, it's difficult to read and understand.

2.4. Advantage: Russian: ☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- What did the author want to demonstrate using the example of measuring the temperature of a healthy person?
- The question should be expressed in a different way, reworded.

Discussion:

Evaluator Rubric (fully coded data)						
Item 37 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑☑					
Content						
Format		☑				
Cult/Ling.						
Other						
						0

2.3. Describe Differences in Detail:

Format:

- There is an extra affix (ending) in the word “эмнеде”k. (what *at what location*)

2.4. Advantage: Russian: Kyrgyz: No Advantage: ☑

2.5. Can these items be reconciled?

- “Эмнеде”k. (at what location) should be just “эмне”k. (what).

Discussion:

Evaluator Rubric (fully coded data)						
Item 38 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content				☑		
Format		☑		☑	☑	
Cult/Ling.						
Other						
						☑☑☑

2.3. Describe Differences in Detail:

Content:

- Incorrect form of the sentence in this task makes understanding the main idea difficult.

Format:

- There is one word missing that is needed. Instead of “токтуу”k. (*no meaning*) – “токтотуу”k. (to stop) is needed.
- There is no such word as “токтуу”k. (*no meaning*). It should have been “токтотуу”k. (to stop)
- In distractor (б) there is a missing letter. It should be “токтотуу”k. (to stop) but “токтуу”k. (*no meaning*) is written.

2.4. Advantage: Russian: ☑☑☑☑ Kyrgyz:

2.5. Can these items be reconciled?

- Yes, use: «Автор эмнеден улам геоинжиниринг долбоорлорунун ийгиликтүү жүрө тургандыгынан күмөн санайт?»
- Edit and check again and again.

Discussion:

Evaluator Rubric (fully coded data)						
Item 39 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑☑☑☑☑					
Content						
Format						
Cult/Ling.						
Other						
						0
2.3. Describe Differences in Detail:						
2.4. Advantage: Russian: Kyrgyz:						
2.5. Can these items be reconciled?						
Discussion:						

Evaluator Rubric (coded summary data)						
Item 40 **2.1. Difference Levels:**	**Identical**	**Somewhat Similar**		**Somewhat Different**	**Different**	**Total Diff.**
	☑☑☑☑					
Content						
Format		☑☑☑				
Cult/Ling.						
Other						
						0

2.3. Describe Differences in Detail:

Format:

- (B) and (r) are not in the place.
- (B) and (r) are mixed up.
- (B) and (r) are not in the sample place.

2.4. Advantage: Russian: Kyrgyz:

2.5. Can these items be reconciled?

- Need to check carefully

Discussion:

Item Analysis Rubric (Reading Comprehension Text)

- In Kyrgyz, “Global Warming” is hard to understand (the title) of the text.
- In my opinion, attention needs to be paid to the size of the text of the Kyrgyz version, because too big of a text can cause difficulties. In general, I have only a few comments about the translation of the text. See text lines: 1, 20, 40, 45, 105, where hyphens are missing which might hinder some understanding.
- Text: there is an extra hyphen between words in some places. In the lines 19, 22, the word “не- гизделген” is written with a hyphen, and “метр-ге”k. is also; this is incorrect.
- The Kyrgyz text had a lot of words, and needs a more careful adaptation. The text needs to be adapted, make the Kyrgyz text shorter so that they will be equivalent.
- In the lines 64-69 (Kyrgyz) the text is more difficult than the Russian variant, the same lines (55-57 Russian) is easier to understand.
- In the Kyrgyz variant (157-159) is written “таппай жатышат” (they can find the answer) but in Russian it is written (131-132) “they do not know.” They need to use the Kyrgyz words “билбей жатышат.”
- In the Kyrgyz version (line 29) “көбөйүүгө”k. is incorrect.
- “более тёплый климат”r. (a warmer climate) is translated incorrectly into Kyrgyz.
- In the Kyrgyz version (line 132) “калифорниядагы” is written as “калифорнистская” in the Russian version. In Kyrgyz this means “locative case” and in Russian is “possessive” case. Need to use the word “калифорнияалык.”
- There are more words in the Kyrgyz text.
- The word “Global Warming” does not have an equivalent in the Kyrgyz language.
- Text needs to be selected carefully. Take a text from Kyrgyz first (original- not translation). This was clearly written in Russian first – which leads to problems. Perhaps take two different versions from both languages about the same theme.
- The text is in scientific format (on the whole). There are many mistakes in the structure of the sentences primarily due to the literal translation which results in incorrect word combinations. For example: «прогноз кылуусуз» (not making a prognosis) should not be translated literally from Russian into Kyrgyz because the result will be poor. This is due to differences in

syntax between the Russian and Kyrgyz languages. In order for the content to be clear, the translation needs to be accessible. About 70-75 percent is difficult to understand.

- Kyrgyz “thinking” is different that Russian “thinking.”
- There is a mistake in the translation: “Many believe that global warming is completely caused by people…” the translation was: “Көптөрү глобалдык жылуулук көбөүгө толугу менен Адам айыпкер деп ишенишет” – but it should have been “Көптөрү глобалдык жылуулуктун жогорулашына толугу менен Адам айыпкер деп ишенет.

REFERENCES

REFERENCES

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29(1),* 67-91.

Agnoff, W.H., & Cook, L.L. (1988). Equating the scores of the prueba de aptitude academica and the scholastic aptitude test. 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, 36(3),* 185-198.

American Councils for International Education[a] (2004, January). *Analiz Obshe Respublikanskova Testirovaniya 2003 goda v Kirgizskoi Respublikii: Perviyee Vzglad* [Analysis of 2003 National Scholarship Testing in the Kyrgyz Republic: A First Look]. Bishkek, Kyrgyzstan: Drummond, T. & Titov, C.

American Councils for International Education[b] (2004, April). Assessing the impact of the new testing and enrollment system of the kyrgyz republic: A measurement of impact through a survey of key stakeholders. Bishkek, Kyrgyzstan: American Councils.

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

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 R. Hambleton, P, Merenda, P. & C, Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, London: Lawrence Erlbaum Associates.

Bereday, G. (1960). *The changing soviet school*. Cambridge, MA: The Riverside Press.

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,* 259-266.

Boljurova, I. (2003, December 2). *Shag za shagom: Obsherespublikanskoe testirovanie 2003 goda.* [National scholarship testing: Step by step]. *Slovo Kyrgyzstan.* pp. 12-13.

Brunner, J. & Tillet, A. (2007). *Higher education in central asia: The challenges of modernization (case studies from kazakhstan, tajikistan, the kyrgyz republic and Uzbekistan).* Washington, DC: The International Bank for Reconstruction and Development/TheWorld Bank.

Camilli, G. & Shephard, L. (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 (2009). *Rezul'tati obsherespublikanskova testirovaniya i zachisleniya na grantovie mesta vuzov v Kirgizskoi Respubliki v 2009 godu*. [Results of national scholarship testing and enrollment in university grant places in the kyrgyz republic in 2009]. Bishkek: CEATM. www.testing.kg

CEATM[a] (2010). *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[b] (2010). *Natsional'noye otsenivanie obrazovatel'nix dostizhenii uchashixsya.* [National assessment of educational quality]. Bishkek: CEATM. www.testing.kg

Census (2010). National statistical committee of the KR: Population and housing census of the kyrgyz republic of 2009. Bishkek: Kyrgyzstan.

Clark, N. (2005, December). Education reform in the former soviet union. *World Education News and Reviews.* WES. 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),* 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,* 202-214.

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

Coffman, W.E. (1961). Sex differences in response to items in an aptitude test. In E.M. Huddleston (Ed.). *The 18th Yearbook of the National Council on Measurement in Education*. Ames, IA: The Council.

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

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

Conference Program (2006, September 19-20). *Comments from the minister, black sea conference on admissions in higher education: Promoting fairness and equity in access to higher education.* Tbilisi, Georgia.

Davidson, D. E. (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, A., & Santos, C. (2004). Central asian educational issues and problems. In S. Heyneman & A. De Young (Eds.), *The Challenge of Education in Central Asia* (pp. 65-80). Greenwich, CT: Information Age Publishing.

De Young, A., Reeves, M. & Valyaeva, G. (2006). *Surviving the transition? Case studies and schooling in the kyrgyz republic since independence*. Greenwich, Connecticut: Information Age Publishing.

De Young, A. (2007, October). *Paradoxes of higher education in the kyrgyz republic*. Paper presented at the meeting of the Central Eurasian Studies Society Eighth Annual Conference, University of Washington, Seattle, WA.

Dienes, L. (1987). *Soviet asia: economic development and national policy choices*. Boulder and London: Westview Press.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H.Wainer (Eds.), *Differential Item Functioning* (pp. 35-66). 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(4),* 465-484.

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

Drummond, T. (2011). Higher education admissions regimes in kazakhstan and kyrgyzstan: Difference makes a difference. In I. Silova (Ed.), *Globalization on the Margins: Education and Post-socialist Transformations in Central Asia* (pp. 117-144). Charlotte, NC: Information Age Publishing.

Duncan, A. (June 14, 2009). "States Will Lead the Way Towards Reform," Address by the Secretary of Education at the 2009 Governors Education Symposium. http://www.ed.gov/news/speeches/states-will-lead-way-toward-reform

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.

Engelhard, G., Hansche, L., & Rutledge, K.E. (1990). Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education, 3,* 347-360.

Engelhard, G., David, M., & Hanshe, L. (1999). Evaluating the accuracy of judgments obtained from item review committees. *Applied Measurement in Education, 12(2),* 199-210.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments, *International Journal of Testing, 2(3&4),* 199-215.

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),* 301-321.

Ercikan, K. & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing, 5(1),* 23-35.

Faranda, R. & Nolle, D.B. (2010). Boundaries of ethnic identity in central asia: Titular and russian perceptions of ethnic commonalities in kazakhstan and kyrgyzstan. *Ethnic and Racial Studies, 34(4),* 620–642.

Fierman, W. (1991). The soviet transformation of central asia. In W. Fierman (Ed.), *Soviet Central Asia: The Failed Transformation* (pp. 11-35). Boulder, CO: Westview Press.

Fierman, W. (1995). Introduction: The division of linguistic space. *Nationalities Papers, 23(3),* 507–513.

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

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, April). Using statistical and judgmental reviews to identify and interpret translation DIF. Paper presented at the Symposium *Translation DIF: Advances and Applications,* Annual Meeting of the National Council on Measurement in Education (NCME), Montreal, Canada.

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

Glenn, C. (1995). *Educational freedom in eastern europe*. Washington, DC: Cato Institute.

Grenoble, L. (2003). *Language policy in the soviet union.* Boston, MA: Kluwer Academic Press.

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

Grisay, A. & Monseur, C. (2007). Measuring the equivalence of item difficulty in the various versions of an international test. *Studies in Educational Evaluation, 33,* 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),* 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. & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11(3),* 147-157.

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

Helimskaia, R. I. (1994). *Taina chong-tasha*. [The Secret of Chong-Tash]. Bishkek: Ilim.

Herczynski, J. (2003). Key issues of governance and finance of kyrgyz education. *Problems of Economic Transition, 45(10),* 58-103.

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.

Hu, Z. & Imart, G. (1989). *A kirghiz reader*. Bloomington, IN: Research Institute for Inner Asian Studies.

Huskey, E. (1995). The politics of language in kyrgyzstan. *Nationalities Papers, 23(1),* 549- 572.

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),* 329-349.

Johnson, M. (2004). The legacy of russian and soviet education. In S. Heyneman and A. De Young (Eds.), *The Challenge of Education in Central Asia* (pp. 21-36). Greenwich, CT: Information Age Publishing.

Joldersma, K. (2008). *Comparability of multi-lingual assessments: An extension of meta-analytic methodology to instrument validation*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Kirk, R.E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56,* 746-759.

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.

Korth, B. (2004). Education and linguistic division in kyrgyzstan. In S. Heyneman and A. De Young (Eds.), *The Challenge of Education in Central Asia* (pp. 97-112). Greenwich, CT: Information Age Publishing.

Korth, B. (2005). *Language attitudes towards kyrgyz and russian: Discourse, education and policy in post-soviet Kyrgyzstan.* Bern: Peter Lang.

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

Landau , J. & Kellner-Heinkele, B. (2000). *Politics of language in the ex-soviet muslim states.* Ann Arbor, MI: University of Michigan Press.

Mambetaliev, R. (2003, September 11). *Nasha zadacha – dostup i kachestvo obrazovaniya.* [Our task - Access and quality of education]. *Obshestveni Rating*, No. 35 (157).

Mazor, K. (1993). *An investigation of the effects of conditioning on two ability estimates in DIF analyses when the data are two-dimensional*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

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,* 131-144.

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

Mazor, Clauser, Hambleton, (1992). The effect of sample size on the functioning of the mantel-haenszel statistic. *Educational and Psychological Measurement, 52,* 443-451.

McGraw, K.O. & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1(1),* 30-46.

Megoran, N. (2002). *The Borders of Eternal Friendship?: The politics and pain of nationalism and identity along the uzbekistan-kyrgyzstan ferghana valley boundary, 1999-2000.* Unpublished doctoral dissertation, Cambridge University, Cambridge.

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

Milroy, J. (2001). Language ideologies and the consequences of standardization. *Journal of Sociolinguistics,* (5/4), 530-555. Oxford: Blackwell publishers.

National Statistical Committee of the USSR (1989). *Narodnoye obrazovanie i kultura v SSSR.* {Education and Culture in the USSR}. Moscow, USSR.

National Statistics Committee (2000). *Obrazovanie v kyrgyzskoi respublike.* {Education in the Kyrgyz Republic}. Bishkek: Kyrgyzstan.

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,* 315-338.

Narayanan, P. & Swaminathan, H. (1996). Identification of items that show non-uniform DIF. *Applied Psychological Measurement, 20(3),* 257-274.

Oruzbaeva, B. (1997). *Kirgizskii Yazyk: Yazyki Mira, Turkskie Yazyki.* [Kyrgyz Language: Languages of the World, Turkic Languages]. Bishkek: Izdatel'skii Dom.

OSI (Open Society Institute) (2002). *Educational development in kyrgyzstan, tajikistan and uzbekistan: challenges and ways forward*. Retrieved January 21, 2011, from http://www.soros.org/initiatives/esp/articles_publications/publications/development_20020401

Organization for Economic Cooperation and Development (OECD), (2003). *PISA Technical Report 2003*. http://www.pisa.oecd.org/dataoecd/49/60/35188570.pdf. pp. 1-426.

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,* 397- 404.

Podol'skaya, D. (03/03/11). *V Rossii na Zarabotkax naxodyatsya 548 tysyach Kyrgyzstantsev*. [There are 548 thousand Kyrgyzstanis Working in Russia] – ИА «24.kg». http://mirror24.24.kg/parlament/94490-v-rossii-na-zarabotkax-naxodyatsya-548-tysyach.html

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* (pp. 237-258). New York: Plenum Press.

Poortinga, Y.H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24,* 737-756.

Presidential Decree No. 91. (2002, April 18). *"O dal'neyshih merax po obespecheniyu kachestva obrazovaniya i sovershenstvovaniyu upravleniya obrazovatel'nymi protsessami v Kyrgyzskoy 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. *International Journal of Continuing Engineering Education and Lifelong Learning, 12(1-4),* 167-180.

Reeves, M. (2005). Of credits, *kontrakty*, and critical thinking: encountering 'market reforms' in kyrgyzstani higher education. *European Educational Research Journal, 4(1).* pp. 5-21.

RIA News, Moscow. (2007, February 5). *Vvedenie v rossii edinogo gosudarstvennogo ekzamena (EGE) yavlyaetsya oshibkoy, ubezhden spiker sovyeta federatsii sergey mironov.* [Speaker of federal Soviet thinks the Unified State Exam (USE) is a 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),* 1-20.

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

Rogers, H.J. (1989). *A logistic regression procedure for detecting item bias*. Unpublished Doctoral Dissertation, University of Massachusetts, Amherst.

Rogers, J. & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement, 17(2),* 105-116.

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

Sait Halma, T. (1981). *201 Turkish verbs fully conjugated in all the tenses*. New York: Barron's Educational Series.

Scheuneman, 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.

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, 25(1),* 1-13.

Schmidt, A.P., & Belistein, C.A. (1987). Factors affecting differential item functioning of black examinees on scholastic aptitude test analogy items (Research Report 87-23). Princeton, NJ: Educational Testing Service.

Shamatov, D. & Niyozov, S. (2010). Teachers surviving to teach: Implications for post-soviet education and society in tajikistan and kyrgyzstan. In J. Zajda (Ed.), *Globalization, Ideology and Education Policy Reforms: Globalization, Comparative Education and Policy Research*. Springer Science + Business Media B.V.

Shamatov, D. (in press). Everyday realities of a young teacher in post-Soviet Kyrgyzstan: A case of a history teacher from a rural school. In P. Akcali, & C.E. Demir (Eds.), *Post-Soviet Kyrgyzstan: Political and Social Challenges*. Routledge.

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.

Silova, I. (2009). The crisis of the post-soviet teaching profession in the caucuses and central asia. *Research in Comparative and International Education, 4(4),* www.wwwords.co.uk/RCIE.

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

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

Solano-Flores, G. (2006). Measurement error in the testing of english language learners, Teachers College Record, Columbia University, 108(11), pp. 2354-2379.

Soktoev, I.A. & Usubaliev, E. T. (1982). *Vyshee shkola sovetskaya kirgizstana. {higher education in soviet Kyrgyzstan}.*Frunze: Kyrgyzstan.

Steiner-Khamsi, G., Teleshaliyev, N., Sheripkanova-MacCleod, G., & Moldokmatova, A. (2011). Ten-plus-one ways of coping with teacher shortage in kyrgyzstan, In I. Silova

(Ed.), *Globalization on the Margins: Education and Postsocialist Transformations in Central Asia* (pp. 203-232). Charlotte, NC: Information Age Publishing.

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,* 301-319.

Swaminathan, H. & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27(4),* 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.

Toursunov, H. (May, 11, 2010). "Jumping Off a Sinking Ship." in *Transitions On-Line*, http://www.tol.org/client/article/21435-jumping-off-a-sinking-ship.html.

USSR Government Committee for Statistics (1989). *Narodnoye Obrazovanie i Kultura v SSSR: Statisticheskii Sbornik.* {Education and Culture in the USSR: Statistical Collection}. Moscow: Finance and Statistics.

Valkova, I. (2001). My symphony: Interview with the minister of education of kyrgyz republic, Camilla Sharshekeeva. *Thinking Classroom, 6.* Vilnius, Lithuania: International Reading Association.

Valkova, I. (2004). *Getting ready for the national scholarship test: Study guide for abiturients.* 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. & Tanzer, N.K., (1999). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 263-279.

Van de Vijver, F. & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting Educational and*

Psychological Tests for Cross-Cultural Assessment. London: Lawrence Erlbaum Associates.

Wright, S. (1999). Kyrgyzstan: The political and linguistic context. *Current Issues in Language & Society, 6(1),* 85-91.

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*. University of Alberta: Centre for Research in Applied Measurement and Evaluation.

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.