2009

The Kingsbury Center at
Northwest Evaluation
Association

Lingling Ma, Ph.D and John Cronin, Ph.D.

[EVALUATING THE EFFECT OF RANDOM SELECTION ON VIRTUAL COMPARISON GROUP CREATION]

This study evaluated the procedures for selecting students to be used as Virtual Comparisons for a study group of 1,000 students in an effort to determine whether the random selection procedures created differences between the VCG sample and the population from which it was drawn.

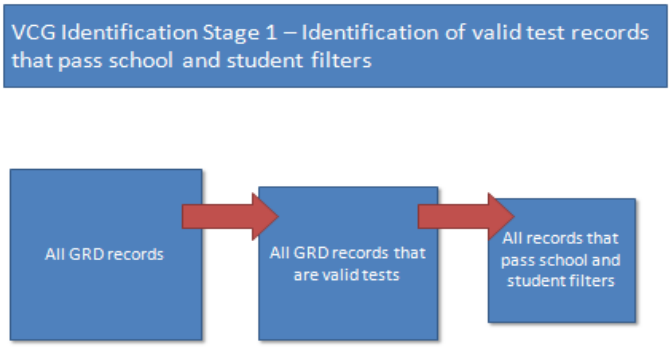
INTRODUCTION

Virtual Comparison Groups (VCG) were developed by the Northwest Evaluation Association as an alternative to conventional controlled experiments for social science researchers working in the field of education. The VCG is generally a group of up to 51 students who are matched, based on key characteristics of the student and school, to a single student who is part of a study group. Because each student has his or her own VCG, a sample of 1,000 students would be matched to 1,000 VCGs of up to 51 students each. In these studies, the growth of study group students is measured using NWEA's Measures of Academic Progress on a cross-grade, equal interval scale developed using a model grounded in item-response theory (IRT) and compared to the growth of the VCG on the same instrument. Ideally this allows us to compare the academic growth of a study group of students in relation to a very large group of comparison students sharing many like characteristics.

Conditions in schools often make it difficult for researchers to conduct large scale, well-controlled experiments involving interventions with students. VCGs are a possible alternative in some circumstances to controlled experiments. By using existing student achievement data and introducing controls that produce a group of students that match the study group population on key characteristics, a well-designed VCG strives to produce the same conceptual comparison as a controlled experiment in which one-group receives a treatment is compared to a control group that shares the same characteristics (because of random assignment) absent access to the treatment.

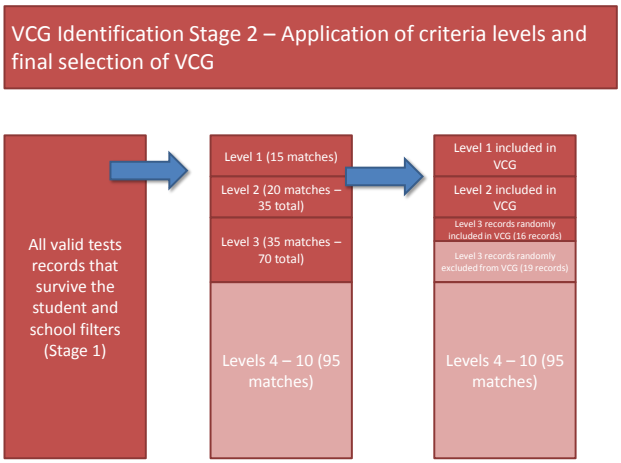
The VCG creation process involves several stages. The first stage involves selecting a pool of potential VCG matches by narrowing the records in the Growth Research Data to those that include valid tests matching the most broadly defined school and student matching criteria (Figure 1).

Figure 1- VCG Identification Stage One.



From this sample, a series of criteria are applied to create the pool from which the VCG will be ultimately collected. The first level criteria are very strict (RIT scores within 1 point of the study group student, school with poverty rate within 5 points, and test dates matching the dates between the study group student's within +/- seven calendar days). If the first level criteria does not produce at least 51 matches, then the criteria are loosened to the point at which a group of at least 51 students is produced (Figure 2). At this point the records are sorted, first in ascending order based on the criteria level applied to acquire the record, and second on a random number that is used to select the records used from the lowest criteria level applied. Thus in the Figure 2 example, all students who were found by applying the level one and two criteria would be included in the VCG. The level three criteria produced matches beyond the 51 needed, so the 35 records acquired from applying level three were sorted by random number, and 16 of these were selected to fill out the VCG.

Figure 2 – VCG Identification Stage 2



RESEARCH QUESTION

One of the assumptions of the Virtual Comparison Group is that the application of criteria and sampling procedures do not, by themselves, produce biases in sampling that compromise the comparability between study group students and their VCG. These biases are most likely to be apparent if differences in growth rates are found between study group students and their VCG that are attributable to the selection methods rather than the treatment.

One risk is that the methodology for randomizing the sample once the final eligible pool of students is acquired may introduce some bias. Figure 2, for example, depicts 35 students who were eligible for inclusion in the VCG after level three criteria are applied. These 35 students were ordered by random number to create the final study group. If the procedure for randomizing introduces its own bias, this would also be problematic. So the research question in this study is whether the process used to randomize records for inclusion in a VCG introduces bias into the results.

If the answer to the question is *no*, then the next logical step would be to investigate the criteria levels and whether the application of these levels may also introduce bias into the creation of VCG samples. For example, when insufficient matches are created from the initial application of the strictest criteria, the requirements for a

match are gradually loosened until sufficient matches are available (or the process is exhausted). The default procedures for this loosen the match on initial test score prior to loosening the criteria for matching instructional time and prior to the criteria for loosening the criteria related to the school's poverty rate (again see Appendix 1 for details on this process). For a student performing at the high end of the performance continuum, loosening the initial score criteria from +/- 1 points to +/- 2 points might cause the inclusion of more students with scores slightly below the student than those above. If this happened the VCG would actually start with slightly worse performance than the study group student, and this introduces a possible bias in results that may be exhibited in differences in growth that are attributable to the difference in starting score. This issue is the subject of research by NWEA but not inside the scope of the current study.

SUMMARY OF THE METHODOLOGY

A group of 1000 fifth grade mathematics students were selected from the GRD as subjects for this analysis. Students performing in the top 5% or the bottom 5% of the fifth grade GRD population were excluded because these students may not have generated adequate matches to test the randomization procedure. The group was further narrowed by applying the *default criteria* for inclusion (see Appendix 1 for a description of these criteria) and the Level 1 matching criteria (also explained in Appendix 1). These are the criteria that would produce the closest match between the characteristics of interest in the study student and his or her VCG. Students who did not have at least 100 matches after application of these filters were excluded as potential subjects. From the remaining students, 250 students were randomly selected from each quartile of the overall population's fall performance. This group of 1000 students became the subjects for the study.

To test the process for random selection, we created a control group and a series of study groups using these subjects. The control group, which will be referred to as the *VCG population group*, contained all students who were available for selection after the application of the application of both the stage 1 and stage 2 criteria. From this population group a random sample of 100 students were drawn. This group became the base for creation of ten study groups, with minimum sizes that ranged from ten to one-hundred in increments of ten. These will be called the *VCG study* groups. The result was a group of 1,000 students each of whom had one *VCG population group* and ten *VCG study groups* of various sizes. The aggregated results of all *VCG population groups* were compared to the aggregated results of the ten *VCG study groups* to get our results.

The VCG population groups varied in size from the minimum 100 students to 986 students, with a mean size of 325 students and a standard deviation of 196. 65% of the cases had more than 200 students, which would be an adequate number of cases to show a substantive probability that a biased selection procedure could cause differences between the population result and a sample result. In addition, the process of randomly selecting the smaller groups (of 10 to 90) within the 100 students also would create an opportunity for differences if bias were introduced in the random selection process.

The study's hypothesis is that, if the randomization procedure used to create the group of 100 students does not create bias, differences in academic growth between the *VCG population group* and the various *VCG study groups* would not be significant. For purposes of this study, academic growth is represented by the *VCG Index* metric. The VCG index is the difference in growth between the student and the average growth of the Virtual Comparison Group.

FINDINGS

Table 2 shows the results of the analysis of the differences between the average of the 1,000 *VCG population groups* and the various *VCG study groups* for student sample. Differences in VCG Index scores between the mean VCG population group index score and the study groups ranged from +0.02 to – 0.02 RIT, meaning that, over 1,000 students, the average difference in VCG index scores was extremely small. The best test for whether randomization biased the sample is the comparison between the VCG population group and the 100 student study sample, since this was student sample that was used to derive the others. For this group the differences in group means was negligible (< .01 RIT) and the standard deviation for individual samples was quite narrow (< 0.6 RIT). The differences in the population and study group means were not statistically significant (t=-.33; p=.74). This result would suggest that the procedure for selecting the study group sample of 100 students did not introduce a bias that caused it to differ significantly from its original population sample. In addition, the narrow standard deviation suggests that differences between each individual study group member and their respective controls were also generally quite small, under .6 RIT for approximately 2/3 of the cases.

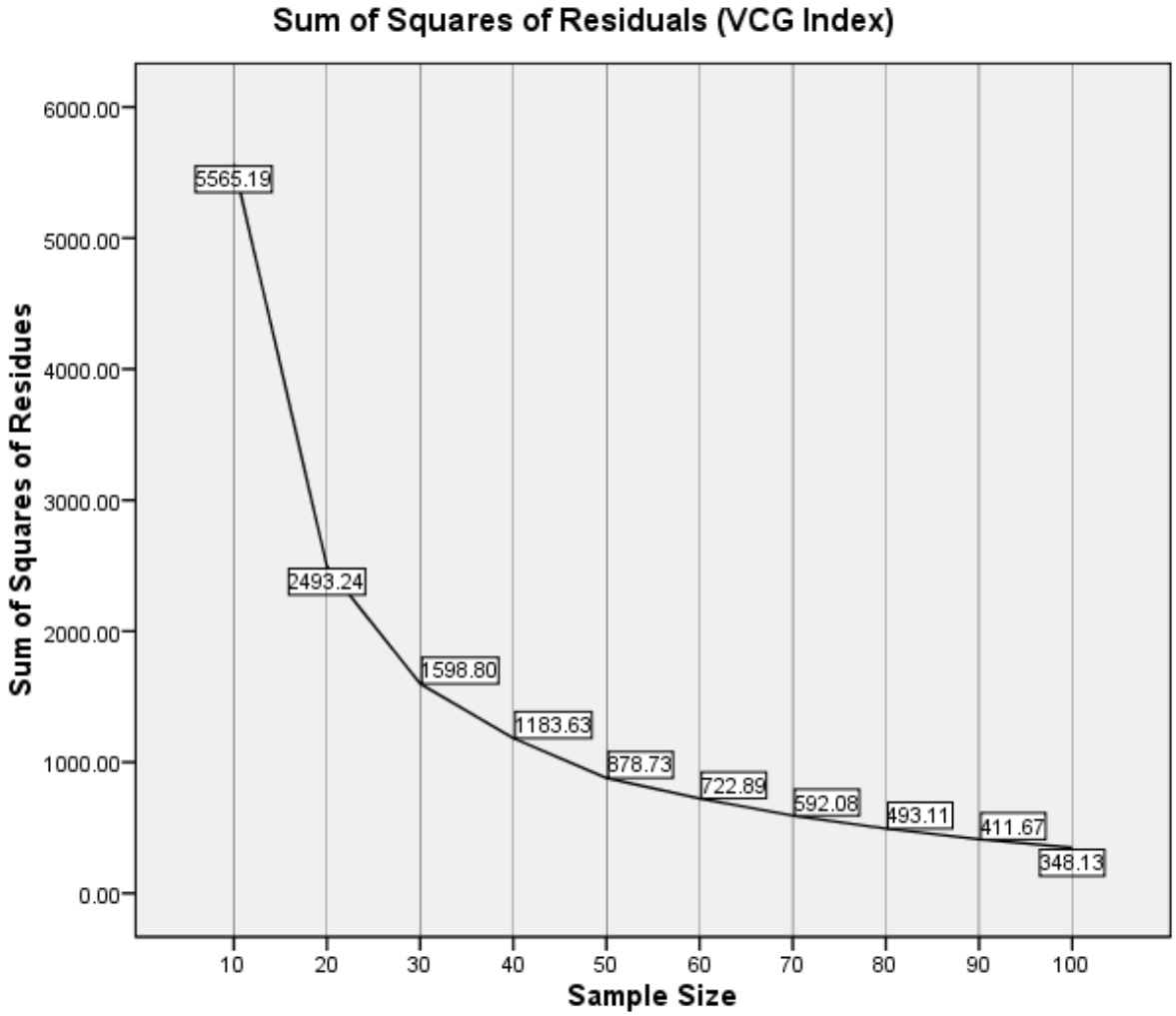
In terms of the study groups randomly derived from the sample of 100, the value of *t* for their differences ranged from +.34 to -.88. None of the differences approached statistical significance at a .05 level. Thus when dealing with sample populations of a moderate size, we found no evidence that the process for randomizing selection of the final sample introduced bias that would inflate or deflate the growth estimate of a VCG group.

Table 1 – Average VCG Index Differences between VCG Population Group and VCG Study Groups for a 1,000 Student Population in Mathematics

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	VCGIndex10 - VCGIndexAll	-.03328	2.36001	.07463	-.17973	.11317	-.446	999	.656
Pair 2	VCGIndex20 - VCGIndexAll	.01697	1.57970	.04995	-.08106	.11500	.340	999	.734
Pair 3	VCGIndex30 - VCGIndexAll	-.01901	1.26493	.04000	-.09751	.05948	-.475	999	.635
Pair 4	VCGIndex40 - VCGIndexAll	-.00721	1.08847	.03442	-.07475	.06034	-.209	999	.834
Pair 5	VCGIndex50 - VCGIndexAll	-.01840	.93770	.02965	-.07659	.03979	-.621	999	.535
Pair 6	VCGIndex60 - VCGIndexAll	-.01721	.85048	.02689	-.06999	.03556	-.640	999	.522
Pair 7	VCGIndex70 - VCGIndexAll	-.00758	.76982	.02434	-.05535	.04019	-.311	999	.756
Pair 8	VCGIndex80 - VCGIndexAll	-.01966	.70230	.02221	-.06324	.02392	-.885	999	.376
Pair 9	VCGIndex90 - VCGIndexAll	-.01057	.64185	.02030	-.05040	.02926	-.521	999	.603
Pair 10	**VCGIndex100 - VCGIndexAll**	**-.00624**	**.59029**	**.01867**	**-.04287**	**.03039**	**-.334**	**999**	**.738**

Changing the size of the VCG did have a significant impact on the variance associated with the VCG study group samples. Figure 3 shows the sum of squares residuals for the various VCG study groups. The results show that increasing the size of the Virtual Comparison Group does yield a meaningful reduction in the amount of variance. For example, variance is reduced by about 85% by raising the size of a VCG from 10 to 50 students. These benefits, as expected, marginally decline as the size of the VCG increases further. Raising the VCG sample size from 50 to 100 students only reduces the original variance by another 9%.

Figure 3 – Sum of Squares of Residuals for the VCG Index



DISCUSSION

The process used to create VCGs is quite elaborate and the randomized selection of a final group of students is but one small part of this process. The evidence from this study would indicate that the part of the process involving random selection of records does not introduce a bias that would cause VCG samples to differ from their respective population groups.

The evidence from the study does introduce a dilemma to be considered as we refine future studies of VCG methodology. Results from the study indicate that larger sample sizes do provide more consistent estimates of the VCG population's actual performance (see Figures 4, 5, and 6). For example, when the sample size is 100 students, approximately 90% of the estimates fall within +/- 1 RIT of the population group estimate (Figure 6). Thus for individual students the 100 student sample size gives us a highly consistent VCG result. As the sample size declines, the variance among the differences in the individual estimates increases greatly. At a sample size of 20 (Figure 4) a large number of individual students vary from their population VCG index by more than +/1 RIT and differences beyond +/- 2 are not uncommon. As a result, smaller samples introduce a greater risk that any single individual's VCG group may vary substantively from his or her population because of these smaller sample sizes. Although these individual differences balance and ultimately wash out when they are observed over enough students, thus preserving the integrity of the group estimates, they are a concern if we want each student to truly be compared to a student that fairly represents the population of students sharing his or her characteristics.

Figure 4 – Difference in mean VCG index score between each population group and study group with a sample size of 20.

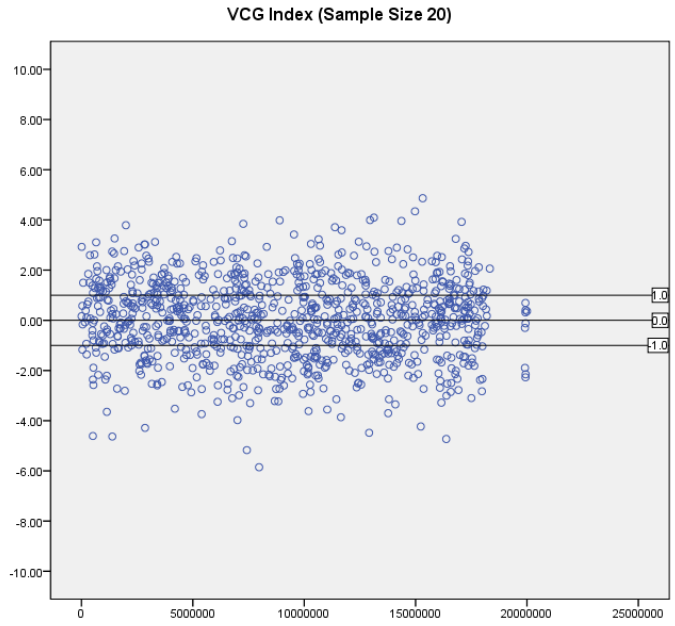


Figure 5 – Difference in mean VCG index score between each population group and study group with a sample size of 50.

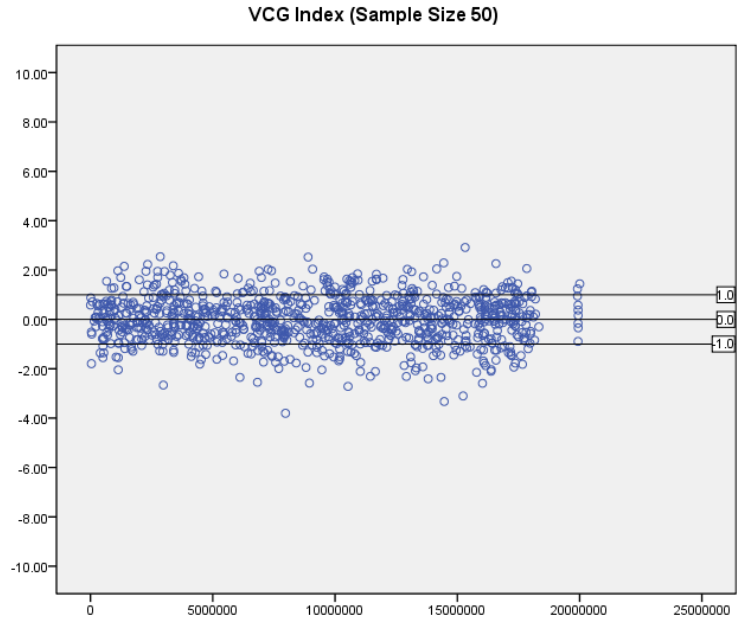
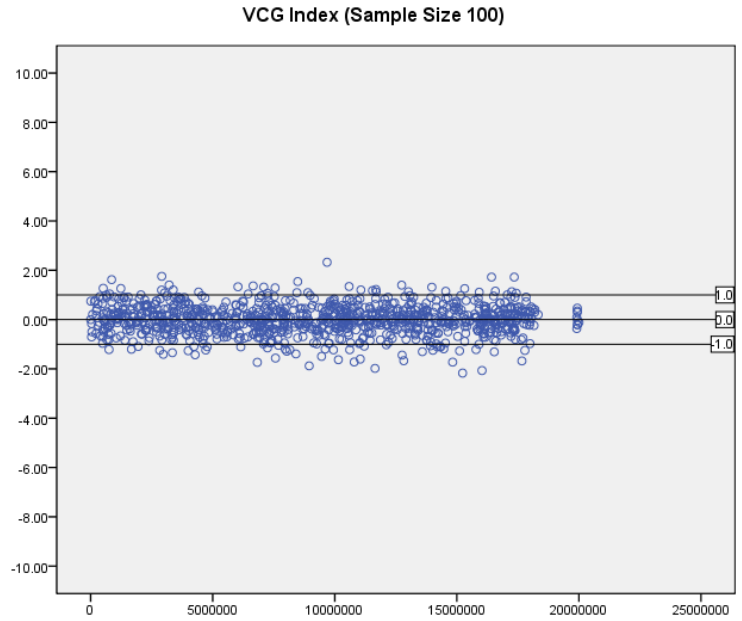


Figure 6 – Difference in mean VCG index score between each population group and study group with a sample size of 100.



The other side of the dilemma has to do with the number of available matches. While larger VCG groups clearly produce more stable comparisons for individual students, some students have characteristics that do not generate

many matches when the tightest criteria are applied. For example, a very high performing student, say a fifth grader with a math RIT of 260 (above the 99th percentile), who also attends a high poverty school, may not generate 51 matches for a VCG, particularly if the student takes the fall and spring assessments at unusual times (say November for fall and June for spring). The current VCG creation process loosens the criteria to the level that creates an adequate number of matches for the student. Thus to gain some stability in the estimate, we accept a poorer fit between the student and the characteristics of the VCG. The way in which criteria are loosened may have a large effect on the final composition of the VCG. For example, if we loosen the required RIT match for our very high performing student, there are more available students under the student's RIT score than above it. If the selection procedure chooses more students with lower RIT scores to fill out the sample, that may introduce a bias in the comparison. If instead we chose to loosen the matching number of instructional days, a student who had a large interval between tests (say over 175 days) might be more likely to be matched to students with shorter intervals than longer ones. This may also introduce a bias to the comparison.

This study gives us some confidence that the process of random selection itself does not bias the estimates of a VCG's performance in a way that causes them to differ from the original matching population from which they were drawn. However, the application of the default criteria used to select a VCG population and the matching procedures used to create the final group may still create a bias, either by the nature of the criteria applied or by their order of application. Thus the next stage of the proposed research would be to pursue studies to evaluate the whether the criteria used to create Virtual Comparison Groups introduce differences between students and their matching population that could be characterized as products of bias. One issue these studies might examine is whether the order in which criteria are applied produce differences that might produce bias. Does, for example, filtering the eligible population according to school poverty rate prior to filtering the population by instructional time, cause VCG results to differ from those obtained if the population was first filtered by instructional time than by school poverty rate? Another question might be to determine the effect that the default criteria have on the character of VCGs and determine whether matching on these characteristics contributes to the robustness of the comparison. For example, requiring VCG students to come from a school with the same urban/rural characteristics as the study group student provides something of a geographic match, but excludes large numbers of potential matches from the available VCG pool, which may compromise the quality of the final VCG selected.

In summary, this study finds that the procedures for randomly selecting cases for membership in a VCG do not create samples that differ significantly from the population group from which they are drawn. The study also shows that while VCG samples above 40 do not necessarily improve the stability of VCG estimates across a large population of students, they greatly improve the stability of individual VCG estimates, thus making the individual comparisons more valuable to teachers. The study informs the need for future research into how the VCG criteria themselves and the order in which they are applied are likely to impact the nature and performance of a Virtual Comparison Group and its validity as a means for comparing student performance to a control in a quasi-experimental setting.

APPENDIX 1 – VIRTUAL COMPARISON GROUP SELECTION PROCEDURE

Virtual Comparison Groups (VCGs) are used to study differences in growth between a population of interest (perhaps a group of teachers implementing a new mathematics program) and a control sample that matches the population of interest on some preidentified characteristics. In general, a separate VCG or control sample of up to 51 students is created for each student in the population of interest. NWEA has developed a set of default criteria that are applied to create the majority of VCGs.

INITIAL POPULATION

Virtual Comparison Groups are drawn from NWEA's *Growth Research Database* a longitudinal depository of student achievement and growth information derived from the organization's Measures of Academic Progress (MAP) and Achievement Level Tests. Approximately two-thirds of NWEA's partners choose to include their information in the Growth Research Database and the GRD now contains over 80,000,000 test results in mathematics, reading, language usage, and science. These results constitute the initial set of eligible records from which a VCG is formed.

TEST FILTERS

Once a student from the population of interest has been identified (hereafter called a study group student) a process to filter the GRD to select the VCG begins. The first set of filters remove test results that do not pass NWEA two test validity filters. Tests with a duration under six minutes are excluded from consideration for a GRD. Tests with a standard error less than 1.5 or greater than 5.5 scale score points (called RIT points) are excluded, unless the student scale score is above 240. These same criteria are applied to the general reporting of test events, thus test results excluded from consideration are also generally not included in official reports of results to schools.

SCHOOL FILTERS

Schools that do not have data inside the National Center for Educational Statistics Common Core of Student Data (CCD) for their student results are excluded from consideration. Schools excluded would normally be limited to new schools who have not been assigned an ID number from the National Center for Educational Statistics or schools that have been changed or reconstituted, for example, a middle school reconstituted as a charter school would not be included if the charter school does not have an NCES ID.

If the Common Core of Data reports a valid percent of students using a free and reduced lunch program, then students from schools whose percentage is within +/- fifteen percent of the free and reduced lunch percentage in the study student's school are included. If this data is not available, then the percentage of children below the poverty line for the school system as reported in the *Small Area Income & Poverty Estimates* of the United States Census is used for the comparison. In this case, records are included if the school system's reported poverty rate (individual school data is not available from the census) is within +/- five to fifteen percent of the rate reported for the study group student's school.

Records that pass this poverty filter must also match on the CCD's school type variable (regular, special education, vocational, other/alternative) and must also match on a recoded variation of the CCD's locale designation. For matching purposes, NWEA consolidates the CCD's locale definitions into two classifications "urban" and "rural". Records from schools whose school type and recoded locale definition match the study group student's school are included.

STUDENT FILTERS

Students remaining in the data pool after application of the school level filters remain as potential candidates for the student's VCG. The next series of filters matches key characteristics of the study group student to this pool. These minimum requires applied by these filters are as follows:

- Students must be enrolled in the same grade as the study group student. Study group students in grades ten, eleven, or twelve need not be matched to students in their enrolled grade, but can be matched to students enrolled in any of those three grades.
- Students must have an initial RIT score within +/- five RITs of the study group student.
- The number of calendar days between the pre and post test must be within +/- eighteen days of the study group student.
- The student must have tested within the same school year or proceeding school year as the study group students.
- Students in the study group cannot serve as VCG members for other students in the study group.

OPTIONAL STUDENT FILTERS

The default VCG criteria do not limit matches by the gender, reporting ethnicity, or program status of the study group students. When these criteria are applied, records are excluded from the study pool if they do not match the study group student's characteristic on the selected characteristic(s).

APPLICATION OF MATCHING CRITERIA

The application of the school and student filters results in a pool of student records that are used to create the final Virtual Comparison Group for the student. The goal of this process is to create a group of 51 students for each study group student who match the student's characteristics of interest as closely as possible. Thus up to ten levels of criteria are applied. The strictest criteria, those that would produce the tightest matches to the study group student are applied first. If these do not produce 51 or more matches, the criteria are loosened in up to ten stages. All records that were returned as valid matches at each prior criteria level are carried over to the next level. Once a criteria level is reached that produces a total of 51 or more matches, that group of students become the sample from which the final VCG is drawn. If the application of all ten levels results in a total of fewer than 51 matches, the available student matches become the student's VCG. If the final number is less than 20, the data is flagged as *low count*, so users will know that the VCG does not produce a sufficient number of students who closely match the study group student's characteristics.

Table 1 shows the various criteria levels and their order of application.

Level	RIT Difference on initial test	School's SES Rating	Number of calendar days between pre and post test	School year
1	+/- 1	+/- 5%	+/- 7	Records included from the current school year only
2				Records included from the prior school year
3	+/-2			
4			+/- 10	
5		+/- 10		
6		+/-15		
7	+/- 3			
8			+/- 18	
9	+/- 4			
10	+/- 5			

(blank cells inherit the values from the cell above them)

RANDOM SELECTION

The application of criteria levels ends at any point in which a cumulative total of 51 or more students are available for the Virtual Comparison Group. At this point, a random value is assigned to all potential VCG matches. Students are sorted in ascending order, first on their criteria level, then on the random value. The first 51 students are then selected as the final VCG.