AFRL-RH-WP-TP-2010-0001

Mixed- & Homogeneous-Culture Military Team Performance on a Simulated Mission: Effects of Age, Computer-Game Experience & English Proficiency

Rik Warren

Warfighter Interface Division
Cognitive Systems Branch

February 2008

Interim Report

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7022

NOTICE

AFRL-RH-WP-TP-2010-0001

//SIGNED//
RIK WARREN
Program Manager
Cognitive Systems Branch

//SIGNED//
DANIEL G. GODDARD
Chief, Warfighter Interface Division
Human Effectiveness Directorate
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
February 2008	Interim	

4. TITLE AND SUBTITLE

Mixed- & Homogeneous-Culture Military Team Performance on a Simulated Mission: Effects of Age, Computer-Game Experience & English Proficiency

5a. CONTRACT NUMBER
In-House

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62202F

6. AUTHOR(S)

Rik Warren

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
71841009

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7022

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHCS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RH-WP-TP-2010-0001

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
NATO-RTO-MP-HFM-142 Symposium "Adaptability in Coalition Teamwork," Copenhagen, Denmark, April 21-23, 2008. 88th ABW/PA cleared on 09 April 2008, WPAFB-08-0723.

14. ABSTRACT

In order to investigate the performance of mixed- versus homogeneous-culture four-person military teams, the NATO Human Factors and Medicine Panel (HFM-138) on "Adaptability in Multinational Coalitions" conducted a computer-based experiment. This paper examines the role of age, computer-game experience, and English proficiency as confounding variables in explaining the results. A key finding is that differences among national groups disappear when the effects of the confounds are removed, but the mixed-culture teams now have the best performance. Some reasons for these findings and the implications for military selection, training, and procedures are discussed.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Rik Warren
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	28	19b. TELEPHONE NUMBER *(include area code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

Mixed- & Homogeneous-Culture Military Team Performance on a Simulated Mission: Effects of Age, Computer-Game Experience & English Proficiency

Rik Warren
U.S. Air Force Research Laboratory
Human Effectiveness Directorate
Wright-Patterson AFB Ohio 45433-7022 U.S.A.

Richard.Warren@wpafb.af.mil

ABSTRACT

In order to investigate the performance of mixed- versus homogeneous-culture military teams, the NATO RTO Research Task Group (HFM-138/RTG) on "Adaptability in Multinational Coalitions" conducted a computer-game experiment involving a modern urban search-for-contraband. Using the Situation Authorable Behavior Research Environment (SABRE), the study used a scenario which required planning, resource allocation, situation awareness, communication, and coordination for good performance. Good performance also required maintaining the good-will of the local populace who could provide useful tips or, the opposite, misinformation to the searchers. Fifty-six 4-person teams of NATO officers each from five nations received training on the game-play, planned for, and conducted their mission. The main hypothesis was that homogeneous-culture teams would perform better than mixed-culture teams. Contrary to expectations, performance was not a simple function of cultural composition. This paper examines the role of age, computer-game experience, and English proficiency as confounding variables in explaining the results. A key finding is that differences among national groups disappear when the effects of the confounds are removed, but the mixed-culture teams now have the best performance. Some reasons for these findings and the implications for military selection, training, and procedures are discussed.

1.0 INTRODUCTION

In order to investigate the performance of mixed- versus homogeneous-culture four-person military teams, the NATO RTO Human Factors and Medicine Panel Research Task Group (HFM-138/RTG) on "Adapatability in Multinational C oalitions" conducted a computer g ame-based experiment (NATO RTO HFM-138/RTG, 2008). U sing t he S ituation A uthorable Behavior Research E nvironment (SABRE) (Warren e t a l., 200 4; Leung, Diller, & Ferguson, 2005), BBN Technologies Inc. developed a modern urban search-for-contraband scenario sp ecifically t ailored f or t his N ATO ex periment (Warren et a l., 2 005) w hich r equired p lanning, resource a llocation, s ituation a wareness, c ommunication, a nd c oordination f or g ood pe rformance. G ood performance also required maintaining the good-will of the local populace who could provide useful tips or, the opposite, misinformation to the searchers.

The principal hypothesis was: Homogeneous-culture teams (i.e., teams whose members are all from the same nation) perform better than mixed culture teams (i.e., teams whose members are from different nations).

Contrary to expectations, performance, as indexed by several different metrics, was not a simple function of culture c omposition. Most s urprizingly, hom ogeneous-culture teams w ere n ot g enerally b etter than m ixed-culture teams. These results are well-illustrated in Figure 1 which shows the relative performance of all 56

teams, grouped by national or mixed-culture composition, on the main performance metric (to be discussed further below).

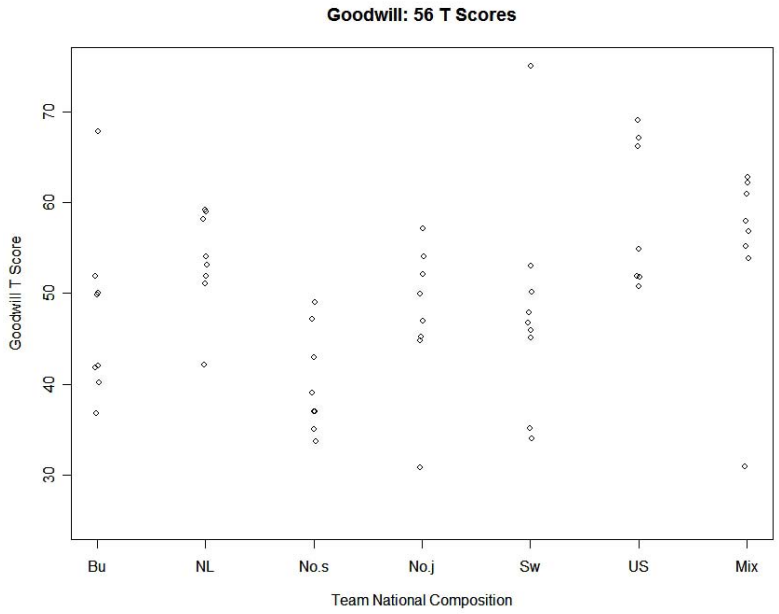

Figure 1: Team "goodwill" performance T-scores (Mean = 50; SD = 10) for each of the 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age

Several non-cultural factors might have contributed to the pattern of results:

- Within teams, participants were of similar ranks/grades and thus similar in age. But between teams, ranks/grades and thus, age and experience, could differ. Relative seniority can be an advantage in a complex task requiring planning. But, relative juniority can be associated with computer-game experience and thus be an advantage.
- The task required playing a complex computer game using many different procedures for communication, movement, and sundry actions. In spite of a two-hour training session, there might be some effect of computer-game experience in achieving a level of mastery permitting participants to concentrate on the task at hand rather than game-play technicalities.
- The game-play was all in English (using keyboard-only communication, so this was monitored and ensured). Hence, in a multi-national population, proficiency in English could affect performance.

We (NATO RTO HFM-138/RTG) anticipated that the two factors of computer game-play experience and English proficiency, in particular, might act as moderator, mediating, or confounding variables and, hence, we collected several relevant questions about each in a pre-game questionnaire. Age and rank data was also collected. As discussed later, it was not possible to select participants with either matching levels or controlled variation in these three factors.

Thus, the purpose of this paper is to explore these possible non-cultural alternative explanations for our pattern of results and to partial-out their effects, if any, using linear regression techniques. (Analysis of

Covariance [ANCOVA] is an alternative approach and is treated in the Discussion section.) Another purpose is to discuss possible non-trivial implications for coalition military team selection, training, and procedures.

2.0 METHOD, ABRIDGED

Before turning to the analysis, I briefly review some details of the experiment. A full description is in NATO RTO HFM-138/RTG (2008).

2.1 Participants & Teams

All 224 participants were volunteers and officers from five NATO nations: Bulgaria, The Netherlands, Norway, Sweden, and the United States. In total, there were 56 teams of 4 persons each: 8 from Bulgaria, 8 from The Netherlands, 16 from Norway, 9 from Sweden, and 7 from the United States. Eight of the Norwegian teams consisted of junior officers or cadets; the 8 other teams were more senior. Hence, some analyses below treat these as two separate "culture" groups: No.j and No.s for "junior" and "senior." Eight additional 4-person teams, the mixed-culture teams, were formed having a composition of one person each from different nations.

Within each team, officers had to be no more than one rank/grade apart, but there was no required specific rank for all teams. No age requirements were set although the imposed similarity of ranks acted to keep ages within a team somewhat similar. Details of the age distributions appear in the age analysis section below.

No requirements were set for computer-game experience nor was game-experience controlled for in the study. However, due to the obvious possible effect on the results, several questions about gaming experience were asked in pre-game-play questionnaires. Details of the gaming-experience distributions appear in the gaming-experience analysis section below.

All had to speak and write English, but no specific proficiency criterion beyond NATO minimums was set. Several questions relating to English proficiency were asked in a pre-game questionnaire. Details of the proficiency distributions appear in the English proficiency analysis section below.

The result of these selection constraints and procedures is that age, English proficiency, and computer-game experience were not independent of each other or national composition. *A few demographic values were missing. Estimated values were included in the current analyses.* Figure 2 is a bubble chart of the three demographic factors with the national composition of each team indicated. Distinct non-balanced non-factorially-crossed patterns can be seen: For example, all seven American teams form a cluster located at the high end of English proficiency and at the middle of the age scale. The bubbles indicate that the Americans also have relatively high levels of computer-game experience. The Dutch teams form another cluster located at the younger end of the age scale and also show high levels of computer-game experience. The senior Norwegian teams, in contrast, form a cluster at the upper end of the age scale and show low levels of computer-game experience.

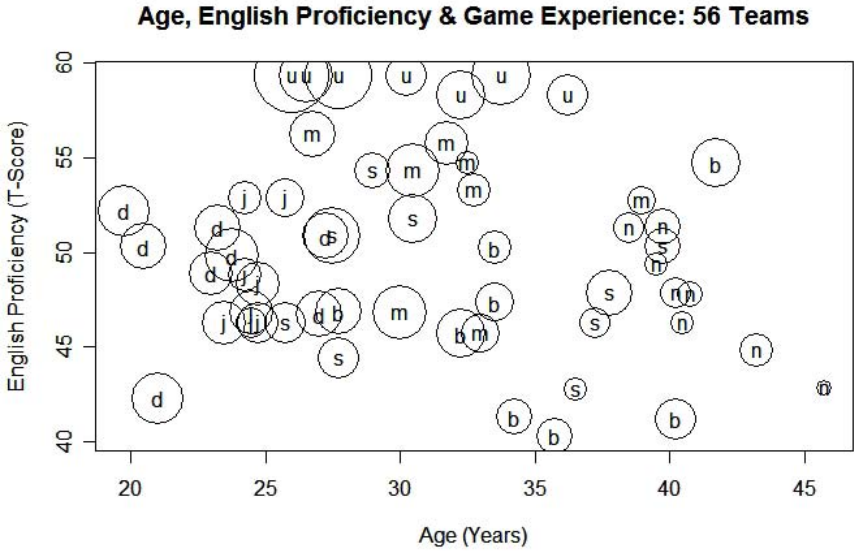

Figure 2: Demographic profiles of the 56 teams. Game experience is proportional to bubble size. Letters indicate national composition of the teams: Bulgaria (b), The Netherlands (d), Norway-senior age (n), Norway-junior age (j), Sweden (s), United States (u), mixed culture (m).

2.2 The Computer Game & Scenario

Details of the computer game and scenario are in NATO RTO HFM-138/138 (2008), Warren et al. (2004), and Warren et al. (2005). Essentially, teams were to find contraband caches hidden in a modern urban environment. The four human players are represented by "avatars" in the game-space. As they explore the cityscape, they meet some of the local populace (played by non-human "non-player characters" or NPC's). Some of the local populace provide "tips" about contraband or suspicious activity. Some of the local populace are truthful, some are not. Teams gain points by finding weapons caches and performing goodwill side-quests for the local populace. Teams lose points for opening empty suspected locations and angering the local populace by how they interact with them.

2.3 Procedure

Each team member was seated at a computer terminal. Same-nation teams were in the same room in their home nation but were visually and auditorily shielded from their other team members. Mixed-nation team members were always in their home nation and played the game over the Internet.

The game is a complex but very absorbing and immersive. Team-members received two-hours of training and learned how to communicate with each other using their computer keyboards. Keyboards and the computer screens were the only means of communication and information sharing. This forced all communication to be in English. It also means that every keystroke was recorded and available for future analysis.

Game-play involved planning, resource allocation, situation awareness, communication, and coordination. Game-play was monitored by a server-computer and almost all activity was recorded. In addition to the game-play, questionnaires were filled-out using the computer. During the game-play, there were probes from a "superior officer" to determine situation awareness at three different times.

2.4 Design & Performance Metrics

The primary independent variable was the homogeneous- versus mixed-culture composition of the 56 teams.

Answers to the pre-game questionnaire were post-game processed to form metrics for game-play experience and English proficiency.

The primary dependent variable was a *team* composite "goodwill" score. Goodwill points were awarded to individual players for such things as finding weapons caches and performing side quests. Points were subtracted for such things as angering the local populace and by opening empty crates. Although we have scores for each of 224 individual players, the four scores *within* a team are not independent of each other. This is because, for example, as one member of a team found a weapons cache, there necessarily was one less cache available for the other team members to find. But another reason is that teams were free to form their own search procedures and that meant that team members could be specialists. Communications Officers and coordinators might find no weapons and have very low scores. Those with weapon sensors would tend to have higher scores. What ultimately matters is how the *team as a whole* did. We thus used the sum of the four individual scores as the team metric. Since the raw scores have no inherent meaning, and to enable ready comparison of relative performance, I rescaled the raw team scores as *T-scores* which are simply re-scaled standardized z-scores with a mean of 50 and a standard deviation of 10 and which preserve the shape of the original distributions.

3.0 SELECTED PERFORMANCE RESULTS & DISCUSSION

Figure 1 showed the mean overall game-play performance for each of the 56 teams grouped by various culture compositions. As pointed out earlier, there is no simple function of cultural composition evident.

To aid in interpreting the data in Figure 1, Figure 3 shows the same 56 composite goodwill scores but with box plots superposed on the score-dots and with the jitter removed. The box plots help the eye remove the influence of outliers from interpretations while at the same time keeping the outliers in mind.

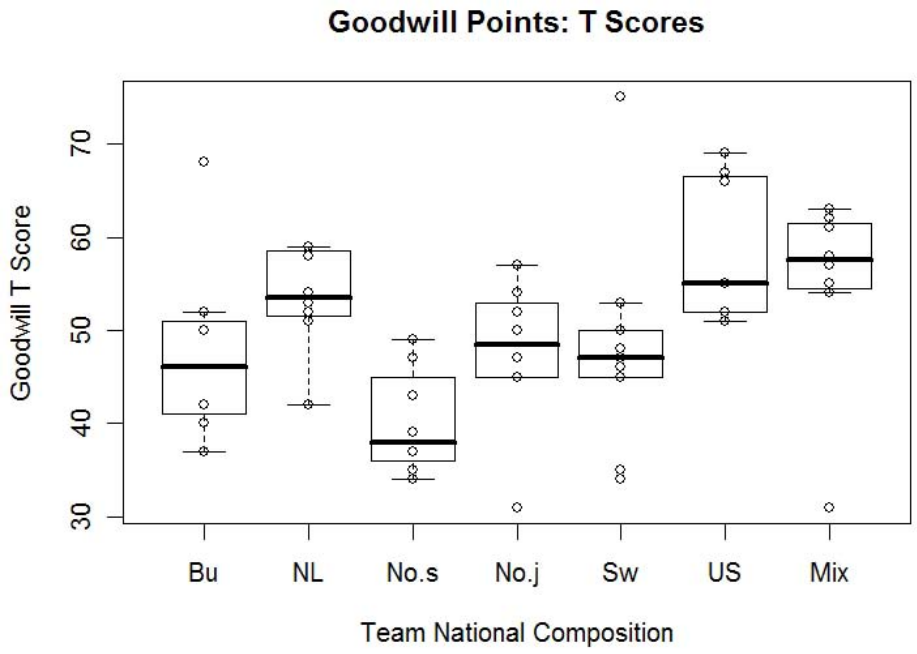

Figure 3: Overall game-play performance T-score (i.e., Mean = 50, SD = 10) for each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix). Same data as in Figure 1 but with jitter removed and box plots superposed on culture groups.

Several features of Figure 3 are relevant to our three variables of interest:

- Just comparing the two Norwegian sets of teams of junior versus senior officers shows a clear performance difference. All things being equal, we might expect the more senior teams to perform better, the results are just the opposite: The younger teams general perform better. Since it would be very counter-intuitive that military experience was not a positive factor, it is reasonable to suppose that an artifact---such as game-play experience---is operating. Thus we suspect that younger teams have more computer-game play experience.
- Note that even in a set of 8 scores it is possible to have outliers as can be seen in the Swedish scores.
- Once the very low-performing mixed-culture team is seen as an outlier, the overall relatively good performance of the mixed-teams is obvious: Although 5 homogeneous-culture teams out of 56 had better performance, the remaining 7 mixed-culture teams all had performance scores above the grand mean. This generally superior performance runs counter to expectations.
- As presumed native speakers of English and as the only native speakers of English, the American teams were expected to have an advantage in playing an English-only game. Figure 3 does show the overall relatively good performance of the American teams, but there are several non-American teams with equal or greater performance than individual American teams. The American teams also showed the most variability in performance as evidenced by the Inter-Quartile Ranges seen in the box plots.

The above points are suggestive.

The plan for analyzing the effects of age, computer-game experience, and English proficiency is to look at each individually in turn and then to look at them in combination.

4.0 ANALYSIS: AGE

This section presents the age profiles for the 56 teams, and the relationship of age and performance.

4.1 Age profiles

The 56 team mean ages ranged from 19.75 to 45.75 years. The median, mean, and SD team ages were 30.50, 31.25, and 6.62 years.

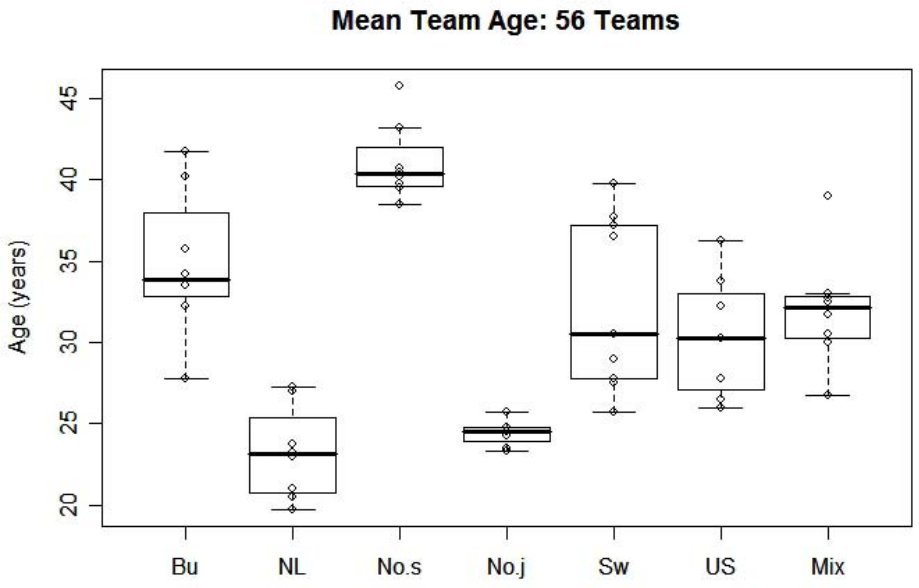

Figure 4: Mean age of each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix).

Figure 4 plots the 56 team age means grouped by national composition with box plots superposed on the culture groupings.

As can be seen in Figure 4, the team-age distributions varied considerable by national composition. Most striking is the large difference between the senior and junior Norwegian groups---justifying the labels "junior" and "senior." The Norwegian senior group is the oldest *as a group* and the Dutch group is the youngest. The mixed-culture teams, in particular, are squarely intermediate in age.

4.2 Age & Performance Relationships

Figure 5 is a scatterplot of age versus performance for all 56 teams.

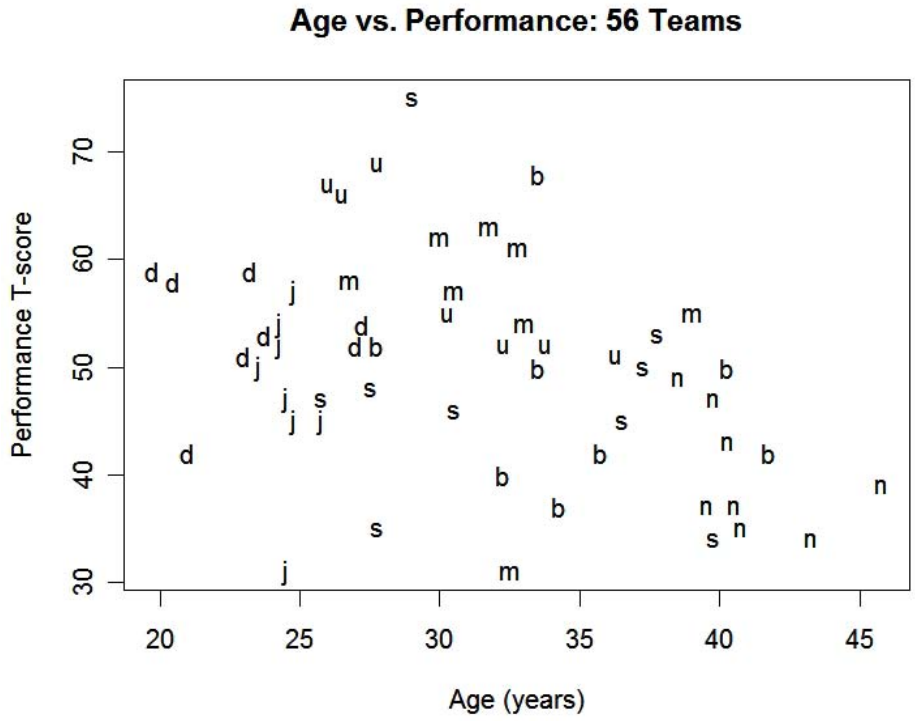

Figure 5: Age versus performance of the 56 teams. Points are coded for national composition of the teams: b: Bulgaria, d: Dutch (The Netherlands), j: Norway(junior teams), n: Norway(senior teams), s: Sweden, u: United States, m: mixed.

Quantitatively, the negative linear correlation between age and performance seen in Figure 5 is moderate and accounts for 15% of the variance ($r(54) = -.387$, r^2 = .1499, $F(1,54) = 9.52$, $p = .003$). The best-fitting linear equation for predicting (team) goodwill performance is

$$\text{Goodwill.T.score} = -.5895 * \text{age} + 68.3641 \qquad \text{Eq. 1}$$

Usually, a prediction equation with an $r^2=.15$ would be considered poor, but in this case it indicates a relatively weak effect of age on performance---which in our case is desirable.

4.3 Age-Adjusted Performance

The negative linear correlation between age and performance can be used to remove the effects of age and leave us with an "age-free" performance index.

This is accomplished by a two-step process: First, subtracting the performance values predicted using Eq. 1 from the original team performance T-scores leaves us with the residual "errors" of the linear prediction of performance using age. The residuals have a mean = 0 (by design) and an SD = 9.30 (empirically).

But the residual "error" in this case is actually goodwill performance---*less the effects of age*---provided we restore the original mean of 50 to all the residuals. This second step (adding 50 to the residuals) results in the age-effect-adjusted performance "T-scores." T-score is in quotes here since the SD has been left as 9.30 instead of being enlarged to 10 (as is needed for a true T-score).

The age-effect-adjusted or age-free T-scores are shown in Figure 6. To make the comparison with the un-adjusted scores easier, Figure 6 also shows corresponding boxplots from Figure 3.

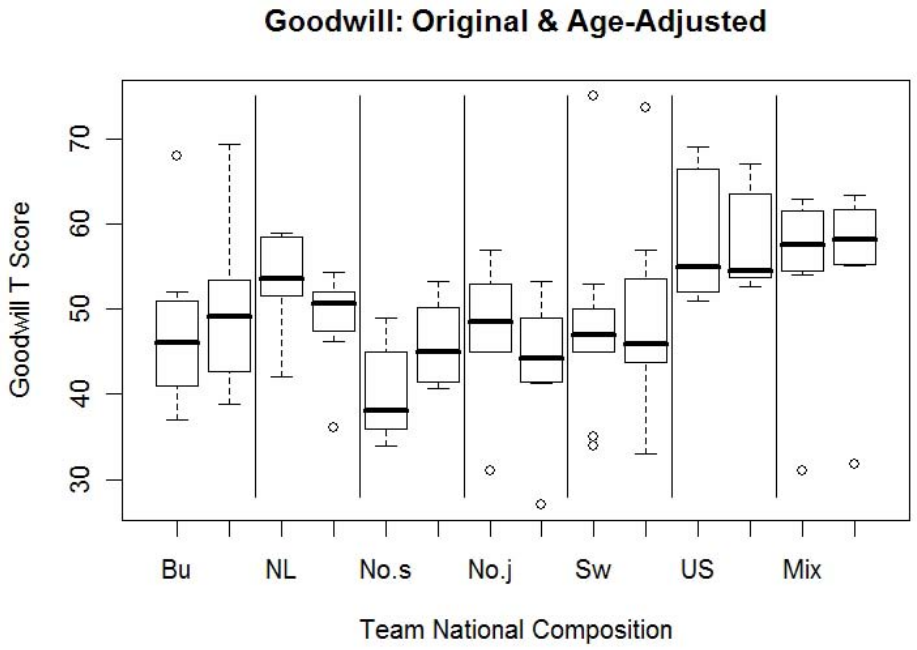

Figure 6: Team performance before and after age-effect adjustment. Original scores on left & adjusted scores on right within each culture sub-panel.

As shown in Figure 6, removal of the negative effect of age on performance raises the adjusted performance scores of the Bulgarian and senior Norwegian teams as a whole since they tended to be older in age. Another result is the adjusted performance of the junior Norwegian teams (as a whole) is lowered and the senior and junior Norwegian teams are more equal on adjusted performance. There is little difference between the performance and adjusted performance scores of the Swedish, American, and Mixed teams (again, considered as groups) since their ages tended to be in the middle of the age distribution.

5.0 ANALYSIS: COMPUTER-GAME EXPERIENCE

This section presents the computer-game experience of the participants and then treats the relationship of game-experience and performance. But unlike the section on age, a metric for computer-game experience had to first be developed.

5.1 Development of a Computer-game Experience Metric

This section only briefly treats the game-experience metric used in the analyses. For details of the metric and its development, see Warren (2008). Since the NATO RTO HFM-138/RTG Study Group anticipated game experience might be a factor, 14 computer and game experience questions were included in the pre-game survey asking about simple usage of games and "chatting" to advanced aspects such as developing "mods" for games. From these, I selected 10 questions in order to develop a game-experience metric:

Questions were scored as sub-scales for each person. Since performance is only meaningful on a team basis, team scores on each of the 9 sub-scales were formed by simply taking means over the 4 members for each of the 56 teams. As can be expected, the resulting sub-scales correlate to varying degrees with each other and also with the overall performance (goodwill) metric.

The composite experience metric which best correlates with performance can be sought using non-linear and smooth regression techniques (Venables & Ripley, 2002). However, I combined the sub-scales into a composite gaming experience metric using simple multiple linear regression and allowed an intercept term for a better fit. This yielded a metric with a correlation with team performance of $r(54) = .538$ and accounting for 28.9% of the variance.

The resulting 56 predicted values using the linear weights have two different interpretations: First, the predictive model being fit is:

$$\text{predicted goodwill.T.score} = \text{Sum}(\text{weight\_i} * \text{sub.scale.score\_i}) + \text{intercept} \qquad \text{Eq. 2}$$

so the resulting values are predicted (goodwill) performance (T-scores) as indicated by the left-side of Eq. 2. But, the right-side of Eq. 2 is a just weighted sum of sub-scale scores and such a weighted sum is exactly what we mean by a composite gaming-experience metric. (The intercept term is just an additive constant.) Hence, *the predicted performance scores also serve as our composite-experience scores*.

Similar to what was noted for Eq. 1 for predicting performance from age which accounted for 15% of the variance, a prediction equation accounting for just 29% of the variance would normally be considered poor. But in this case it indicates a weak to moderate effect of gaming experience on performance---and in our case, the weaker the effect the better.

5.2 Gaming-experience profiles

The 56 team mean composite gaming-experience scores ranged from 36.89 to 65.78. The median, mean, and SD team scores were 49.74, 49.95, and 5.42.

As was the case for age, there are wide differences in the experience distributions of the national-composition groups.

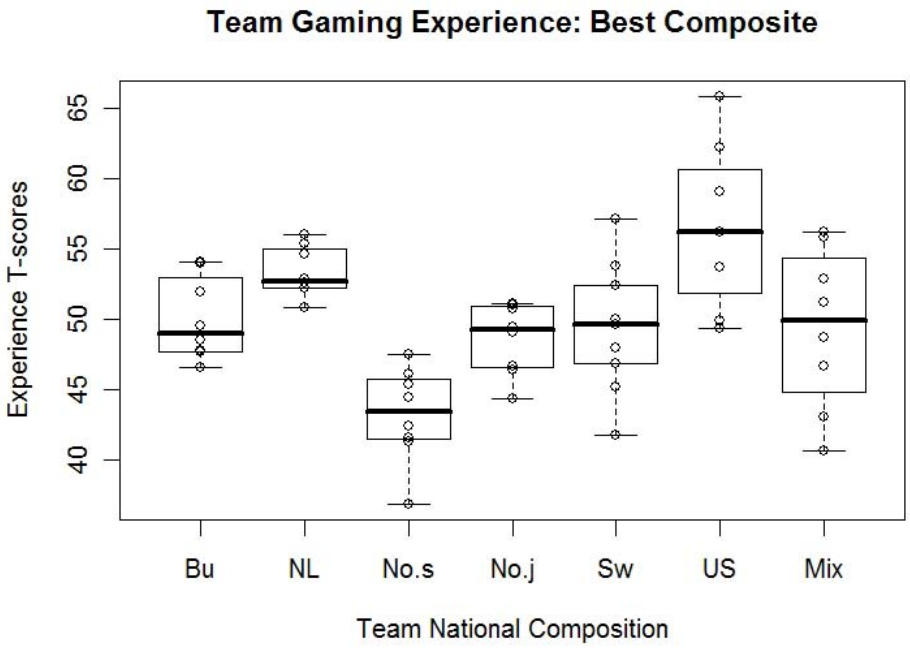

Figure 7: Composite gaming experience: 56 teamsComposite gaming-experience of each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix). Boxplots are superposed on datapoints.

Figure 7 plots the 56 team gaming-experience means grouped by national composition with box plots superposed on the culture groupings. As can be seen in Figure 8, the team-experience distributions varied considerable by national composition. The Norwegian senior group had the least gaming-experience *as a group* and the Americans the most. The mixed-culture teams, in particular, are squarely intermediate in experience.

5.3 Experience & Performance Relationship

Nations with more gaming experience tend to perform better than nation with less experience. This is consistent with the overall correlation of composite-experience with performance for the 56 teams which is shown as a scatterplot in Figure 8.

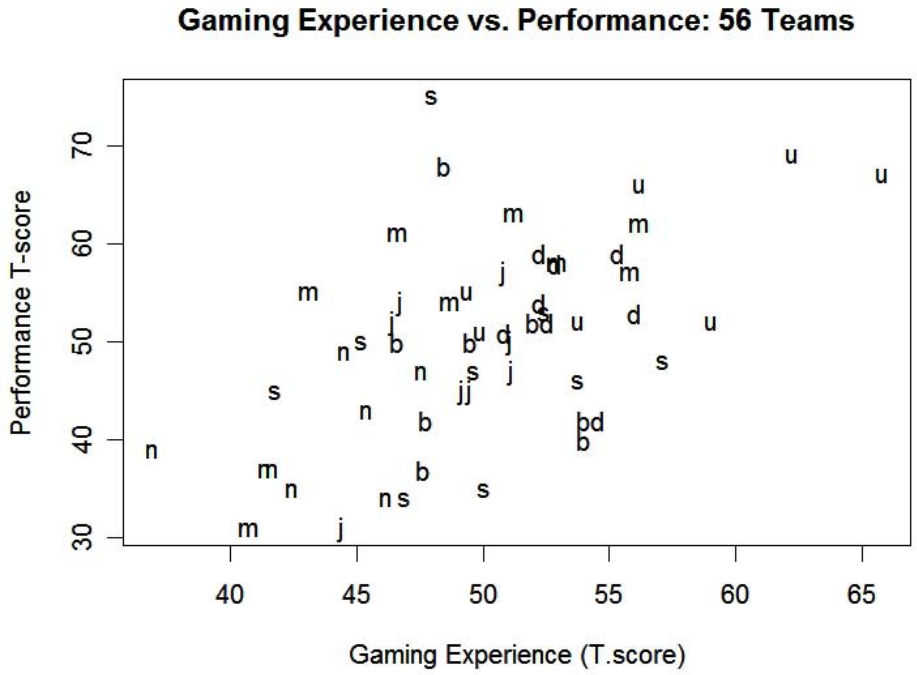

Figure 8: Gaming-Experience versus performance of the 56 teams. Points are coded for national composition of the teams: b: Bulgaria, d: Dutch (The Netherlands), j: Norway(junior teams), n: Norway(senior teams), s: Sweden, u: United States, m: mixed.

5.4 Gaming-experience adjusted performance

The positive linear correlation between gaming-experience and performance can be used to remove the effects of experience and leave us with an "experience-free" measure of performance. The procedure is the same one used in extracting an age-free performance measure. Since the experience scores are also the predicted performance scores in the multiple regression of the 9 experience sub-scales with performance, the residual "errors" formed by subtracting predicted performance from actual performance are then simply performance scores less the effects of experience. These residual performance or experience-free scores, as residuals, have a mean of 0 (by design) and an SD = 8.50. By adding 50 to all the residuals, we obtain a distribution with mean=50 and SD=8.50 --- a distribution of experience-free adjusted performance T-scores. These are shown in Figure 9 along with corresponding original performance boxplots from Figure 3 to make comparisons easier.

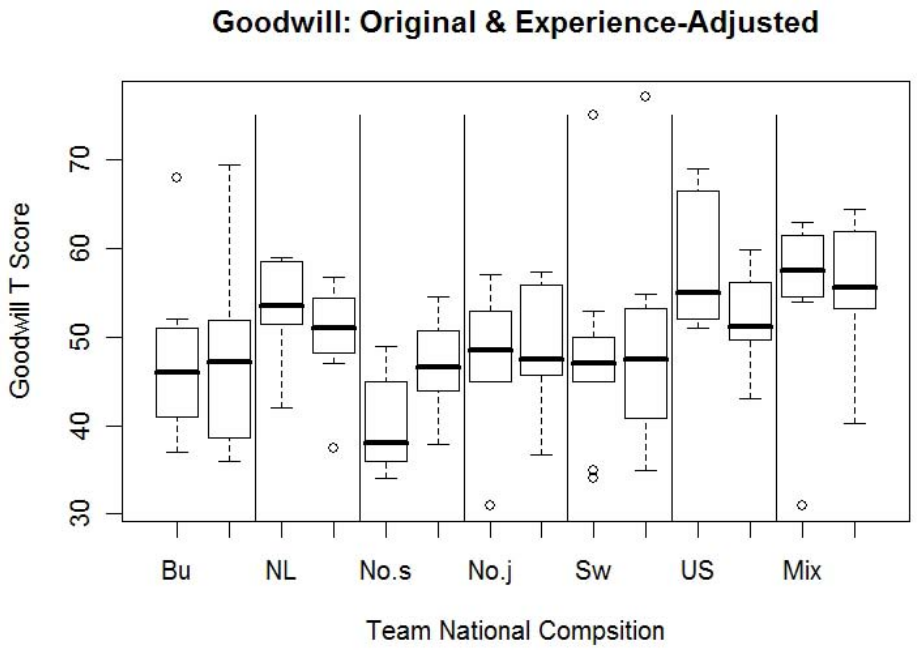

Figure 9: Team performance before and after removal of effect of gaming experience. Original scores on left & adjusted scores on right within culture sub-panels.

As can be seen in Figure 9, removal of the positive effect of prior gaming experience tends to equalize the performance of the teams considered as cultural groups. Most noticeable is the increase in the adjusted performance score of the senior Norwegian teams since they had the least gaming experience as a group. Also, the relative performance of the American teams is adjusted downward since they tended to have the most gaming experience as a group.

6.0 ANALYSIS: ENGLISH PROFICIENCY

This analysis paralleled that for gaming experience:

6.1 An English proficiency metric

I developed an English proficiency metric and prediction equation using responses on a pre-game questionnaire. For details, see Warren (2008). The resulting correlation of English proficiency and team performance was $r(54) = .5049$ ($t(54)=4.30$, $p=.00007$) and accounts for 25.5% of the variance. As with gaming, *the predicted performance scores also serve as the composite English-proficiency scores*.

As noted earlier for Eqs. 1 and 2 for age and gaming experience, a prediction equation accounting for 25% of the variance would normally be considered poor. But in this case it indicates a weak to moderate effect of English proficiency on performance---and again, the weaker the effect the better.

6.2 English proficiency profiles

The 56 team mean composite English-proficiency scores ranged from 36.89 to 65.78. The median, mean, and SD team scores were 49.74, 49.95, and 5.42.

As was the case for age and gaming experience, there are wide differences in the English proficiency distributions of the national-composition groups.

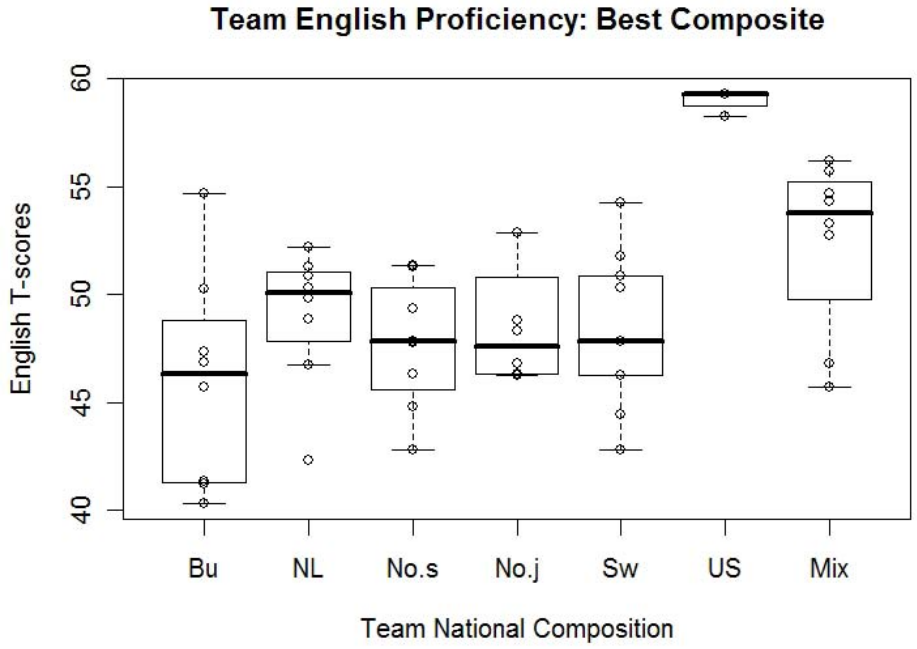

Figure 10: English proficiency of each of 56 teams grouped by national composition. Adjusted T-scores (Mean=50, SD=5.09). Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix). Boxplots are superposed on datapoints.

Figure 10 plots the 56 team English-proficiency mean adjusted T-scores grouped by national composition with box plots superposed on the culture groupings. As can be seen in Figure 10, the team-proficiency distributions varied considerable by national composition. The Bulgarians had the least English proficiency as a group and the Americans the most. The mixed-culture teams, in particular, are relatively proficient.

6.3 English Proficiency & Performance Relationships

Nations with more proficiency tend to perform better than nations with less proficiency. This is consistent with the overall correlation of composite English proficiency with performance for the 56 teams which is shown as a scatterplot in Figure 11.

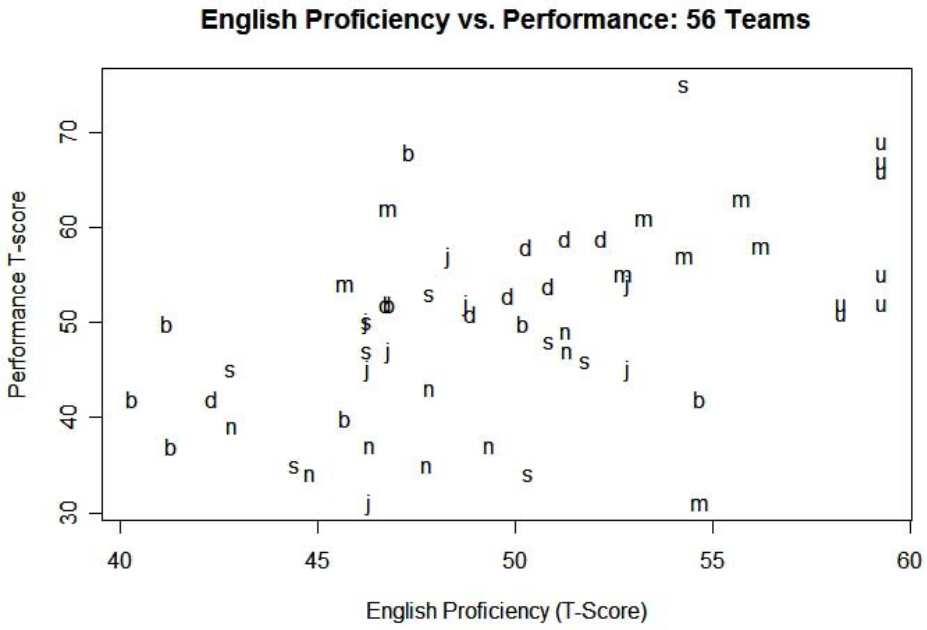

Figure 11: English proficiency versus performance of the 56 teams. Points are coded for national composition of the teams: b: Bulgaria, d: Dutch (The Netherlands), j: Norway(junior teams), n: Norway(senior teams), s: Sweden, u: United States, m: mixed.

6.4 English-proficiency-adjusted performance

The positive linear correlation between English proficiency and performance can be used to remove the effects of experience and leave us with an "English-proficiency-free" measure of performance. The procedure is the same one used in extracting the age-free performance and gaming-experience-free measures. Since the proficiency scores are also the predicted performance scores in the multiple regression of the 3 proficiency sub-scales with performance, the residual "errors" formed by subtracting predicted performance from actual performance are then simply performance scores less the effects of English proficiency. These residual performance or proficiency-free scores, as residuals, have a mean of 0 (by design) and an SD = 8.70. By adding 50 to all the residuals, we obtain a distribution with mean=50 and SD=8.70 --- a distribution of English-proficiency-free adjusted performance T-scores. These are shown in Figure 12, and to make comparisons easier, alongside the original performance scores.

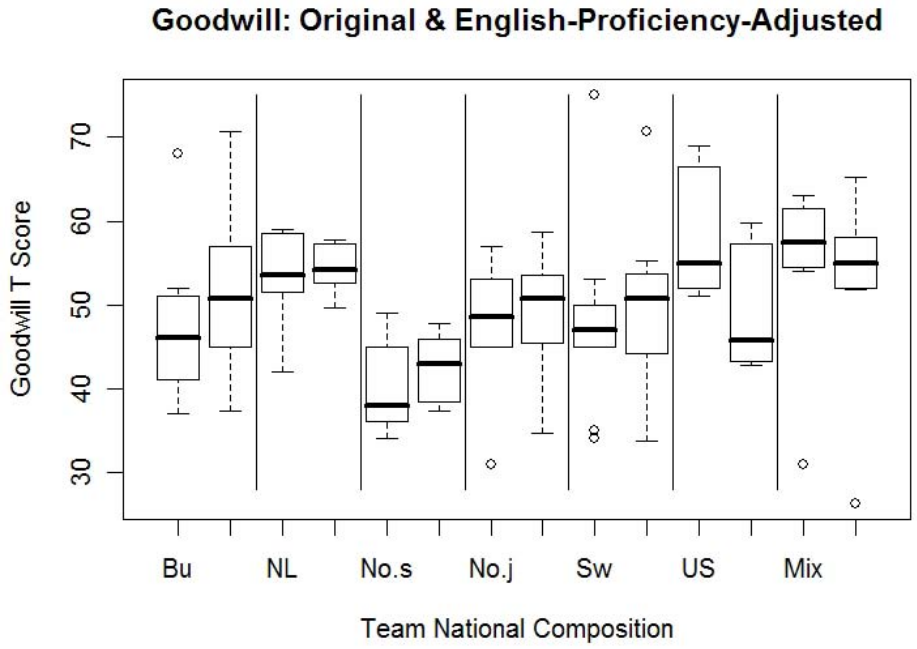

Figure 12: Team performance BEFORE & AFTER adjusting for English proficiency. Original scores on left & adjusted scores on right within culture sub-panels.

As can be seen in Figure 12, removal of the positive effect of English proficiency tends to equalize the performance of the teams considered as cultural groups. The performance of the Bulgarian and senior Norwegian groups have been adjusted upwards since they had relatively less English proficiency than the other groups. The relative performance of the American and Mixed-culture groups is adjusted downward since they tended to have the most English proficiency as groups. This pattern is similar to that found for gaming experience although the magnitude of the adjustments is less since the correlation of performance with English proficiency is less that with gaming experience.

7.0 ANALYSIS: AGE, GAMING & ENGLISH COMPOSITE EFFECTS

In the previous sections, the effects of age, gaming experience, and English proficiency on performance were assessed individually. In each case, a single effect was subtracted from overall performance to yield performance scores free of any effect of the specific chosen factor. But the resulting performance scores, although free of the effects of one confounding factor, still contain effects due to the other confounding factors.

In this section, the compound effect of all three confounding factors acting together are assessed and these compound effects are then subtracted from the original performance scores. The result is a measure of performance that is free of any effects of all three confounding factors acting simultaneously. As discussed below, these metrics are not independent of each other and have high intercorrelations.

7.1 Best-linear confound-free game-performance metric

As was the case for age, gaming experience, and English proficiency assessed individually, I used simple linear multiple regression to find the best composite linear predictor of performance and which thus has the maximum correlation with performance. An intercept term was allowed for a better fit.

As before, *the predicted performance scores also serve as the composite English-proficiency scores*. The resulting linear correlation of the aggregate confounding factors and performance yields a "grand" correlation of $r(54) = .6352$ ($t(54)=6.04$, $p=1.4E-7$) and accounts for 40% of the variance in performance compared to 15%, 25%, and 29% for age, English proficiency, and gaming experience respectively treated individually. As previously noted, a prediction equation accounting for 40% of the variance would normally be considered poor. But in this case it indicates a moderate effect of the combined confounds on performance---and again, the weaker the effect the better.

7.2 Scale intercorrelations

The grand confound composite scores and the sub-components of age, gaming experience, and English proficiency correlate to varying degrees with each other and also with the overall performance (goodwill) metric. Table 1 shows these correlations based on the scores of the 56 teams. To better assess the strength of association, Table 2 presents the squares of these correlations. Column 1 is of particular interest as it summarizes the variance of the performance scores accounted by the confounding factors singly and in grand combination. The factor accounts for less than the sum of its three components since the component confound are themselves intercorrelated. Of note is the large negative correlation of gaming experience and age ($p <$ .001 as are all first-column correlations). Also of interest is the insignificant correlation of English proficiency and age.

Table 1: Grand Inter-Correlation of Confounds & Performance Based on Mean Scores of 56 Teams

Scale	*Gdw*	*G.Exp*	*English*	*Age*	*Grand*
Goodwill	1.00	.54	.50	-.39	.64
Gaming Experience	.54	1.00	.43	-.53	.85
English Proficiency	.50	.43	1.00	-.15	.79
Age	-.39	-.53	-.15	1.00	-.61
Grand Composite	.64	.79	.85	-.61	1.00

Critical value: $r(54) = .263$, $p=.05$; $r(54) = .341$, $p=.01$

Table 2: Performance Variance Accounted By Confounds Based on Mean Scores of 56 Teams

Scale	*Gdw*	*G.Exp*	*English*	*Age*	*Grand*
Goodwill	1.00	.29	.25	.15	.40
Gaming Experience	.29	1.00	.18	.28	.72
English Proficiency	.25	.18	1.00	.02	.63
Age	.15	.28	.02	1.00	.37
Grand Composite	.40	.72	.63	.37	1.00

7.3 All-Confounds adjusted performance

The linear relation between the grand composite confounds "factor" and performance can be used to remove the effects of all three confounds and leave us with a "confound-free" measure of performance. The procedure is the same one used in extracting the previous individual confounding factor effects.

The residual "errors" formed by subtracting predicted performance (using the grand confound factor as the predictors) from actual performance are then simply performance scores less the effects of all 3 confounds. These residual performance or confound-free scores, as residuals, have a mean of 0 (by design) and an SD = 7.87. By adding 50 to all the residuals, we obtain a distribution with mean=50 and SD=7.87 --- a distribution of confound-free performance adjusted T-scores. These are shown in Figure 13, and to make comparisons easier, again in Figure 14 alongside the original performance scores.

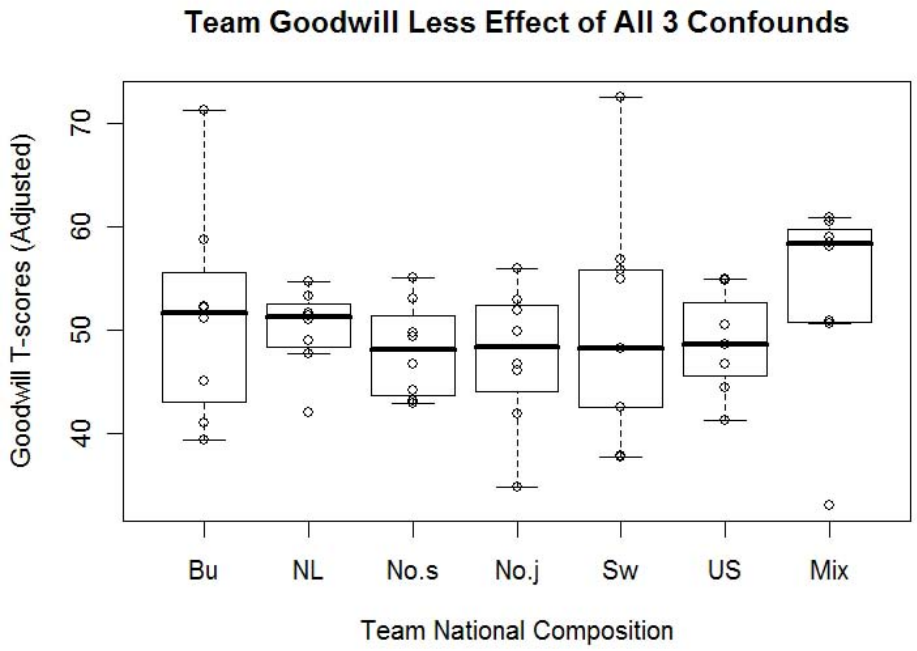

Figure 13: Game-play performance less effects of all 3 confounds (Adjusted T-scores, i.e., Mean = 50, SD = 7.87) for each of 56 teams grouped by national composition. Key: Bulgaria (Bu), The Netherlands (NL), Norway-senior age (No.s), Norway-junior age (No.j), Sweden (Sw), & the United States (US), Mixed culture (Mix). Compare with Figure 3. Box plots superposed on culture groups.

As can be seen in Figure 13, removal of the composite effect of all three confounds tends to equalize the performance of the (non-mixed) national teams considered as cultural groups. In fact, the central tendency of all six national groups is virtually the same (although there are differences in the within-group variabilities). It is interesting that the mixed culture teams as a whole now are at a performance level noticeably above the national groups. Possible reasons for this are considered in the Discussion section.

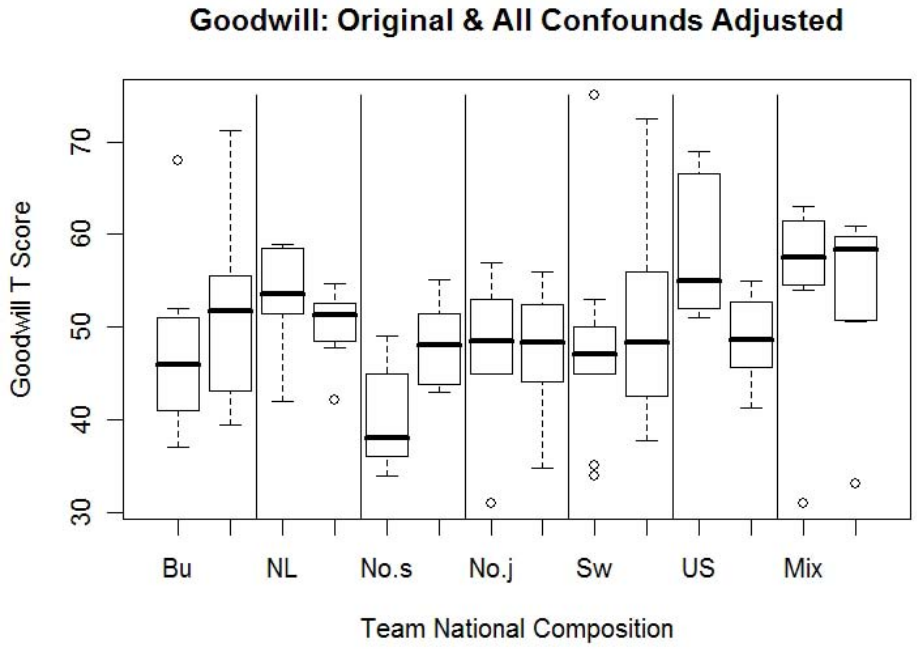

Figure 14: Team performance BEFORE & AFTER removal of effects of all 3 confounds. Original scores on left & adjusted scores on right within culture sub-panels.

Figure 14 shows that the biggest group-wise adjustments are those for the Bulgarian, senior Norwegian, and American groups. The performance of the Bulgarian teams *as a whole* have been adjusted upwards since they had relatively less English proficiency and game experience than the other groups. The performance of the senior Norwegian group also has been adjusted upwards primarily due to compensations for age and lack of game experience. The downwards performance adjustment of the American teams reflects compensation for native English proficiency and considerable computer-game experience.

The boxplots in Figures 13 and 14 visually, and the previous discussion in words, emphasize the relative positions and shifts of position for the national groups *considered as wholes*. That was deliberate as I wanted to focus on the general positions and shifts of the cultural groups as wholes. But we know from the outliers and other factors that not all teams within a national group conform to the pattern of their parent nation. A case in point is the particular Bulgarian team which was second in overall performance both before and after the removal of the three confound effects. This is clearly seen in Figure 15 which plots the before and after confound-removal goodwill performance of all the 56 teams. In Figure 15, vertical distance from the main diagonal indicates whether a particular team was moved upwards or downwards in performance after removal of the confound effects. Notice that some teams that were above the mean in original performance (indicated by the horizontal line) have been shifted to be even more above the mean after adjustment for confound effects. And as already pointed out for one Bulgarian team, some teams are shifted in the opposite direction from that of their parent group as a whole.

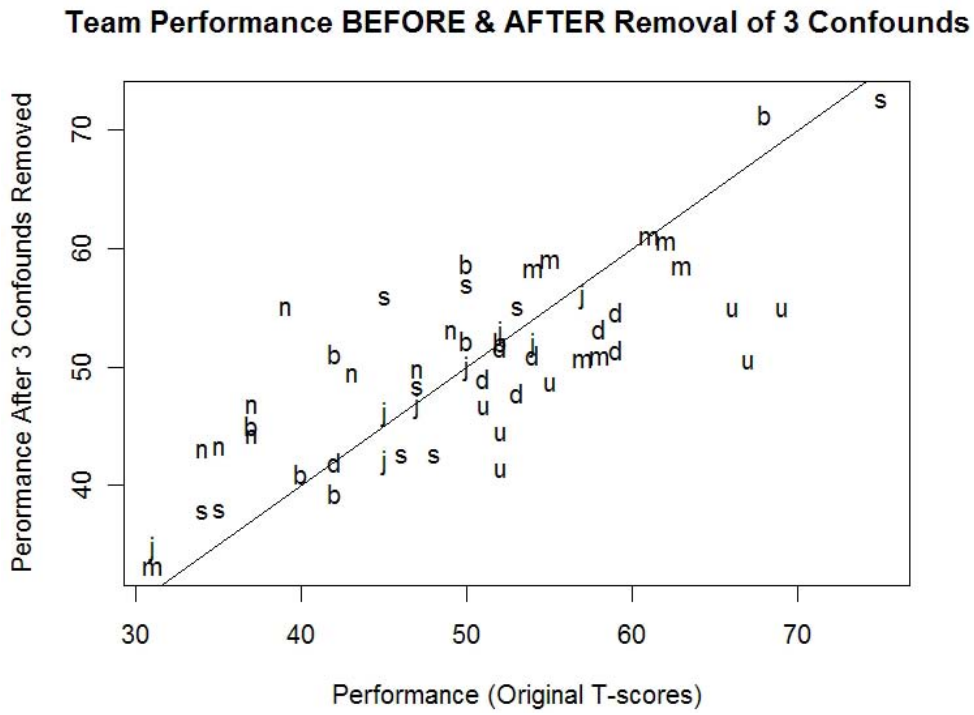

Figure 15: Scatterplot: Team performance BEFORE & AFTER removal of the effects of the 3 confounds. Vertical distance from main diagonal indicates amount of adjustment.

Both the parental group tendencies and the idiosyncratic behavior of individual teams have significance for recommendations and need to be further discussed.

8.0 DISCUSSION

It is reasonable to expect communication to be crucial for team members conducting a complex military task such as searching for hidden weapons in an urban environment. It is also reasonable to presume that effective communication should be easiest for people who share a common culture.

Hence, the principal hypothesis explored by the NATO RTO HFM-138/RTG study group was, as stated in the Introduction: *Homogeneous-culture teams (i.e., teams whose members are all from the same nation) perform better than mixed culture teams (i.e., teams whose members are from different nations).*

The hypothesis suggests an experimental design with a between groups factor, "Type of Team," with just two levels, namely, teams with a homogeneous culture and teams with a mixed culture. The hypothesis does allow us to "nest" sub-levels comprised of specific national compositions within the homogeneous culture level and thus allow for national differences to emerge. But the homogeneous culture teams, as a whole, are still expected to perform better than the mixed-culture teams.

As shown by Figures 1 and 3, the results were contrary to expectations: performance was not a simple function of team culture composition. Indeed, homogeneous-culture teams were not generally better than mixed-culture teams.

8.1 Questions raised by the results

What can account for the results? One possibility is that there were sources of non-random non-systematic variation between teams other than national composition. That this is the case is illustrated by Figure 2 which shows the profiles of the 56 teams with respect to age, English proficiency, and computer-game experience. The national and mixed teams exhibit clusters that are correlated with these factors.

The purpose of the current analysis was to assess the effects of these three confounding factors singly and in combination. A second purpose was to examine the results after the removal of the effects of the confounds. A number of questions can be asked:

- How strong are the confound effects and what are their relative importance?
- Since some effects were anticipated, why were the teams not better matched or the factors included in the design of the experiment?
- What accounts for the superiority of the mixed teams after the confound effects are removed?
- How can regression and ANCOVA be used for living with confound problems?
- What are the implications of the existence of the confounds?
- Irrespective of confounds, what makes some teams better than others?

8.2 The confound effects & their strengths

As shown in the first columns of Tables 1 and 2, the correlation of age and performance is negative and accounts for 15% of the variance. Next in strength is English proficiency which accounts for 25% of the variance. The strongest single effect is that of computer game experience which accounts for 29% of the variance by itself. All together, and due to the interactions among the three confounding variables, they account for 40% of the variance in performance.

The relationship of English proficiency to performance is not unexpected in a game permitting English-only communication. However, the 25% associative strength is not overwhelming and attests to the relatively high levels of English proficiency exhibited by the European participants. In fact, some of the 25% associative strength of English may be due to the 18% variance shared by English and computer-game-play experience. Hence, the "true" advantage of English proficiency might even by less.

The "contribution" of age to performance appears to be due almost entirely to its high negative correlation with gaming experience (r=-.53). The older generation has less computer-game experience than the younger generation. Interestingly, English proficiency and age are weakly and negatively correlated (r=-.15, not significant). Hence, any discussion of the implications of gaming experience to performance must keep the relationship of age and gaming experience in mind.

The most important confound that emerges is that of computer-game-play experience. It is not the only factor which must be considered since the impact of all three confounds taken together (40%) does exceed the impact of game experience by itself (29%) by 11%. But as a single factor, it can be expected to have an effect on performance in other computer-oriented tasks regardless of the language being used (even if no English is used whatsoever).

8.3 Why not use matching, counter-balancing, or factorial crossings?

Given that the effects of English proficiency and computer-game experience were somewhat anticipated, why were these variables not deliberately counter-balanced, varied factorially or at least matched in sample selection? This is not really a question of hindsight. The simple answer is that the subject pool (NATO officers matched in rank with reasonable English proficiency) is already highly limited. Adding other constraints such as certain levels of computer-game experience would greatly diminish an already scarce resource.

Even if a large pool of people were available, another problem, as the Analyses sections show, is that metrics for English proficiency and computer-game experience are determined after the fact from questionnaires administered after people have agreed to participate. No one single question or demographic datum (such as age) can provide the necessary information on which to match people or assign them to groups in a factorial design. Assuming a large enough subject pool exists (which is not the case), some metrics for English might be argued to exist (such as scores on a standardized test of English). But there is today no largely available and universally accepted computer-game-play scale which many people would already have taken and which could be used for pre-selection or factorial assignment purposes.

Even if such a readily available gaming-skill scale existed and people's skill levels known prior to participation, there is still a major barrier impeding the assignment of participants to an elegant experimental design: If we need teams of, say, four people with certain characteristics, we schedule six people to be prudent. However, all too often just three report for the experiment! This frustrating problem of "no shows" is endemic to team research and is independent the size of the available population.

Given this problem, it is remarkable that we were able to obtain 224 officers to from 56 intact teams for the experiment.

8.4 Confounds, regression techniques, & ANCOVA

Military teams are made of bright, creative, and well-trained individuals. When the team performance we are interested in researching is to be relevant to the real world, we must use complex scenarios and tasks which permit innovation and unpredictable behaviors to emerge. Further, when the teams may be geographically distributed and be comprised of members from multiple nations, confounds will be real, significant, omnipresent, and inescapable.

The researcher's task becomes not how to avoid the confounds, but rather how to gather useful information in spite of them. Since matching, counterbalancing, and factorial-crossing are not possible, we have a powerful ally in two statistical techniques: regression and the analysis of covariance (ANCOVA). ANCOVA itself is a combination of regression and analysis of variance. It capitalizes on the linear correlation of "covariates" with the dependent variable to eliminate systematic variance due to the covariates and thereby to reduce the within group error variance (Stevens, 2002).

Similar to analysis of variance, the focus is on the assessment of differences among means. Also similar to analysis of variance, ANCOVA requires that certain assumption be met. According to Stevens (2002, p. 347), ANCOVA rests on the same assumptions as ANOVA plus three additional assumptions concerning the regression aspects: (1) Linearity between the dependent variable and the covariates; (2) Homogeneity of the regression lines, planes, or hyperplanes (depending on the number of covariates); and (3) That the covariates are measured without errors. According to Stevens, violation of the assumptions is serious. Used properly, ANCOVA is a powerful and sophisticated technique for dealing with confounds.

However, I chose to use regression techniques without ANCOVA for a number of reasons. The relatively small number of values (7 to 9) for a relatively large number of national groups (7) means that the population estimates based on the samples may have large amounts of error associated with them. The sample sizes make use of exploratory data analysis (EDA) techniques more appropriate. Another reason is that the strict assumptions of ANCOVA, such as homogeneity of regression variance and error-free measurement of the covariates, were unlikely to have been met.

In addition to the technical reasons, a key reason for not using ANCOVA in the current analysis is that the focus here is not just on differences among means, but on comparing the full distributions within and among the national groups. Differences in the variances and skews of the group distributions are as great of interest as differences in central tendency. Especially with such small national group sizes (7 to 9), attention to the presence of outliers is crucial to proper assessment.

But the point here is that both techniques, ANCOVA and exploratory regression, can be powerful allies in studying team performance in complex situations in which confounding variables are manifold and rampant.

8.5 Why are mixed teams superior after de-confounding?

Figure 13 shows that the mixed teams, as a group, are superior to the homogeneous-culture groups after de-confounding. Indeed, Figure 13 shows that the median de-confounded performance score is above the 75th percentile of each of the distributions of all the other national groupings after de-confounding. This is exactly the opposite of what was expected.

Since the rationale behind the hypothesis (that communication is critical and that same-culture teams would have better communication) is still cogent, I will risk three speculations. Two are related and arise from the methodology and the third relates to possible consequences of group diversity.

The current analysis examined three possible confounds, namely, age, English proficiency, and computer-game experience. It is possible that yet two more confounds exist due to a procedural difference in the way data was collected for homogeneous culture versus mixed-culture teams:

Homogeneous-culture teams were geographically co-located and were tested in their respective home nations in the same building and often in same laboratory suite. Mixed-culture teams were geographically distributed (one person each in their home nation) and were tested over the Internet. Although all players were tested in their own cubicles and only communicated by keyboard during game play, same-site players were briefed together at the start of testing and could interact during breaks and lunch, whereas distributed-site players necessarily took their breaks and lunch apart from each other. Same-site players were instructed to not discuss the game/experiment during their breaks, but there was no way to monitor this. Further, some same-site players knew each other by virtue of working at the same site, whereas no distributed-site players knew each other before (or during) the game.

Hence, I speculate that:

- Distributed-play with strangers over the Internet sets up an atmosphere engendering a sense of seriousness of purpose and professionalism greater than that which might exist for colleagues playing at the same site.
- Since the distributed-site strangers are known to be from other nations, such a game environment might foster a sense of duty to perform at one's best out of national pride.

I emphasize that these two items are about increases in seriousness and motivation based on national pride. There is no suggestion here whatsoever that the homogeneous teams lacked seriousness or professionalism. Indeed, one of the reasons for using immersive role-play problem-solving games for research is that their very nature engenders a strong desire to perform well.

Although these putative two procedural confounds are almost untestable, they can be mitigated against in future research by testing all players over the Internet in different buildings even when they are from the same site. The identities of same-site players can be kept from each other as well.

Yet one more possible non-procedural reason for the superior performance of the mixed teams is that:

- Strangers, especially those from different nations, are likely more diverse in their backgrounds and training with respect to problem solving than team members from the same nation and even place of work. This greater cognitive diversity of the mixed teams might to lead to better decision making.

The facilitating effects of group diversity on decision making are based on many studies. See Surowiecki (2004/2005) for a popular review whose title *The wisdom of crowds* captures the essence of the effect.

8.6 Implications of the existence of confounds

Statistically removing the effects of some confounds from the data sets does not remove the reality of the effects of such variables such as age, computer-game experience, and English proficiency on performance. Age and gaming experience differences are real. Language differences are real. And distributed operations using mixed-nation teams are real.

The current analysis does not suggest avoidance of, or "work-arounds" to, the confounds. Rather it calls for an awareness of their presence and effects so that their consequences may be consciously taken into account in team-formation, team training, team operations, and team performance assessment.

For example, tomorrow's military recruits are today playing multi-player computer games over the Internet with team members they have never met face-to-face in contradistinction to the recruits of yesterday. The skill sets and mind sets of these recruits must be taken into account and capitalized on.

8.7 What makes some teams better than others?

The removal of confounding effects erases differences *between* the national groups, but it does *not* remove differences *within* national groups. As can be seen in Figures 13 and 14, there is still considerable variability among the 56 teams in overall goodwill performance.

Possible motivational and team diversity reasons for the differences have already been discussed. What has not been discussed are the variables explored in the main report of NATO RTO HFM-138/RTG (2008). These include quality and quantity of mission planning, quantity of communications, quality of communication content, team organization and assignment of sub-tasks, and team situation awareness. All these variables remain pertinent to our need to understand why some teams perform better than others. Age, computer-game experience, and English proficiency are just a part of what differentiates teams.

REFERENCES

[1] Leung, A., Diller, D., & Ferguson, W. (2005). SABRE: A game-based testbed for studying team behavior. *Proceedings of the Fall Simulation Interoperability Workshop (SISO).* Orlando, FL, September 18-23, 2005.

[2] NATO RTO HFM-138/RTG (2008). Final Report of the NATO RTO Human Factors and Medicine Panel Research Task Group 138 on "Adaptability in Multinational Coalitions." Brussels: NATO.

[3] Stevens, J.P. (2002). *Applied multivariate statistics for the social sciences*, (4th Ed.). Mahwah, NJ: Erlbaum.

[4] Surowiecki, J. (2004/2005). *The wisdom of crowds*. New York: Anchor Books.

[5] Venables, W.N., & Ripley, B.D. (2002). *Modern applied statistics with S*, (4th Ed.). New York: Springer.

[6] Warren, R. Sutton, J., Diller, D., Ferguson, W., & Leung, A. (2004). A game-based testbed for culture & personality research. *Proceedings of the NATO Modeling and Simulation Group -- 037 Workshop: Exploiting Commercial Games for Military Use*. The Hague, The Netherlands, 20-21 Oct 2004.

[7] Warren, R., Diller, D.E., Leung, A., Ferguson, W., & Sutton, J.L. (2005). Simulating scenarios for research on culture & cognition using a commercial role-play game. In M.E. Kohl, N.M Steiger, F.B. Armstrong, and J.A. Jones, (Eds.), *Proceedings of the 2005 Winter Simulation Conference*. Orlando, FL.

[8] Warren, R. (2208). Mixed- & Homogeneous-Culture Military Team Performance on a Simulated Mission: Effects of Age, Computer-Game Experience & English Proficiency (AFRL-RH-TR-08-xxxx). Wright-Patterson AFB: Air Force Research Laboratory. Approval pending.