Improving Scoring Consistency of Flight Performance through Inter-Rater Reliability Analyses

Matthew V. Smith, Mary C. Niemczyk, and William K. McCurry
Arizona State University

ABSTRACT

Students, as well as the other stake-holders of flight schools, must be sure that the scoring of flight performance is such that the scores are a meaningful indicator of the student's performance rather than an arbitrary indicator of the instructor's perception. The scores should be somewhat consistent from one instructor to another. The apparent inconsistency in scoring from one instructor to another can be examined by conducting inter-rater reliability (IRR) analyses. Inter-rater reliability measures the extent of agreement between two or more individual raters – it is used to measure the consistency of a scoring or rating system, and those who use it. This foundational investigation was designed to assess inter-rater reliability between instructor pilots when observing 10 sample flights performed by student pilots. Results of the study indicated that inter-rater reliability was low. Suggestions for improving the consistency of flight instructor scoring are discussed, as well as recommendations for future research.

INTRODUCTION

There are many different organizations that offer flight training, whether it is a local Fixed Base Operator (FBO), or a two – or four-year college program. Though ground school and written exams issued by the Federal Aviation Administration (FAA) are standardized, training from school to school may not identical, even though fully compliant with FAA regulations. Even within a flight school that has very exacting standards, training may vary between flight instructors for any number of reasons, such as the instructors' abilities, experience level, and perhaps interests. Regardless of their personal characteristics, all instructors must do one thing: evaluate student performance. And yet, because of their personal characteristics, experience, and training, instructors may perceive student performances differently from one another. The reasons for differences in instructor perception of student performance can be systematic or arbitrary, conscious or subconscious, innocuous or malicious; one simply cannot catalog another's motives, but one can see the result of the instructors' perceptions: difference.

When scoring a student pilot, there is the student pilot's performance, which is objective, and the instructor pilot's perception of that performance, which is subjective. In the best of circumstances, the performance and the recorded perception of that performance share a high degree of similarity. That is, the instructor ought always to record a score that accurately and precisely reflects the student's performance. However, this is not always the case. Some perceptions of performance are too forgiving, while others are overly critical. In other words, the same student pilot can receive a passing score from an overly forgiving instructor and a failing score from an overly critical instructor for an identical or near-identical performance, leaving the student confused or frustrated.

Students, as well as the other stake-holders of flight schools, must be sure that the scoring system is such that the scores are a meaningful indicator of the student's performance rather than an arbitrary indicator of the instructor's perception. Furthermore, the scores should be consistent from one instructor to another.

The apparent inconsistency in scoring from one instructor to another can be examined by conducting inter-rater reliability (IRR) analyses. Inter-rater reliability is "used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon" (Trochim, 2001, p.96). Inter-rater reliability measures the extent of agreement between two or more individual raters – it is used to measure the consistency of a scoring or rating system, and those who use it (DeVellis, 2005; Trochim, 2001). The purpose of this investigation, then, is to determine the inter-rater reliability of instructor pilots when evaluating student pilot performance.

REVIEW OF THE LITERATURE

After an extensive review of the literature, published articles focusing on IRR in aviation were not found. There were, however, many other examples of IRR studies conducted extensively in other fields, such as sports, psychology, health care, and education.

Inter-rater Reliability in Sports

Flying and sports are related activities in that they are both simultaneously physical and mental, or psychomotor, to denote the inseparability between the physical and mental aspects. One such study, *Development of an Instrument to Assess Jump-Shooting Form in Basketball* (Lindeman, Libkuman, King, & Kruse, 2000), examined the physical form and movements of a jump shot. A scoring instrument for assessing jump-shots was developed based on the expertise of several recognized basketball coaches. Four raters viewed video tapes of 32 shooters and rated the shooters' form and movement according to the scoring instrument developed. The conclusion was that the instrument may help discern a correlation between the shooter's form and the shooter's success rate. This study shows the applicability of an inter-rater reliability analysis when evaluating psychomotor activity scoring. An inter-rater reliability study may, therefore, be appropriate when evaluating flight performance scoring, since flying an aircraft is also a psychomotor activity.

Inter-rater Reliability in Psychology

Inter-rater reliability studies are often used in psychology to determine if scales and other methods of measuring patient behavior are reliable means of assessment. These studies have been used to assess rating scales and assessment methods related to sleep disorders (Ferri, Bruni, Miano, Smerieri, Spruyt & Terzano, 2005), mental capacity (Raymont, Buchanan, David, Hayward, Wessley & Hotopf, 2006), agoraphobia (Schmidt, Salas, Bernert & Schatschneider, 2005), delusions (Bell, Halligan & Ellis, 2006 and Meyers, English, Gabriele, Peasley-Milkus, Heo, Flint, et al., 2006), social dysfunction in schizophrenia and related illnesses (Monroe-Blum, Collins, McCleary, & Nuttall, 1996), and other means of rating psychological disorders (Drake, Haddock, Terrier, Bentall & Lewis, 2007).

Using inter-rater reliability studies to validate psychological testing is not limited to the United States. It has also been used in China (Leung & Tsang, 2006), Korea (Joo, Joo, Hong, Hwang, Maeng, Han, et al., 2004), Japan (Kaneda, Ohmoria & Fujii, 2001), in the Arabic language (Kadri, Agoub, El Gnaoui, Mchichi Alami, Hergueta & Moussaoui, 2005), Turkey (Tural, Fidaner, Alkin & Bandelow, 2002), Greece (Papavasiliou, Rapidi, Rizou, Petrapoulou & Tzavara, 2007 and Kolaitas, Korpa, Kolvin & Tsiantis, 2003), and France (Thuile, Even, Friedman & Guelfi, 2005). In all of these articles, scales or other methods of assessment were tested, and validated using inter-rater reliability studies.

Inter-rater Reliability in Health Care

Training health care practitioners also has parallels to training pilots. Both health care and flying require mental aptitude and physical skills. Bann, Davis, Moorthy, Munz, Hernandez, Khan, Datta, and Darzi (2005) studied 11 surgical trainees and put them through a 15 minute, six-station rotation of basic surgical tasks. One of the results of this experiment confirmed that video assessment is a reliable means of assessing performance. A similar study concluded that inter-rater reliability of video taped cases was excellent, having a reliability coefficient of .93 (Hulsman, Mollema, Oort, Hoos & de Haes, 2006)

Inter-rater reliability studies are not used solely in the training of health care professionals, but also to verify the rubrics for rating the effectiveness of out-of-hospital CPR (Rittenberger, Martin, Kelly, Roth, Hostler, & Callaway, 2006) and for rating the severity of rosacea (Bamford, Gessert, & Renier, 2004). Bamford, Gessert, and Renier (2004) reported that a scoring rubric with a scale ranging from 1 to 10 may tend to provide an unreliable rating, but when the scale was reduced to a range from 1 to 5, the inter-rater reliability coefficient was much greater, indicating reliability.

Inter-rater Reliability in Education

In *An analysis of statistical techniques used in the Journal of Educational Psychology, 1979-1983*, Goodwin and Goodwin (1985) reported

that from 1979-1983, 40 out of 92 reliability studies in the Journal of Educational Psychology were inter-rater reliability studies, comprising nearly half of the studies, by far the greatest percentage. Considering how commonly researchers use inter-rater reliability studies to establish or verify reliability in an educational setting, the Goodwin's article indicates that performing an inter-rater reliability study at flight schools is a legitimate pursuit.

CALCULATING INTER-RATER RELIABILITY

In his 2005 entry into the Encyclopedia of Social Measurement, Robert F. DeVellis reported that there are two influences at work in the process of measuring scores: "(1) the true score of the object, person, event, or other phenomenon being measured, and (2) error (i.e. everything other than the true score of the phenomenon of interest)" (p. 315). A true score is considered to be an objective performance with the opportunity for error resulting from the instructor's perception. The instructor's perception is susceptible to error, thus the disconnect between the true score (objective performance) and the recorded score (instructor's perception). Error is simply a phenomenon to be dealt with through statistical processes and analysis.

In order to get a clear depiction of the level of agreement between raters, consideration must be given to agreement between raters due to chance; chance being a type of error. A thorough review of the inter-rater reliability literature found that Cohen's kappa coefficient was used extensively to test for chance-corrected agreement. Though there are other means (coefficients) of determining inter-rater reliability, Cohen's kappa was used in this study due to its wide use in other IRR investigations.

Cohen's Kappa Coefficient

In the late 1950's and throughout the 1960's, Jacob Cohen conducted seminal research focusing on inter-rater reliability. Cohen proposed a coefficient represented by the Greek letter kappa (κ), as the standard coefficient for inter-rater reliability, with $\kappa \geq .70$ being considered reliable. This is not merely a 70% agreement, because agreement can happen by chance, instead, kappa accommodates the expected frequency of ratings; thus eliminating mere chance agreement (Cohen, 1960; Gwet, 2002b).

A study conducted by Holey and Watson (1995) provided a stark example of the necessity for kappa rather than using mere percentage of agreement when performing an inter-rater reliability study. In their study, some cases resulted in a percentage of agreement between raters of 100%, while the kappa coefficient, which accounts for chance agreement, was 0.01, the absolute lowest number possible.

The purpose of the kappa statistic is to account for and eliminate agreement by chance, chance being a type of error, so that the researcher can get a clearer idea of how much agreement there really is between raters. The coefficient, then, distinguishes between purposeful agreement and accidental agreement. In a reliability formula, the quantified possible error becomes the denominator, while the quantified true score is the numerator. Thus, whatever reliability coefficient is used it is the "ratio of variability ascribable to the true score relative to the total variability of the obtained score" (DeVellis, 2005). Or, in the terms chosen for this investigation, it is the ratio of the pilot's objective performance and the instructors' recorded perception of that performance. In this study, it is assumed that any disconnect in the relationship between the pilot's performance (true score) and the instructors' recorded perception (obtained score) is due to the raters, not the pilot.

The way to find this coefficient, then, is to measure rater against rater rather than pilot against rater. Each rater observes the same flight performance; therefore, the raters ought to record identical scores. In practice they may or may not. This is why one performs an inter-rater reliability study, to discover these discrepancies between true score and obtained score, should discrepancy (error) exist.

METHODOLOGY

This investigation was designed to assess inter-rater reliability between instructor pilots when observing flights performed by student pilots. The study included videotaping the performance of student pilots flying an industry

standard instrument flight rules (IFR) pattern. Four instructor pilots reviewed the recorded flight performance footage and scored the performance of 10 student pilots' on a scale of 1 to 5. A score of 1 represented an unsatisfactory performance; 2, marginal; 3, good; 4, very good; and 5, excellent.

Flight Pattern

In *The Pilot's Manual: Instrument Flying* (Kirshner, 1990) there are several flight patterns to choose from. The pattern used for this investigation is referred to as Pattern D. It was chosen because it is long enough to give the raters something substantial to score, yet not so time-consuming as to prove burdensome.

Pilot Participants

Student pilots enrolled in a flight program at a four-year research university participated by flying the aforementioned flight pattern using a PCATD. The researcher explained to the students that they were being videotaped for the purpose of investigating inter-rater reliability. They were assured that these scores, good or bad, would not figure into their course average. Their identities were protected by preventing any distinguishing features from being recorded on video. Also, the order in which the flight performances were viewed was different from the order they were recorded. Thus, the student who flew the first flight on the day of recording might have actually have been the last flight viewed by the raters.

Rater Participants

The rater-participants were selected from the pool of instructor pilots at the flight school. All instructor pilots were offered a chance to participate, resulting in four volunteers. These instructor pilots watched and scored the videotaped flights. The raters were assured of their anonymity and that their performance in this study would not impact their employment at the flight school. Also just as with the student pilot participants, the researcher did not collect or record any demographic data about the rater participants. There is nothing to indicate that the results would have been better or worse with more or fewer raters because the literature found did not suggest an optimal number of raters to use. A future researcher could find an optimal number based of further experimentation.

Scoring Rubric

In order to measure inter-rater reliability, a scoring mechanism, such as a rubric, must be used. The flight school at which this study was performed already had a scoring rubric and that same rubric was used in this investigation.

Flying the Pattern

Prior to sitting at the PCATD, the researcher briefed the student pilots. The pattern is rather complex, and depending on the skill of the student pilot, the researcher gave verbal instructions, if necessary. The student pilots' ability to perform the flight pattern well or poorly was immaterial. The raters were entirely unaware of which student referred to the pattern and which students performed the pattern from memory.

DATA COLLECTION PROCEDURE

The experiment was conducted in a classroom equipped with a PC, projector, and movie screen. The four raters sat in the same room, but were seated far apart to prevent communication between them. They were given instructions and a score sheet and were briefed by the researcher about how to behave during the test (i.e. no talking, gesturing, or using other means of communicating during flights, no talking about the flights during break times, etc.). It took three hours to watch all of the flights. Two short breaks and one longer break were included.

RESULTS

Raw Scores

The raters watched the flights and marked their scores on the score sheet that was provided. These scores are not averages of aspects of the flights such as altitude, heading or air speed scores, but rather single scores for the entire flight. The raw scores are shown in Table 1.

The numbers 1 through 5 indicate the scores the raters gave to each of the 10 flight performances. A score of 1 represents an unsatisfactory performance; a 2, marginal; a 3, good; a 4, very good; and a 5, excellent.

Table 1. *Flight Performance Scores by Rater*

	Sample Flight									
Rater	A	B	C	D	E	F	G	H	I	J
1	4	5	2	1	4	3	2	3	1	5
2	4	5	1	1	4	4	2	4	1	3
3	3	3	1	1	3	4	1	4	1	2
4	3	5	1	1	3	3	2	4	1	4

At first glance, these scores appear to show good agreement, especially in sample flights C, D, G, H and I. A brief examination of the raw scores also reveals that Rater 1 evenly distributed the scores; the only rater to do so. Raters 2 and 4 had very similar results, with only disagreement being between a score of 3 and 4. Rater 3 gave the most scores of 1, and gave no scores of 5. However, to properly analyze the data for inter-rater reliability, the raw scores were analyzed using the methodology of Cohen's kappa coefficient.

Contingency Tables Used to Calculate Cohen's Kappa Coefficient

Cohen's kappa coefficient is derived using only two raters, therefore, six contingency tables were developed. Table 2 is the contingency table for Rater 1 and Rater 2 and is provided as an example.

Table 2. Agreement/Disagreement between Rater 1 and Rater 2

		Rater 1							
	Score	1	2	3	4	5	Row Totals:	*a*	*ef*
Rater 2	1	2	1	0	0	0	3	2	.6
	2	0	1	0	0	0	1	1	.2
	3	0	0	0	0	1	1	0	.2
	4	0	0	2	2	0	4	2	.8
	5	0	0	0	0	1	1	1	.2
Column Totals:		2	2	2	2	2	*N*	*Σa*	*Σef*
							10	6	2

Given: $N = 10$, $\Sigma a = 6$, $\Sigma ef = 2$

$\kappa = (\Sigma a - \Sigma ef) \div (N - \Sigma ef) = (6 - 2) \div (10 - 2) = 4 \div 8 =$ **.50**

In Table 2, (Rater 1 versus Rater 2), {1,1}, meaning that both Rater 1 and Rater 2 each provided two sample flight performances with a score of 1, unsatisfactory. Both Raters had one agreement of a score of 2, marginal {2,2}; no agreement for a score of 3, good {3, 3}; two agreements for a score of 4, very good {4,4}; and one agreement for a score of 5, excellent {5,5}. The total number of agreements (Σa) between Rater 1 and Rater 2 was six.

As shown in Table 2, *N* equals 10, the number of sample flight performances. Column *a* is the number of agreements. This number is simply the cells showing agreement (e.g. 2, 1, 0, 2, 1) transferred over to a single column. In order to account for chance agreement, the expected frequency (*ef*) is determined by dividing the product of the row and column totals by the number of samples, (*N*), 10. This is the expected frequency by chance.

To find kappa, then, the difference of Σa minus Σef is divided by the difference of *N* (number of samples) minus Σef (sum of expected frequency). That is: $\kappa = (\Sigma a - \Sigma ef) / (N - \Sigma ef)$.

Kappa is evaluated next. As was stated previously, a kappa of .70 or greater is considered satisfactory; less than .70 is not.

This calculation was done for each possible permutation without replicating pairs. After the result of each table was tallied, the resultant coefficients were then analyzed to determine the inter-rater reliability of the instructor pilots in comparison with each other.

Summary of Results

The scores were tallied and the kappa for each rater pair calculated. As stated previously, the minimum desirable kappa coefficient is .70. The results in this study were markedly lower.

Table 3. Summary of Kappa for Each Rater Pair

Rater Pair	Kappa
Rater 1 vs. Rater 2	.50
Rater 1 vs. Rater 3	.00
Rater 1 vs. Rater 4	.50
Rater 2 vs. Rater 3	.38
Rater 2 vs. Rater 4	.47
Rater 3 vs. Rater 4	.44
Average	.38

The best kappa was .50, and the worst, 0. The average kappa coefficient was .38, just over half of the desired .70.

Although all of the rater pairings in this study fell far below .70, one rater, Rater 3, seemed the least reliable of the four. The three pairings in which Rater 3 was involved were the least reliable, one of which had a kappa of 0, entirely unreliable. Rater 1, with whom Rater 3 shared the kappa of 0, enjoyed the two highest reliability scores, .50, with Raters 2 and 4.

Each rater was paired three times. When each rater's three pairings were averaged, Rater 1 scored a .33, Rater 2, .45, Rater 3, .27, and Rater 4, .37. However, removing Rater 3 from the averages, so that each rater was only paired twice, Rater 1's average rose to .50, Rater 2 to .48 and Rater 4 to .48. Among Raters 1, 2 and 4, the scores are extremely similar (pair 1 & 2 .50, pair 1 & 4 .50 and pair 2 & 4 .47). Thus it seems that removing Rater 3 improved the inter-rater reliability in this study. Without Rater 3 the overall average reliability increased from .38 to .49. This is still well below .70, but much better.

DISCUSSION

This investigation was designed to assess inter-rater reliability between instructor pilots when observing 10 sample flights performed by student pilots. Four instructor pilots reviewed the recorded flight performance footage and scored the performance on a scale of 1 to 5. A score of 1 represented an unsatisfactory performance; 2, marginal; 3, good; 4, very good; and 5, excellent. Inter-rater reliability was determined by using Cohen's Kappa coefficient. Ultimately, the study indicated that the inter-rater reliability was low; having an average kappa of .38, well below the desired .70.

The resultant coefficients are such that the study did not yield good inter-rater reliability. Because of this, steps should be taken to improve inter-rater reliability at the flight school. Two suggestions are to engage in extensive recurrent training and to improve the scoring rubric.

Recurrent Training

These scores show low inter-rater reliability which may indicate the need for recurrent training, which may help the flight school reinforce the scoring criteria. In the case of Rater 3, more training would be required than for Raters 1, 2 and 4. In sample C, while Raters 1, 2 and 4 agreed upon a score of 5, Rater 3 awarded a score of 3. In sample G where all others gave a score of 2, Rater 3 gave a 1. And in Sample J, where there was no agreement among any raters, Rater 3 gave the low score of 2. After examining the raw scores, it is evident that the most common disagreement was between the scores 3 and 4. It may be that Raters 1, 2 and 4 need to review the scoring standards to help them differentiate between performances that rate a 3 rather than a 4, while Rater 3 needs a greater amount of training to align that rater's expectations of student performance with flight school standards.

It may also be helpful to begin training instructor pilots how to interpret the standards used to score student pilot performance first using simple maneuvers and working their way up to complex patterns, just as the students

themselves must work their way up from simple maneuvers to complex patterns. This recurrent training may be of little use unless the standards are better defined through an improved scoring rubric.

Scoring Rubric Improvements

It could also be that the scoring rubric needs improving. There seems to be a disconnect between the description of the quality of performance and quantifiable data. For example, "An 'Excellent' (5) grade will be issued when a student's performance far exceeds and is well above the completion standards." Unfortunately, there is little to define exactly what makes a performance far exceed or well above the completion standards. The same can be said for scores 4, 3, 2, and 1. The definitions of the scores may be too broad.

The scoring sheet provided the rater the completion standards from the lesson in which Pattern D is taught. The altitude standard states only that a student pilot must remain within plus or minus 200 feet of the starting altitude. This standard is very broadly defined and leaves too much open to interpretation by individual instructor pilots and hence affects inter-rater reliability. An example of how to fine tune the altitude standards could include the following scores:

- a score of 5 should require the student remain within plus or minus 50 feet;
- a 4, plus or minus 100 feet;
- a 3, plus or minus 150 feet;
- a 2, plus or minus 200; and
- a 1 indicates that the student violated the 200 foot limit in either direction, and therefore is unsatisfactory.

The other standards, heading, bank angle and airspeed, could also be redefined to more precisely indicate how skilled the student is, rather than leaving a broad range that is susceptible to loose interpretation. Perhaps by fine-tuning the standards and requiring the instructor pilots to be retrained in these newer, more precisely defined, standards would help to improve inter-rater reliability. Fine-tuning these standards may require further research.

Recommendations for Further Research

This investigation represents a foundational study, meant to lay the groundwork and establish a method to study inter-rater reliability at flight schools.

The first recommendation is to expand the number of samples, the number of raters, or both. It may also be beneficial to utilize other means of measuring inter-rater reliability. Other possible statistical techniques include calculating alpha and rho. In the interest of finding the best analytical method, alpha, rho, and other coefficients should be tested along with the increase in samples and raters until an agreed upon method is derived.

The second recommendation is to choose different flight patterns. One suggestion is to begin testing particular maneuvers such as shallow, medium and steep turns, ascending and descending turns, or constant airspeed climbs. These are just examples, and a future researcher could experiment with particular maneuvers rather than entire patterns. At the same time, one could also consider choosing from a catalog of other instrument patterns, more or less challenging than Pattern D.

It may also be beneficial to collect demographic information on the flight instructors. Differences in scoring may be dependent on experience levels, previous training, and other similar factors.

REFERENCES

Bamford, J.T.M., Gessert, C.E., & Renier, C.M. (2004) Measurement of the severity of rosacea. [Electronic Version]. *Journal of the American Academy Dermatology, 51*(5), 697-703.

Bann, S., Davis, I.M., Moorthy, K., Munz, Y., Hernandez, J., Khan, M., Datta, V., & Darzi, A. (2005). The Reliability of multiple objective measures of surgery and the role of human performance. [Electronic version]. *The American Journal of Surgery, 189*, 747-752.

Bell, V., Halligan P.W., & Ellis, H.D. (2006). Diagnosing Delusions: A review of inter-rater reliability. [Electronic version]. *Schizophrenia Research, 86*, 76-79.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational Psychological Measurement, 20*(1), 37-46.

DeVellis, R.F. (2005). Inter-Rater Reliability. [Electronic version]. *In Encyclopedia of Social Measurement* (Vol. 2, pp. 317-322). New York: Elsevier Inc.,

Dionne, C.P., Bybee, R.F., & Tomaka, J. (2006). Inter-rater reliability of McKenzie assessment in patients with neck pain. [Electronic version]. *Physiotherapy, 92*, 75-82.

Drake, R., Haddock, G., Terrier, N., Bentall, R., & Lewis, S. (2007). The Psychotic Symptom Rating Scales (PSYRATS): Their usefulness and properties in first episode psychosis. [Electronic version]. *Schizophrenia Research, 89*, 119-122.

Ferri, R., Bruni, O., Miano, S., Smerieri, A., Spruyt, K., & Terzano, M. (2005). Inter-rater reliability of sleep cyclic alternating pattern (CAP) scoring and validation of a new computer-assisted CAP scoring method. [Electronic version]. *Clinical Neurophysiology, 116*, 696-707.

Goodwin, L.D. & Goodwin, W.L. (1985). An Analysis of Statistical Techniques Used in the Journal of Educational Psychology, 1979-1983. [Electronic version]. *Educational Psychologist, 20*(1), 13-21.

Gwet, K. (2002a) Kappa statistic is not satisfactory for assessing the extent of agreement between raters. Retrieved December 15, 2006, from http://www.stataxis.com/files/articles/ kappa_statistic_is_not_satisfactory.pdf.

Gwet, K. (2002b) Cohen's Kappa. Retrieved December 15, 2006, from http://www-class.unl.edu/psycrs/handcomp/hckappa.pdf.

Holey, L.A., & Watson, M.J. (1995) Inter-rater reliability of connective tissue zones recognition. [Electronic version]. *Physiotherapy, 61*(7), 369-372.

Hulsman, R.L., Mollema, E.D., Oort, F.J., Hoos, A.M., & de Haes, J.C.J.M. (2006) Using standardized video cases for assessment of medical communication skills: Reliability of an objective structured video examination by computer. [Electronic version]. *Patient Education and Counseling, 60*, 24-31.

Joo, E.-J., Joo, Y.-H., Hong, J.-P., Hwang, S., Maeng, S.-J., Han J.-H., Yang, B.-H., Lee, Y.-S., & Kim, Y.-S. (2004). Korean Version of the Diagnostic Interview for Genetic Studies: Validity and Reliability. [Electronic version]. *Comprehensive Psychiatry, 45*(3), 225-229.

Kadri, N., Agoub, M., El Gnaoui, S., Mchichi Alami, Kh., Hergueta, T., & Moussaoui, D. (2005). Moroccan colloquial Arabic version of the Mini International Neuropsychiatric Intervire (MINI): qualitative and quantitative validation. [Electronic Version]. *European Psychiatry, 20*, 193-195.

Kaneda, Y., Ohmoria, T., & Fujii, A. (2001). The serotonin syndrome: investigation using the Japanese version of the Serotonin Syndrome Scale. [Electronic version]. *Psychiatry Research, 105*, 135-142.

Kirshner, W.K. (1990) *The Pilot's Manual: Instrument Flying* (4th ed.). Ames, IA: Iowa State Press

Kolaitas, J., Korpa, T., Kolvin, I., & Tsiantis, J. (2003). Letter to the Editor. [Electronic version]. *European Psychiatry, 18,* 374-375.

Kolt, G.S., Brewer, B.W., Pizzari, T., School, A.M.M., & Garrett, N. (2006). The Sport Injury Rehabilitation Adherence Scale: a reliable scale for use in clinical physiotherapy. [Electronic version]. *Physiotherapy 93*(1), 17-22.

Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. [Electronic version]. *Assessing Writing, 9,* 4-26.

Leung, T.K.S. & Tsang H.W.H. (2006). Chinese version of the Assessment of Interpersonal Problem Solving Skills. [Electronic version]. *Psychiatry Research 143,* 189-197.

Lindeman, B., Libkuman, T., King, D., & Kruse B. (2000). Development of an Instrument to Assess Jump-Shooting Form in Basketball. [Electronic version]. *Journal of Sports Behavior. 23*(4), 335-348.

Meyers, B.S., English, J., Gabriele, M., Peasley-Miklus, C., Heo, M., Flint, A.J., Mulsant, B.H., & Rothschild, A.J. (2006). A Delusion Assessment Scale for Psychotic major Depression: Reliability, Validity, and Utility. *Biological Psychiatry, 60,* 136-1342.

Michelson, J.D. (2006). Simulation in Orthopaedic Education: An Overview of Theory and Practice. [Electronic version]. *The Journal of Bone & Joint Surgery. 88-A* (6), 1405-1411.

Monroe-Blum, H., Collins, E., McCleary, L., & Nuttall, S. (1996). The social dysfunction index (SDI) for patients with schizophrenia and related disorders. [Electronic version]. *Schizophrenia Research. 20,* 211-219.

Papavasilou, A.S., Rapidi, C.A., Rizou, C., Petrapoulou, K., & Tzavara, Ch. (2006). Reliability of Greek version Gross Motor Function Classification System. [Electronic version]. *Brain & Development, 29,* 79-82

Penny, J., Johnson, R.L., & Gordon, B. (2000) The effect of rating augmentation on inter-rater reliability: and empirical study of a holistic rubric. [Electronic version]. *Assessing Writing, 7,*143-164.

Raymont, V., Buchanan, A., David, A.S., Hayward, P., Wessley, S., & Hotopf, M. (2006). The inter-rater reliability of mental capacity assessments. [Electronic version]. *Law and Psychiatry, 30,* 112-117

Rittenberger, J.C., Martin, J.R., Kelly, L.J., Roth, R.N., Hostler, D., & Callaway, C.W. (2006). Inter-rater reliability for witnessed collapse and presence of bystander CPR. [Electronic version]. *Resuscitation, 70,* 410-415.

Schmidt, N.B., Salas, D., Bernert, R., & Schatschneider, C. (2005). Diagnosing agoraphobia in the context of panic disorder: examining the effect of the DSM-IV criteria on diagnostic decision-making. [Electronic version]. *Behavior Research and Therapy, 43,* 1219-1229.

Thuile, J., Even, C., Friedman, S., & Guelfi, J.-D. (2005). Inter-rater reliability of the French version of the core index for melancholia. [Electronic version]. *Journal of Effective Disorders, 88,* 193-208.

Trochim, W.M.K. (2001). *The Research Methods Knowledge Base* (2nd ed.). Mason, OH: Thomson

Tural, U., Fidaner, H., Alkin, T. & Bandelow, B. (2002). Assessing the severity of panic disorder and agoraphobia: Validity, reliability and objectivity of the Turkish translation of the Panic and Agoraphobia Scale (P & A). [Electronic version]. *Journal of Anxiety Disorders, 16,* 331-340.

Worster, A., Sardo, A.A., Fernandes C.M.B., Eva, K., & Upadhy, S. (2007). Triage tool inter-rater reliability: a comparison of live versus paper case scenarios. [Electronic version]. *Journal of Emergency Nursing, 33*(4), 319-323.