Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 10 Number 5, June 2005 **ISSN 1531-7714**

The Standard Error of a Proportion for Different Scores and Test Length

David A. Walker, Northern Illinois University

This paper examines Smith's (2003) proposed standard error of a proportion index associated with the idea of reliability as sufficiency of information. A detailed table indexing all of the standard error values affiliated with assessments that range from 5 to 100 items, where students scored as low as 50% correct and 50% incorrect to as high as 95% correct and 5% incorrect, calculated in increments of 1 percentage point, is presented, along with distributional qualities. Examples using this measure for classroom teachers and higher education instructors of assessment are provided.

In a recent issue of *Educational Measurement: Issues and Practice*, Smith (2003) contended that the concept of reliability in classroom assessment should be reexamined through a different framework; one that is "… based on the argument that at a rudimentary level, reliability theory is based on the notion of having enough information to make decisions…" (p. 26). Thus, this reconceptualization of reliability in classroom assessment can be thought of as different from conventional ideas and assumptions affiliated with classical test theory and measurement. That is, Smith argued that research in classroom assessment continues to change and move, in a direction away from classical measurement ideas, more toward a view of assessment as a component within "… the service of instruction and/or learning. Assessments are viewed as tools to inform the teacher about strengths and weaknesses of individual students as well as the class as a whole…" (p. 27).

In considering this view of classroom assessment, of which reliability is a factor, Smith (2003) proposed an error index to accompany his argument of regarding reliability within the classroom as a concept pertaining to sufficiency of information. Building on Smith's conceptualization of reliability, this article will present a detailed table indexing a standard error of the proportion (SEP_1) for an obtained score of correct answers from a multiple choice test ranging from 5 to 100 items. Smith defined this SEP_1 measure as:

> *The score for a student could be represented as the proportion of items answered correctly…. If a given student's response to each item is considered to be independent of his or her response to the other items, and if each item is equally difficult for the student, then a standard error of measurement would simply be the standard error of the proportion of items correct… (p. 30)*

An assumption associated with this idea of reliability would be that the items on a test were a random sample of the population of items that could be administered. The SEP_1 functions under another assumption of equal item difficulty. This is an idealistic assumption in the sense that items are almost never equally difficult, especially in the classroom on teacher-made tests that consist frequently of various items that range from the easy to the difficult level. However, since the SEP_1

method is an approximation under this very improbable assumption, it still has value in the K-12 classroom and higher education classroom assessment course for teacher trainees. Finally, it should be noted that although the SEP_1 is defined for a single student, which is independent of a second student's SEP_1, we can make the assumption that if enough SEPs of minimal error appear on an in-class examination, we can begin to address Smith's reliability notion of sufficiency of information.

The SEP_1 is an approximation by which the proportion of items that a student can answer correctly is estimated by the proportion answered correctly on the sample. Thus, SEP_1 is distributed as a *t* instead of a *z* value, which allows for its use with very small test lengths. Smith's SEP_1 is represented as:

$$SEP_1 = \sqrt{pq / k} \quad (1)$$

where

p = proportion correct
q = proportion incorrect or *q = 1-p*
k = number of items

This paper illustrates the uses of the SEP_1 via an examination of Smith's (2003) contention that … *the accuracy of the measurement for students scoring 70% or better does not increase substantially between a 40-item assessment and a 100-item assessment* (p. 31). Secondly, a detailed table indexing all of the standard error values affiliated with assessments that range from 5 to 100 items, where students scored as low as 50% correct and 50% incorrect to as high as 95% correct and 5% incorrect, calculated in increments of 1 percentage point.

RESULTS

Tables 1A and 1B show the SEP_1 index for examinations consisting of 100 and 40 items, respectively that range from 95% to 5% correct in increments of 5%. From both tables, one can see that as the number of items decreases, the amount of error increases, which is to be expected. Further, the two tables indicate that the SEP_1 is symmetrical around p = .50 and, thus, either half of the data presented in Appendix A can be used. For example, in Table 1A for a 100-item test, a student with 55% correct and 45% incorrect on said examination would obtain a SEP_1 = .0497. Proportionately, if this situation were reversed, where a second student on the same 100-item examination achieved 45% correct and 55% incorrect, the SEP_1 value would still remain identical at .0497. To add to the proposed SEP_1 index in terms of its properties, Figure 1 is shown as a scatterplot of the plotted SEP_1 values for 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100 item assessments from 0% proportion correct to 100% correct, in increments of 10%. From Figure 1, we can see this index's symmetrical properties as well.

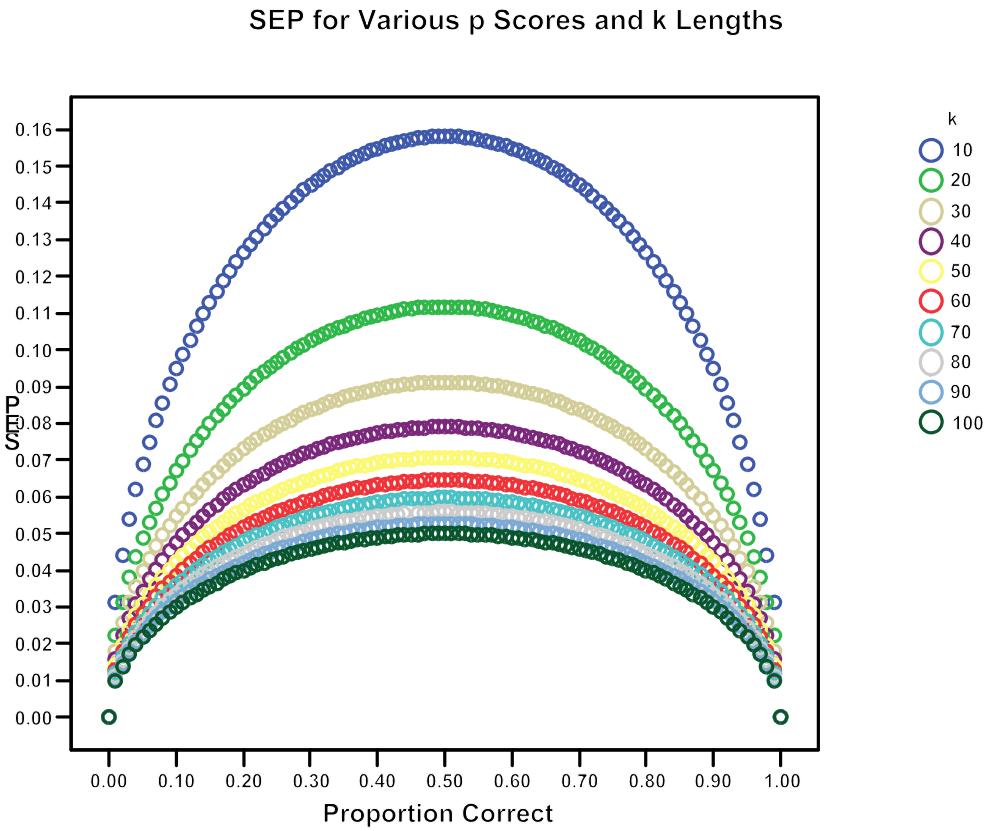
Table 1A. SEP_1 Index for 100 Items Ranging from 95% to 5% Correct

Proportion of Correct Items	SEP_1		Proportion of Correct Items	SEP_1
95%	.0218		5%	.0218
90%	.0300		10%	.0300
85%	.0357		15%	.0357
80%	.0400		20%	.0400
75%	.0433		25%	.0433
70%	.0458		30%	.0458
65%	.0477		35%	.0477
60%	.0490		40%	.0490
55%	.0497		45%	.0497
50%	.0500		50%	.0500

Table 1B. SEP_1 Index for 40 Items Ranging from 95% to 5% Correct

Proportion of Correct Items	SEP_1		Proportion of Correct Items	SEP_1
95%	.0345		5%	.0345
90%	.0474		10%	.0474
85%	.0565		15%	.0565
80%	.0632		20%	.0632
75%	.0685		25%	.0685
70%	.0725		30%	.0725
65%	.0754		35%	.0754
60%	.0775		40%	.0775
55%	.0787		45%	.0787
50%	.0791		50%	.0791

Figure 1. Scatterplot of SEP_1 10 to 100 Items and 0% to 100% Proportion Correct



Appendix A presents a detailed table that the user can apply to examine quickly the standard error of a student's score on an assessment that ranged from 5 to 100 items in regard to their proportion of correct to incorrect items varying between the typical classroom extents of 50% to 95%. With this standard error information, the classroom teacher may begin to make decisions based on evidence concerning how much data are enough with respect to a sufficiency of information (i.e., reliability) to answer, "'Do I have enough information here to make a reasonable decision about this student with

regard to this domain of information?"' (Smith, 2003, p. 30). It should be noted that when regarding this table, certain test lengths, especially for the shorter lengths, the table's proportions correct are not very feasible. For example, it is nearly impossible on a 5-item quiz to have a 94% correct rate that is very meaningful. However, this information was placed into the table to look at the full range of the proposed SEP_1's properties.

Appendix B contains syntax to create a detailed table of the SEP_1 values affiliated with a student's score on an assessment ranging from 5 to 100 items and assuming the student received between 50% correct and 95% correct on the assessment. The syntax in Appendix B will provide the user with standard errors for a specified percentage correct/percentage incorrect for an assessment ranging from 5 to 100 items. To change the syntax, the user can type in the data area their desired percentage correct to create numerous SEP_1 indices for various classroom situations.

CLASSROOM USAGE

Appendix A presents a detailed table containing all of the possible SEP_1 values for assessments comprised of between 5 and 100 items, where a student may have received a score ranging from a low of 50% correct to a high of 95% correct. In using this table, a teacher or instructor of classroom assessment would determine the number of items on an examination or quiz (i.e., k), which is the left column of the table ranging from 5 to 100. Then, the teacher or instructor would identify the percentage of correct responses for a student (i.e., p), which is the top row of the table ranging from .95 to .50. Finally, the user, via an intersection of the k and p data, would find within the table the SEP_1.

Classroom Teacher Examples

Suppose that the parent of a K-12 student wanted to know how much error was on their child's in-class, multiple-choice examination and also the status of their child in comparison to a standard of performance established for that particular exam, a classroom teacher could use this table in the following manner. Initially, the teacher would need to establish a classroom-wide performance standard for the examination of, for instance, a 60% correct response rate (k = 50, p = .60, SEP_1 = .0693) as the criterion for passing. Knowing that the number of items on the test was 50, the teacher would identify the proportion of correct answers for the particular student in question, which was .82 or a raw score of a 41. Intersecting 50 and .82, this specific student had a SEP_1 of .0543 or nearly 5%. Also, confidence intervals expressed as two standard errors around this student's score of 41 indicates that their true test score would lie between (40.8914, 41.1086) or .0543 x 2 = 41 ± .1086. Thus, the teacher could relay to the parent of this student that on this particular in-class examination, we could be confident that the student's score of 82% or a 41 reflected a very good comprehension of the domain being measured with a little over 5% error. However, what if another student on the same in-class examination had a .64 proportion of correct answers or a raw score of a 32, which meant that they had a SEP_1 of .0679, is this acceptable error rate? Confidence intervals of two standard errors around the second student's score of 32 specifies that their true test score would lie between (31.8642, 32.1358) or .0679 x 2 = 32 ± .1358. Again, we could be fairly confident that the student's score of 64% reflected an acceptable level of comprehension of the domain being measured with almost 7% error.

College or University Instructor of Assessment Example

For a college or university instructor of classroom assessment, these examples could demonstrate to teacher trainees that the K-12 classroom teacher could begin to respond positively, for both students, to Smith's (2003, p. 30) original query pertaining to the concept of reliability within the classroom, or "'Do I have enough information here to make a reasonable decision about this student with regard to this domain of information?"' Indeed, as a teacher attempting to make both a student learning decision and an instructional delivery assessment, there appears to be enough information, with minimal error pertaining to the items measured on the examination, to decide that these particular students had acceptable levels of performance on the domain covered, with the first student out-performing the second. Additionally, if there were enough students whose performance on this 50-question examination surpassed the standard

set for it at 60% with an SEP_1 = .0693, we could be fairly confident that coverage of the learning domain was met for many of the students, as well as the instructional intent for the classroom teacher, and, therefore, fewer pieces of information would be needed to satisfy the reliability criterion of sufficiency of information in this case. Theoretically, the classroom teacher could make the decision to move beyond the domain represented in the examination via an understanding that a sufficient amount of the students' learning transpired at an acceptable level along with a desired instructional process.

Finally it should be reiterated that the SEP_1 functions under the assumption of equal item difficulty, which coupled with teacher-made tests is almost never met because teachers often write items on a single test that range from easy to difficult to measure various cognitive abilities for an array of student aptitudes. Nonetheless, the SEP_1 does have merit as an index for considering reliability within the classroom as a concept pertaining to sufficiency of information.

REFERENCE

Smith, J. K. (2003). Reconsidering reliability in classroom assessment and grading. *Educational Measurement: Issues and Practice, 22*(4), 26-33.

APPENDIX A

Table Indexing SEP_1 for Different Scores and Test Length (.81 to .95)

k|p	.95	.94	.93	.92	.91	.90	.89	.88	.87	.86	.85	.84	.83	.82	.81
5	.0975	.1062	.1141	.1213	.1280	.1342	.1399	.1453	.1504	.1552	.1597	.1640	.1680	.1718	.1754
10	.0689	.0751	.0807	.0858	.0905	.0949	.0989	.1028	.1063	.1097	.1129	.1159	.1188	.1215	.1241
15	.0563	.0613	.0659	.0700	.0739	.0775	.0808	.0839	.0868	.0896	.0922	.0947	.0970	.0992	.1013
20	.0487	.0531	.0571	.0607	.0640	.0671	.0700	.0727	.0752	.0776	.0798	.0820	.0840	.0859	.0877
25	.0436	.0475	.0510	.0543	.0572	.0600	.0626	.0650	.0673	.0694	.0714	.0733	.0751	.0768	.0785
30	.0398	.0434	.0466	.0495	.0522	.0548	.0571	.0593	.0614	.0634	.0652	.0669	.0686	.0701	.0716
35	.0368	.0401	.0431	.0459	.0484	.0507	.0529	.0549	.0568	.0587	.0604	.0620	.0635	.0649	.0663
40	.0345	.0375	.0403	.0429	.0452	.0474	.0495	.0514	.0532	.0549	.0565	.0580	.0594	.0607	.0620
45	.0325	.0354	.0380	.0404	.0427	.0447	.0466	.0484	.0501	.0517	.0532	.0547	.0560	.0573	.0585
50	.0308	.0336	.0361	.0384	.0405	.0424	.0442	.0460	.0476	.0491	.0505	.0518	.0531	.0543	.0555
55	.0294	.0320	.0344	.0366	.0386	.0405	.0422	.0438	.0453	.0468	.0481	.0494	.0507	.0518	.0529
60	.0281	.0307	.0329	.0350	.0369	.0387	.0404	.0420	.0434	.0448	.0461	.0473	.0485	.0496	.0506
65	.0270	.0295	.0316	.0336	.0355	.0372	.0388	.0403	.0417	.0430	.0443	.0455	.0466	.0477	.0487
70	.0260	.0284	.0305	.0324	.0342	.0359	.0374	.0388	.0402	.0415	.0427	.0438	.0449	.0459	.0469
75	.0252	.0274	.0295	.0313	.0330	.0346	.0361	.0375	.0388	.0401	.0412	.0423	.0434	.0444	.0453
80	.0244	.0266	.0285	.0303	.0320	.0335	.0350	.0363	.0376	.0388	.0399	.0410	.0420	.0430	.0439
85	.0236	.0258	.0277	.0294	.0310	.0325	.0339	.0352	.0365	.0376	.0387	.0398	.0407	.0417	.0426
90	.0230	.0250	.0269	.0286	.0302	.0316	.0330	.0343	.0354	.0366	.0376	.0386	.0396	.0405	.0414
95	.0224	.0244	.0262	.0278	.0294	.0308	.0321	.0333	.0345	.0356	.0366	.0376	.0385	.0394	.0402
100	.0218	.0237	.0255	.0271	.0286	.0300	.0313	.0325	.0336	.0347	.0357	.0367	.0376	.0384	.0392

Note: k = number of items, p = percentage correct, and the tabled numbers represent the SEP_1 values.

Table Indexing SEP_1 for Different Scores and Test Length (.66 to .80)

k|p	.80	.79	.78	.77	.76	.75	.74	.73	.72	.71	.70	.69	.68	.67	.66
5	.1789	.1822	.1853	.1882	.1910	.1936	.1962	.1985	.2008	.2029	.2049	.2068	.2086	.2103	.2118
10	.1265	.1288	.1310	.1331	.1351	.1369	.1387	.1404	.1420	.1435	.1449	.1463	.1475	.1487	.1498
15	.1033	.1052	.1070	.1087	.1103	.1118	.1133	.1146	.1159	.1172	.1183	.1194	.1204	.1214	.1223
20	.0894	.0911	.0926	.0941	.0955	.0968	.0981	.0993	.1004	.1015	.1025	.1034	.1043	.1051	.1059
25	.0800	.0815	.0828	.0842	.0854	.0866	.0877	.0888	.0898	.0908	.0917	.0925	.0933	.0940	.0947
30	.0730	.0744	.0756	.0768	.0780	.0791	.0801	.0811	.0820	.0828	.0837	.0844	.0852	.0858	.0865
35	.0676	.0688	.0700	.0711	.0722	.0732	.0741	.0750	.0759	.0767	.0775	.0782	.0788	.0795	.0801
40	.0632	.0644	.0655	.0665	.0675	.0685	.0694	.0702	.0710	.0717	.0725	.0731	.0738	.0743	.0749
45	.0596	.0607	.0618	.0627	.0637	.0645	.0654	.0662	.0669	.0676	.0683	.0689	.0695	.0701	.0706
50	.0566	.0576	.0586	.0595	.0604	.0612	.0620	.0628	.0635	.0642	.0648	.0654	.0660	.0665	.0670
55	.0539	.0549	.0559	.0567	.0576	.0584	.0591	.0599	.0605	.0612	.0618	.0624	.0629	.0634	.0639
60	.0516	.0526	.0535	.0543	.0551	.0559	.0566	.0573	.0580	.0586	.0592	.0597	.0602	.0607	.0612

Table Indexing SEP_1 for Different Scores and Test Length (.66 to .80)

k|p	.80	.79	.78	.77	.76	.75	.74	.73	.72	.71	.70	.69	.68	.67	.66
65	.0496	.0505	.0514	.0522	.0530	.0537	.0544	.0551	.0557	.0563	.0568	.0574	.0579	.0583	.0588
70	.0478	.0487	.0495	.0503	.0510	.0518	.0524	.0531	.0537	.0542	.0548	.0553	.0558	.0562	.0566
75	.0462	.0470	.0478	.0486	.0493	.0500	.0506	.0513	.0518	.0524	.0529	.0534	.0539	.0543	.0547
80	.0447	.0455	.0463	.0471	.0477	.0484	.0490	.0496	.0502	.0507	.0512	.0517	.0522	.0526	.0530
85	.0434	.0442	.0449	.0456	.0463	.0470	.0476	.0482	.0487	.0492	.0497	.0502	.0506	.0510	.0514
90	.0422	.0429	.0437	.0444	.0450	.0456	.0462	.0468	.0473	.0478	.0483	.0488	.0492	.0496	.0499
95	.0410	.0418	.0425	.0432	.0438	.0444	.0450	.0455	.0461	.0466	.0470	.0475	.0479	.0482	.0486
100	.0400	.0407	.0414	.0421	.0427	.0433	.0439	.0444	.0449	.0454	.0458	.0462	.0466	.0470	.0474

Note: *k* = number of items, *p* = percentage correct, and the tabled numbers represent the SEP_1 values.

Table Indexing SEP_1 for Different Scores and Test Length (.65 to .50)

k|p	.65	.64	.63	.62	.61	.60	.59	.58	.57	.56	.55	.54	.53	.52	.51	.50
5	.2133	.2147	.2159	.2171	.2181	.2191	.2200	.2207	.2214	.2220	.2225	.2229	.2232	.2234	.2236	.2236
10	.1508	.1518	.1527	.1535	.1542	.1549	.1555	.1561	.1566	.1570	.1573	.1576	.1578	.1580	.1581	.1581
15	.1232	.1239	.1247	.1253	.1259	.1265	.1270	.1274	.1278	.1282	.1285	.1287	.1289	.1290	.1291	.1291
20	.1067	.1073	.1080	.0787	.1091	.1095	.1100	.1104	.1107	.1110	.1112	.1114	.1116	.1117	.1118	.1118
25	.0954	.0960	.0966	.0787	.0975	.0980	.0984	.0987	.0990	.0993	.0995	.0997	.0998	.0999	.1000	.1000
30	.0871	.0876	.0881	.0886	.0891	.0894	.0898	.0901	.0904	.0906	.0908	.0910	.0911	.0912	.0913	.0913
35	.0806	.0811	.0816	.0820	.0824	.0828	.0831	.0834	.0837	.0839	.0841	.0842	.0844	.0844	.0845	.0845
40	.0754	.0759	.0763	.0767	.0771	.0775	.0778	.0780	.0783	.0785	.0787	.0788	.0789	.0790	.0790	.0791
45	.0711	.0716	.0720	.0724	.0727	.0730	.0733	.0736	.0738	.0740	.0742	.0743	.0744	.0745	.0745	.0745
50	.0675	.0679	.0683	.0686	.0690	.0693	.0696	.0698	.0700	.0702	.0704	.0705	.0706	.0707	.0707	.0707
55	.0643	.0647	.0651	.0654	.0658	.0661	.0663	.0666	.0668	.0669	.0671	.0672	.0673	.0674	.0674	.0674
60	.0616	.0620	.0623	.0627	.0630	.0632	.0635	.0637	.0639	.0641	.0642	.0643	.0644	.0645	.0645	.0645
65	.0592	.0595	.0599	.0602	.0605	.0608	.0610	.0612	.0614	.0616	.0617	.0618	.0619	.0620	.0620	.0620
70	.0570	.0574	.0577	.0580	.0583	.0586	.0588	.0590	.0592	.0593	.0595	.0596	.0597	.0597	.0597	.0598
75	.0551	.0554	.0557	.0560	.0563	.0566	.0568	.0570	.0572	.0573	.0574	.0575	.0576	.0577	.0577	.0577
80	.0533	.0537	.0540	.0543	.0545	.0548	.0550	.0552	.0554	.0555	.0556	.0557	.0558	.0559	.0559	.0559
85	.0517	.0521	.0524	.0526	.0529	.0531	.0533	.0535	.0537	.0538	.0540	.0541	.0541	.0542	.0542	.0542
90	.0503	.0506	.0509	.0512	.0514	.0516	.0518	.0520	.0522	.0523	.0524	.0525	.0526	.0527	.0527	.0527
95	.0489	.0492	.0495	.0498	.0500	.0503	.0505	.0506	.0508	.0509	.0510	.0511	.0512	.0513	.0513	.0513
100	.0477	.0480	.0483	.0485	.0488	.0490	.0492	.0494	.0495	.0496	.0497	.0498	.0499	.0500	.0500	.0500

Note: *k* = number of items, *p* = percentage correct, and the tabled numbers represent the SEP_1 values.

APPENDIX B

SPSS Syntax for SEP_1 calculation

```
**************************************************************************
Author: David A. Walker, dawalker@niu.edu
        Northern Illinois University
**************************************************************************.
INPUT PROGRAM.
  LOOP #CASE = 100 to 5 BY -5.
    COMPUTE k = #CASE.
**************************************************************************
NOTE: Change the LOOP to reflect how you would like to see the data
presented. For example, if you wanted the SEP value for 60 to 80 items,
change the LOOP to #CASE = 60 TO 80 or 80 TO 60 BY -1.
**************************************************************************.
    END CASE.
  END LOOP.
  END FILE.
END INPUT PROGRAM.
EXECUTE.
COMPUTE p = .80.
**************************************************************************
NOTE: Change the proportion correct (p) above to reflect your situation
**************************************************************************.
COMPUTE q = 1-p.
COMPUTE SEP = SQRT((p * q) / k).
EXECUTE.
* FINAL REPORTS *.
FORMAT SEP  (f9.4) p q (f9.2) k (f8.0).
VARIABLE LABELS p 'Proportion Correct'/
                q 'Proportion Incorrect'/
                k 'Number of Items'/
                 SEP 'Standard Error of a Proportion' /.
REPORT FORMAT=LIST AUTOMATIC ALIGN(CENTER)
  /VARIABLES= k p q SEP
  /TITLE "SEP for an Individual Test Score".
```

Citation

Walker, David A. (2005). The Standard Error of a Proportion for Different Scores and Test Length. *Practical Assessment Research & Evaluation*, 10(5). Available online: http://pareonline.net/getvn.asp?v=10&n=5

Author

David A. Walker, Ph.D.
Northern Illinois University
ETRA Department
101J Gabel
DeKalb, IL 60115

dawalker@niu.edu