Weighting of items in a tutorial performance evaluation instrument: Statistical analysis and results

Authors:
Melanie L. Lack[1]
Judith C. Bruce[1]
Piet J. Becker[2]

Affiliations:
[1]Department of Nursing Education, University of the Witwatersrand, South Africa

[2]Department of Biostatistics, University of the Witwatersrand, South Africa

Correspondence to:
Judith Bruce

e-mail:
judith.bruce@wits.ac.za

Postal address:
Department of Nursing Education, Faculty of Health Sciences, 7 York Road, Parktown, Johannesburg 2193, South Africa

Keywords:
weighting; tutorial performance; problem-based learning; evaluation instrument; statistical analysis; Delphi technique

Dates:
Received: 15 Oct. 2007
Accepted: 27 Oct. 2008
Published: 21 May 2009

How to cite this article:
Lack, M.L., Bruce, J. & Becker, J., 2009, 'Weighting of items in a tutorial performance evaluation instrument: Statistical analysis and results', *Health SA Gesondheid 14*(1), Art. #408, 5 pages. DOI: 10.4102/hsag.v14i1.408

This article is available at:
http://www.hsag.co.za

ABSTRACT

Weighting of items in an evaluation instrument contributes to more meaningful and valid interpretations of student performance in respect of each learning outcome or item being assessed. It follows that the validity of instruments is important for meaningful inferences about students' learning performance, including their performance in tutorial groups. The Delphi technique was used to elicit experts' subjective judgement of the content validity of items in the tutorial performance evaluation instrument in rounds one and two. A sample of eight experts (n = 8) was selected by purposive, maximum variation sampling.

In round three Delphi a weighted score was determined for each of the instrument items, subitems and Likert scale points through pairwise comparison by the experts. Mathematical modelling of experts' weighting comparisons, recorded on visual analogue scales, resulted in proportional weights for each item; these weights are expressed as a percentage.

The final instrument comprised weighted items measured on a rating scale with points that are not equidistant. A computerised tutorial performance evaluator (TPE) was developed for accurate, economical and efficient calculation of student scores. The purpose of this article is to report on the statistical analysis and results of the weighting of items in an instrument to assess and evaluate baccalaureate nursing students' performance in problem-based learning tutorials.

OPSOMMING

Die waardebepaling van items in 'n evalueringsinstrument dra by tot 'n meer betekenisvolle en geldige interpretasie van studente se vermoë ten opsigte van elke leeruitkoms of item wat geassesseer word. Hieruit volg dit dat die geldigheid van instrumente belangrik is vir betekenisvolleafleidings betreffende studente se leervermoë, insluitend hulle prestasie in leergroepe. Die Delphi-tegniek is gebruik in rondtes een en twee om kundiges se subjektiewe oordeel oor die inhoudsgeldigheid van items in die leergroepgedragsevalueringsinstrument te bekom. 'n Steekproef van agt kundiges (n = 8) is deur doelgerigte, maksimale-variasiesteekproeftrekking gekies.

Die kundiges het in rondte drie van die Delphi-tegniektoepassing deur middel van gepaarde vergelyking 'n waarde bepaal vir elk van die instrumentitems, sub-items en Likertskaalpunte. Die wiskundige modellering van kundiges se waardebepalingsvergelykings, aangeteken op visuele analogiese skale, het proporsionele gewigte vir elke item tot gevolg gehad. Hierdie gewigte word persentasiegewys voorgestel. Die finale instrument het bestaan uit items wat gemeet word teen 'n graderingskaal met punte wat nie op 'n gelyke afstand van mekaar is nie. 'n Rekenaargebaseerde leergroepgedragsevalueringsinstrument is ontwikkel vir die akkurate, ekonomiese en doeltreffende berekening van studente se punte.

Die doel van die artikel is om verslag te lewer oor die statistiese analise en resultate van die waardebepaling van items in 'n instrument om baccalaureate-verpleegkundestudente se prestasie in probleemgebaseerde leergroepe te meet en te evalueer.

INTRODUCTION

Weighting of items in any assessment tool is difficult in the absence of scientific evidence and is a perennial challenge facing nurse educators. Assessing and evaluating student performance require scores that are accurate, valid and free from bias for meaningful interpretations and conclusions about students' learning performance. It follows that instruments used to assess performance in any learning environment must produce results from which valid and unbiased inferences can be drawn. In problem-based learning (PBL) students' group skills or group behaviours in small group tutorials are important indicators of learning. Tutorial behaviours generally assessed include self-directed learning, communication, small group interaction, reasoning and autonomy (Niemenin, Saure & Lonka 2006:65; Rideout 1999:216). The content of this paper is derived from a study that sought to determine, in part, the validity of a tutorial performance evaluation instrument to evaluate group skills in a PBL context; the validation processes are described in a previous article.

The study institution, a university department of nursing, uses a PBL curriculum for the preparation of its undergraduate baccalaureate nursing students. One of the main features associated with this learning approach is the tutorial, where small groups of students discuss and analyse clinical and community problems. These small group discussions are facilitated by a nurse educator (called a facilitator) whose primary role is to foster cooperation, stimulate thinking, promote enquiry and facilitate problem solving through the search for, application and integration of knowledge.

These processes require the individual learner to possess or, in the longer term, to develop a range of skills necessary for effective group functioning. Appropriate communication skills are required together with having to learn the 'new language' associated with health sciences. Growth of the

Journal of Interdisciplinary Health Sciences
Article #408

student and of the group is encouraged and promoted to improve self-confidence and to motivate the individual to become a self-directed learner. Good problem-solving skills together with critical thinking skills are paramount and, if not present, need to be developed. These skills should assist the individual to arrive at a logical conclusion when analysing and seeking solutions to a problem. Most of these skills are abstract and difficult to measure and, in a learning programme, must be assessed to determine and provide feedback on students' learning progress. Originally, a 36-item instrument was developed to assess the development of group skills within PBL tutorials without evidence of its validity to assess student learning in tutorial groups. After determining its content validity this instrument was subjected to processes and statistical procedures to determine the value of each item in relation to other items in the instrument, ultimately to assign individual item weights. This paper reports on the statistical analysis and procedures for the weighting of items in a tutorial performance evaluation instrument.

Definition of key concepts

The following definitions applied to this study:

- Validity refers to the appropriateness, meaningfulness and usefulness of inferences drawn from instrument scores.
- Construct refers to a main variable in an evaluation instrument within which measurable criteria or subitems are located. Construct is used interchangeably with main item in this study.
- Weighting refers to the value assigned to an item and subitem based on its importance in a set of items in an evaluation instrument to enhance its internal structure.
- Tutorial performance refers to student behaviour in a small-group learning context, which facilitates individual learning, group learning and team work (Rideout, 1999:233).

The tutorial performance evaluation instrument

Historically, students' performance in PBL tutorials had been assessed using an original tutorial performance evaluation instrument comprising seven main items or constructs and 36 subitems. These items are equally weighted and rated against an eight-point Likert scale, with equidistant points. After content validity had been determined in an earlier part of the study this evaluation instrument was sent to experts for them to estimate the value of each item, subitem and the Likert scale through pairwise comparisons. Statistical procedures applied to the experts' estimated values resulted in relative weights being assigned to each item. In preparing the instrument for weighting procedures the seven main items (constructs) were labelled A–G and the subitems inside each construct as 1, 2, and so forth, according to the number of subitems present.

METHODS

The purpose of this part of the research was to determine the relative weights of items in a tutorial performance evaluation instrument with a view to enhancing its internal structure and thus its validity.

Design and sample

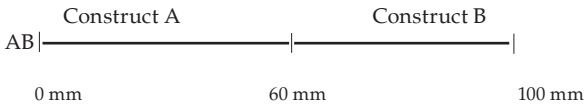
Within an overarching quantitative design the Delphi technique was used to elicit experts' subjective judgement (Crawford & Williams 1985:3; Miranda 2001:87) regarding the validity of unweighted items in the tutorial performance evaluation tool. To this end, experts estimated the relative weight of each item in relation to another by a process of pairwise comparison. A sample of eight experts (n = 8) from two South African universities was selected by purposive, maximum variation sampling (Patton 2002:234).

Data collection procedure

Three rounds of the Delphi technique were used for collecting the data. In rounds one and two content validity was established; the seven main items remained and the subitems were reduced from 36 to 34. The Likert scale was reduced from eight to four points with descriptors as determined and verified by the experts. Once content validity had been established the pre-eminent tutorial performance evaluator (TPE) had the distinct disadvantage of all items carrying the same weight. In round three of the Delphi technique, a weighted score incorporating the weights for each of the constructs or main items (WC), subitems (WI) and Likert scale points (WL) was determined. This was achieved by the experts' subjective judgement, through pairwise comparison (David 1963:9) of the relative value of each of WC, WI and WL. These subjective ratings were recorded on visual analogue scales. The procedure for weighting through pairwise comparison was as follows:

Firstly, each expert was asked to rate the value or importance of one item relative to another for each possible pair of main items (constructs) on a 100-mm visual analogue scale; in other words, experts were asked to give a subjective judgement of the weight of one construct against a second construct in a pairwise fashion until all constructs had been rated.

Example:

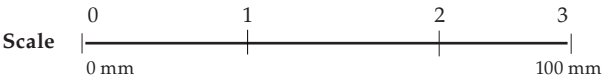


Construct A = 60 mm; therefore, Construct B = 40 mm.

Secondly, and in a similar way, experts were required to judge and rate the relative weights of pairs of subitems within each construct until all possible combinations of subitems had been rated. These two steps in the data collection procedure produced 100 visual analogue scales per expert: 21 for main items and 79 for subitems. This step resulted in a total of 800 units of analysis.

Thirdly, experts were required to conduct a similar weighting assessment of the four-point Likert scale by marking the distance between the four points (0 to 3) on a 100-mm visual analogue scale.

Example:



This step in the data collection procedure resulted in an additional eight units of analysis (n = 8).

After pairwise comparisons had been completed all visual analogue scales (n = 808) were returned to the researcher. Visual analogue scales from pairwise comparison of main and subitems (n = 800) were accurately measured for the distance between 0 mm and the experts' marks. These measurements were entered onto an Excel spreadsheet for statistical analysis by a resident statistician. Visual analogue scales (n = 8) for determining the Likert scale weighting were accurately measured for the distance between each point of the scale. Similarly, all measures were entered onto a second Excel spreadsheet for analysis.

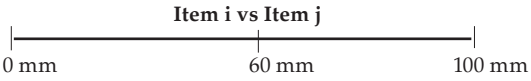
Statistical analysis

A linear regression model was fitted to the logarithms of the relative weights obtained during pairwise comparison of main items (WC) and subitems (WI). Regression coefficients were exponentiated and standardised to add up to 100%. Each subitem within a main item (construct) was weighted and expressed as a percentage, the sum of which equals 100%. Each construct now had its own unique weighting represented as a

percentage, the sum of which equals 100%. In estimating item weights a mathematical logarithm was used.

Example:

a_{ijk} is the ratio between w_i and w_j



$a_{ijk} = (w_i / w_j) = 60/40$

In the above example the ratio between Item *i*/ Item *j* is 1.5/1.0.

RESULTS

Statistical analysis and mathematical modelling of instrument items now produced ordinal scale data for all main items (constructs) (n = 7) and subitems (n = 34) with points that are not equidistant. Thereafter, each main and subitem had its own proportional value. Once the subitems in a particular construct or main item had been calculated to an overall percentage, the latter was further calculated in accordance with the percentage specific to that construct.

Example:

(WI)(WC) / 100%

Results of mathematical modelling of experts' weighting of the four-point rating scale (0–3) were as follows: Assuming that 0 = 0% and 3 = 100% a rating of 1 was weighted as 28% and a rating of 2 as 69%; thus the Likert scale points were no longer equidistant. The results of item weighting are shown in Figure 1. Every tick-box on the instrument, now called the TPE, has a unique weight equal to the product of weight of main item/ construct (WC), weight of subitem (WI) and weight of rating scale (WL), in other words. (WC)(WI)(WL)/100%.

Constructing and implementing the TPE

Each of the seven (7) main items together with the specific subitems relative to that main item (construct) had their own given percentage. Being assigned their relative weights the main items were ranked from the highest to the lowest

MAIN ITEMS/CONSTRUCTS	Weight	RATING SCALE 0	1	2	3
			28%	69%	100%
PROBLEM SOLVING SKILLS	17.38				
1. Identifies possible solutions to the problem			6.36	15.69	22.74
2. Helps select strategy to solve the problem			6.02	14.83	21.50
3. Manages to identify the problem			5.78	14.25	20.66
4. Implements a solution to the problem			5.09	12.54	18.18
5. Discusses the best solution to the problem			4.73	11.67	16.92
CONTRIBUTIONS	16.65				
1. Takes into account other disciplines when appropriate			6.80	16.78	24.32
2. Is comprehensive depending on subject learnt			5.91	14.57	21.13
3. Integrates ethics into discussion			5.65	13.93	20.20
4. Integrates legislation into discussion when appropriate			5.43	13.39	19.41
5. Integrates health service principles			4.18	10.30	14.94
COMMUNICATION	14.58				
1. Demonstrates verbal skills appropriate to the situation			17.86	44.02	63.81
2. Demonstrates non-verbal skills appropriate to situation			10.13	24.97	36.19
CRITICAL THINKING SKILLS	13.44				
1. Identifies a problem, question or issue			7.94	19.58	28.38
2. Uses evidence to support an argument or position			7.43	18.32	26.56
3. Suggests and proposes alternatives			7.04	17.36	25.17
4. Verbally analyses the problem, question or issue			5.56	13.71	19.88
LEARNING SKILLS	13.21				
1. Should be able to use more than one resource			9.50	23.42	33.95
2. Demonstrates ability to understand concepts and theories			9.33	23.01	33.35
3. Is able to interpret a learning objective			9.15	22.56	32.70
PERSONAL GROWTH	13.16				
1. Displays active listening			3.79	9.36	13.57
2. Presents constructive interventions and feedback			3.78	9.32	13.51
3. Shows ability to brainstorm			3.77	9.29	13.47
4. Acknowledges contributions from group members			3.65	9.01	13.06
5. Offers encouragement and support to group members			3.56	8.79	12.74
6. Able to summarise discussion			3.35	8.27	11.99
7. Willing to work with group members			3.27	8.06	11.69
8. Assists and manages ground rules for the group			2.79	6.87	9.97
LEADERSHIP	11.58				
1. Offers facts, suggestions, opinions to group members			4.54	11.20	16.24
2. Identifies learning issues			4.49	11.08	16.06
3. Initiates the undertaking of tasks			4.35	10.72	15.55
4. Shows a caring attitude			3.99	9.84	14.27
5. Drives the process forward			3.95	9.75	14.14
6. Takes decisions			3.36	8.30	12.03
7. Shows assertiveness			3.27	8.07	11.71

FIGURE 1
Relative weights of Main-items and sub-items

TABLE 1
Relative weights of Main items expressed as percentage

MAIN ITEMS/CONSTRUCTS	PERCENTAGE
A. Problem solving skills	17.38
B. Contributions	16.65
C. Communication	14.58
D. Critical thinking skills	13.44
E. Learning skills	13.21
F. Personal growth	13.16
G. Leadership	11.58
TOTAL	100%

TABLE 2
TPE Rating scale descriptors

SCORE	DESCRIPTORS
0	Lack of ability
1	Limited ability
2	Improved ability
3	Good ability

percentage (see Table 1). Once completed, the subitems within each main item were also ranked from the highest to the lowest percentage.

Using the TPE to score students' learning performance requires the nurse educator/facilitator to calculate the (WC)(WI)(WL)/100% for each subitem. This would be time consuming if done manually. Additionally, error may occur rendering the composite score inaccurate, unreliable and invalid. Practicality and ease of use of the TPE could not be overlooked for successful implementation. A computer-based TPE (see Figure 2) was designed to allow for these calculations to be done efficiently, accurately and quickly for meaningful interpretation of students' scores.

The TPE is used for formative assessment purposes by both the student (self-assessment) and the nurse educator (facilitator assessment). During this process, which is a paper assessment, each subitem on the TPE is given a rating of 0, 1, 2 or 3 by the individual carrying out the assessment according to descriptors of the rating scale (see Table 2). An agreed-upon rating between the student and facilitator is then entered onto the computerised TPE by clicking in the corresponding box. The calculations are computed automatically by identifying the value of a 0, 1, 2 or 3 rating and converting the rating to the relative percentage. The sum of the percentages is computed to produce the total percentage.

DISCUSSION

Emerging thoughts on validity suggest that validity is not a property of an evaluation instrument but of instrument scores and interpretation of scores (Beckman, Cook & Mandrekar 2005:1159; Cook & Beckman 2006:166.e7). In this regard validity has become a unitary concept to describe the degree to which a score can be interpreted as representing the activity being measured (Cook & Beckman 2006:166.e8) – in this case, PBL tutorial performance. Sources of validity evidence are many; noteworthy, and for the purpose of this study, is the internal structure of an instrument (Beckman *et al*. 2005:1160). Internal structure refers to the degree to which individual items fit the underlying construct and is usually determined using factor analysis (Beckman *et al*. 2005:1160). Many factors in the instrument itself may threaten its internal structure; equality in weighting or unweighted items has been described as one such factor. Subjective judgement as an alternative to factor analysis has been used in this study to improve the internal structure of a tutorial performance evaluation instrument by way of item weighting. However, the credibility of subjective judgement as method has aroused increasing criticism (Crawford 985:3). As a consequence, quantification of experts' subjective judgements has been posited as a valid and reliable method to assist with weighting and preferential ranking of instrument items. Statistical analysis of subjective data from experts is thus important for valid inferences about student learning: in this instance, learning performance in PBL groups.

Individualised weighting of each item and subitem in the TPE together with the four points on the rating scale provides useful, differentiated information about specific aspects of student learning within PBL groups. Scaling each set of subitems within a construct or main item to a value of 100% allows the facilitator to view the student score for each set of subitems or learning domain on the TPE. Remediation and/or support can be offered to the student in areas where a low percentage has been obtained. Additionally, the hierarchical arrangement of constructs and their subitems enables the facilitator to be selective and to prioritise when giving academic support to a student. Depending on the level or year of study the facilitator and student can initially concentrate on domains that are seen to be of greater importance than others.

CONCLUSION

Statistical procedures applied to quantify experts' pairwise comparison of the relative value of instrument items resulted in the weighting and preferential ranking of items. Paired comparisons for the weighting of items produce more valid estimates than no comparison at all or reliance on experts' intuition. It may be concluded that, based on its internal structure, the computerised TPE has validity by virtue of the relative weight of items obtained during pairwise comparisons. The TPE is accurate, economical and easy to use by both the student and the facilitator; entering ratings onto the computer is a quick process with automatic calculation and conversion of scores. Reducing the rating options from eight points to four points and providing concrete descriptors for each point on the rating scale make item ratings more objective and reliable. The validity of a tutorial performance evaluation instrument thus also enhances the reliability of the processes during which scores are produced.

REFERENCES

Beckman, T.J., Cook, D.A. & Mandrekar, J.N., 2005, 'What is the validity evidence for assessments of clinical teaching?', *Journal of General Internal Medicine* 20, 1159–1164.

Cook, D.A. & Beckman, T.J., 2006, 'Current concepts in validity and reliability for psychometric instruments: Theory and application', *The American Journal of Medicine* 119(2), 166.e7–166.e16.

Crawford, G. & Williams, C., 1985, *The analysis of subjective judgment matrices. A Project Air Force report,* Rand, Santa Monica.

David, H.A., 1963, *The method of paired comparison*, Charles Griffin and Co. Ltd., London.

Miranda, E., 2001, *Improving subjective estimates using paired comparisons,* IEEE Software, Ericsson Research Canada, Mississauga, Ontario.

Niemenin, J., Saure, P. & Lonka, K., 2006, 'On the relationship between group functioning and study success in problem-based learning', *Medical Education* 40, 64–71.

Patton, M.Q., 2002, *Qualitative research and evaluation methods*, Sage, Thousand Oaks.

Rideout, E., 1999, *Transforming nursing education through problem-based learning*, Jones and Bartlett Publishers International, London.

Savin-Baden, M., 2000, *Problem-based learning in higher education: Untold stories*, viewed 7 July 2006, from: www.mcgraw-hill.co.uk/openup/chapter.

Article #408 Journal of Interdisciplinary Health Sciences

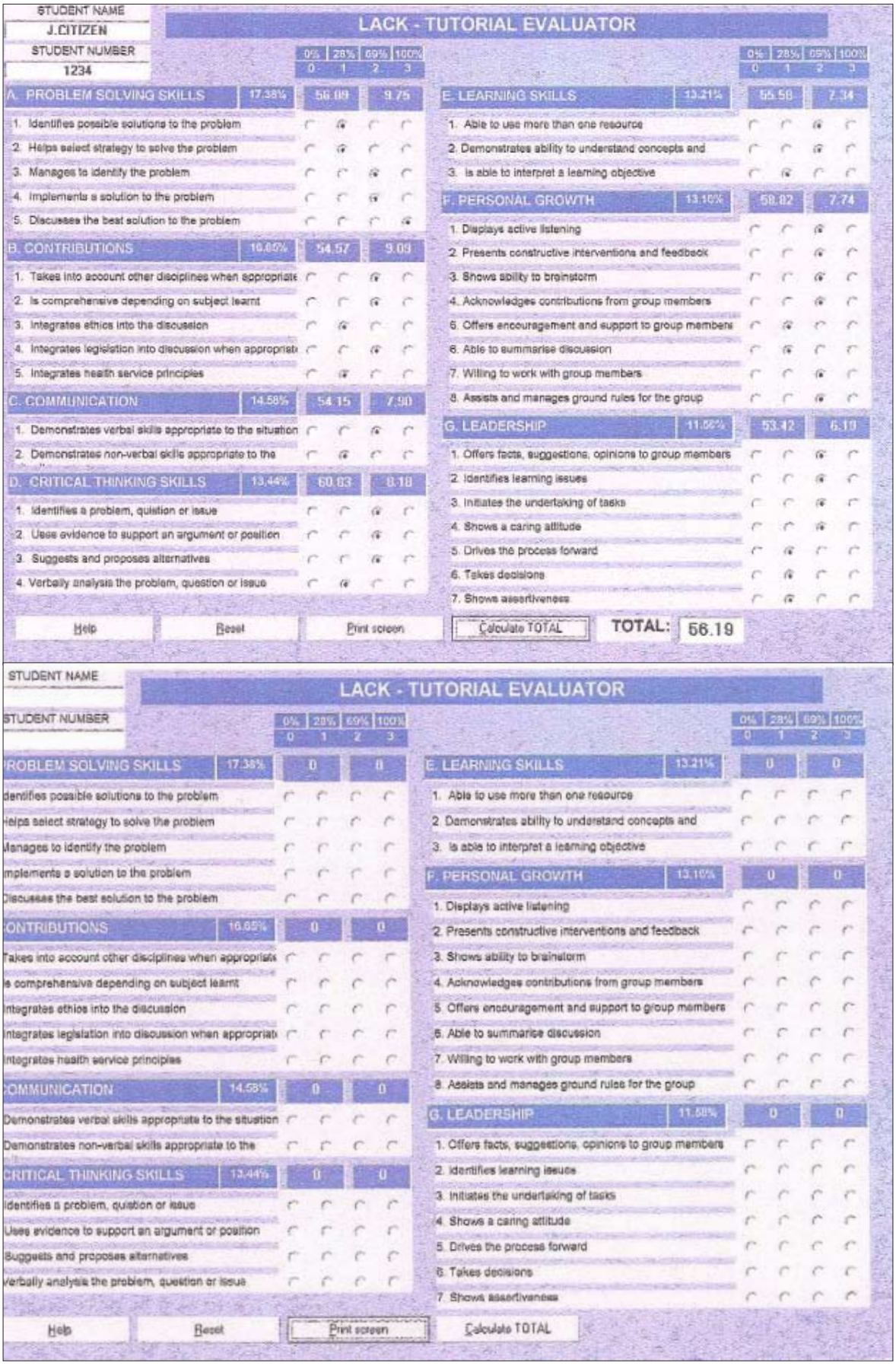

STUDENT NAME
J.CITIZEN
STUDENT NUMBER
1234
LACK - TUTORIAL EVALUATOR
0% 28% 69% 100%
0 1 2 3
A. PROBLEM SOLVING SKILLS 17.38% 56.89 9.75
1. Identifies possible solutions to the problem
2. Helps select strategy to solve the problem
3. Manages to identify the problem
4. Implements a solution to the problem
5. Discusses the best solution to the problem
B. CONTRIBUTIONS 16.65% 54.57 9.09
1. Takes into account other disciplines when appropriate
2. Is comprehensive depending on subject learnt
3. Integrates ethics into the discussion
4. Integrates legislation into discussion when appropriate
5. Integrates health service principles
C. COMMUNICATION 14.58% 54.15 7.90
1. Demonstrates verbal skills appropriate to the situation
2. Demonstrates non-verbal skills appropriate to the
D. CRITICAL THINKING SKILLS 13.44% 60.63 8.18
1. Identifies a problem, quistion or issue
2. Uses evidence to support an argument or position
3. Suggests and proposes alternatives
4. Verbally analysis the problem, question or issue
E. LEARNING SKILLS 13.21% 55.56 7.34
1. Able to use more than one resource
2. Demonstrates ability to understand concepts and
3. Is able to interpret a learning objective
F. PERSONAL GROWTH 13.10% 58.82 7.74
1. Displays active listening
2. Presents constructive interventions and feedback
3. Shows ability to brainstorm
4. Acknowledges contributions from group members
5. Offers encouragement and support to group members
6. Able to summarise discussion
7. Willing to work with group members
8. Assists and manages ground rules for the group
G. LEADERSHIP 11.58% 53.42 6.19
1. Offers facts, suggestions, opinions to group members
2. Identifies learning issues
3. Initiates the undertaking of tasks
4. Shows a caring attitude
5. Drives the process forward
6. Takes decisions
7. Shows assertiveness
Help
Reset
Print screen
Calculate TOTAL
TOTAL: 56.19
STUDENT NAME
STUDENT NUMBER
LACK - TUTORIAL EVALUATOR
Help
Reset
Print screen
Calculate TOTAL

FIGURE 2