Open-source and free software for in-class online surveys and data analyses

Author: **Enrique Frio**

Affiliate (adjunct) faculty, University of the Philippines Open University (UPOU)
and
Adjunct instructor, University of Maryland University College (UMUC)

E-mail: Enrique.Frio@upou.edu.ph

Mailing address: UPOU Headquarters, Los Baños, Laguna, Philippines
Tel.: +63(49) 536-6001 to 06 loc 340
Telefax: +63(49) 536-5484

Abstract

Data analysis of surveys is a valuable tool in seeking trends in knowledge, attitudes, and behavior in respondents to certain variables that researchers wish to study. Commercial software is very popular for this purpose, with IBM's SPSS as the leading standard software for survey data analyses. A new open-source software program called "PSPP" from the GNU foundation has been written to allow free and unlimited data analyses capabilities for everyone on most platforms (Unix, PC, Mac).

Survey questionnaire design (which includes question generation, variable coding, and online questionnaire implementation) and data gathering are areas where a lot of time and effort is spent by the researcher. A free website for quick questionnaire generation, *Kwiksurveys.com*, presents a fast and efficient alternative to the time-consuming manual aspects of questionnaire design and implementation.

Together, *Kwiksurveys.com* and PSPP present an alternative and free platform for generating surveys, descriptive statistics, and performing deeper data analyses with no associated software costs. This presents everyone, and particularly third world academic institutions, a statistical analysis platform that can be useful and cost-free for any activities dealing with data analyses. Students can also benefit from these in an educational setting in that they can download and install the PSPP program on their local machines without being limited by cost limitations and time- or machine-locked licenses.

Some sample analyses of survey results from the author's UPOU (University of the Philippines Open University) Science, Technology, and Society (STS) class are presented, including selected examples of data analyses of knowledge, attitude, and behavior related to Facebook and social network usage by students.

Introduction

Data analysis of surveys is a common task in analytics in almost all sectors of academic and commercial endeavor, and it will only continue to gain more importance and usage in the future. Commercial software is very popular for this purpose, with IBM's SPSS as the leading application used by academe and industry alike. Commercial software programs always have limitations of cost, time or machine locks, and are not a viable solution for many third world countries and academic institutions. This paper aims to present an alternative free platform for questionnaire design and implementation and statistical analyses of survey results.

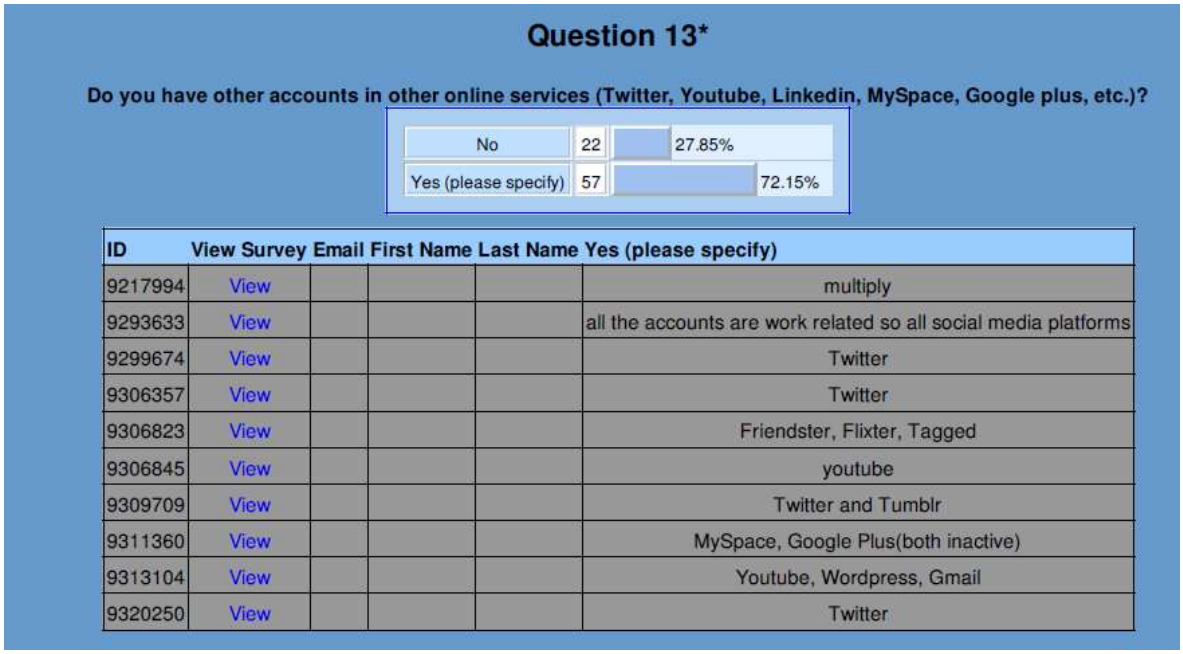
Methodology

Questionnaire design and data collection

The website Kwiksurveys.com (*http://www.kwiksurveys.com/*) was used to design a questionnaire for determining different aspects of knowledge, attitudes, and behaviors of Science, Technology, and Society (STS) students on social networks in general and on Facebook usage in particular. All questions were set to be required for answering, and all were in the multiple choice format, with some allowances for open-ended responses. The survey editor is very intuitive to use and questions can be added at any part of the questionnaire (see Figure 1 below).

Figure 1 . The Kwiksurvey.com survey editor page. A toolbar atop each question is present for very quick and efficient editing.

After the survey, all the data obtained were downloaded as a CSV (comma-separated values) file and imported into PSPP. The questionnaire can also be downloaded as an MSWord file or as a PDF file. Below (*Figure 2*) is a screen capture of a summary table for one question in the study:



Question 13*

Do you have other accounts in other online services (Twitter, Youtube, Linkedin, MySpace, Google plus, etc.)?

No	22		27.85%
Yes (please specify)	57		72.15%

ID	View Survey	Email	First Name	Last Name	Yes (please specify)
9217994	View				multiply
9293633	View				all the accounts are work related so all social media platforms
9299674	View				Twitter
9306357	View				Twitter
9306823	View				Friendster, Flixter, Tagged
9306845	View				youtube
9309709	View				Twitter and Tumblr
9311360	View				MySpace, Google Plus(both inactive)
9313104	View				Youtube, Wordpress, Gmail
9320250	View				Twitter

Figure 2 . Kwiksurvey.com summary display of answers to question 13 which allows open-ended responses. The open-ended answers are listed clearly along with a simple bar graph of the results of answers to the question.

Variable coding in PSPP

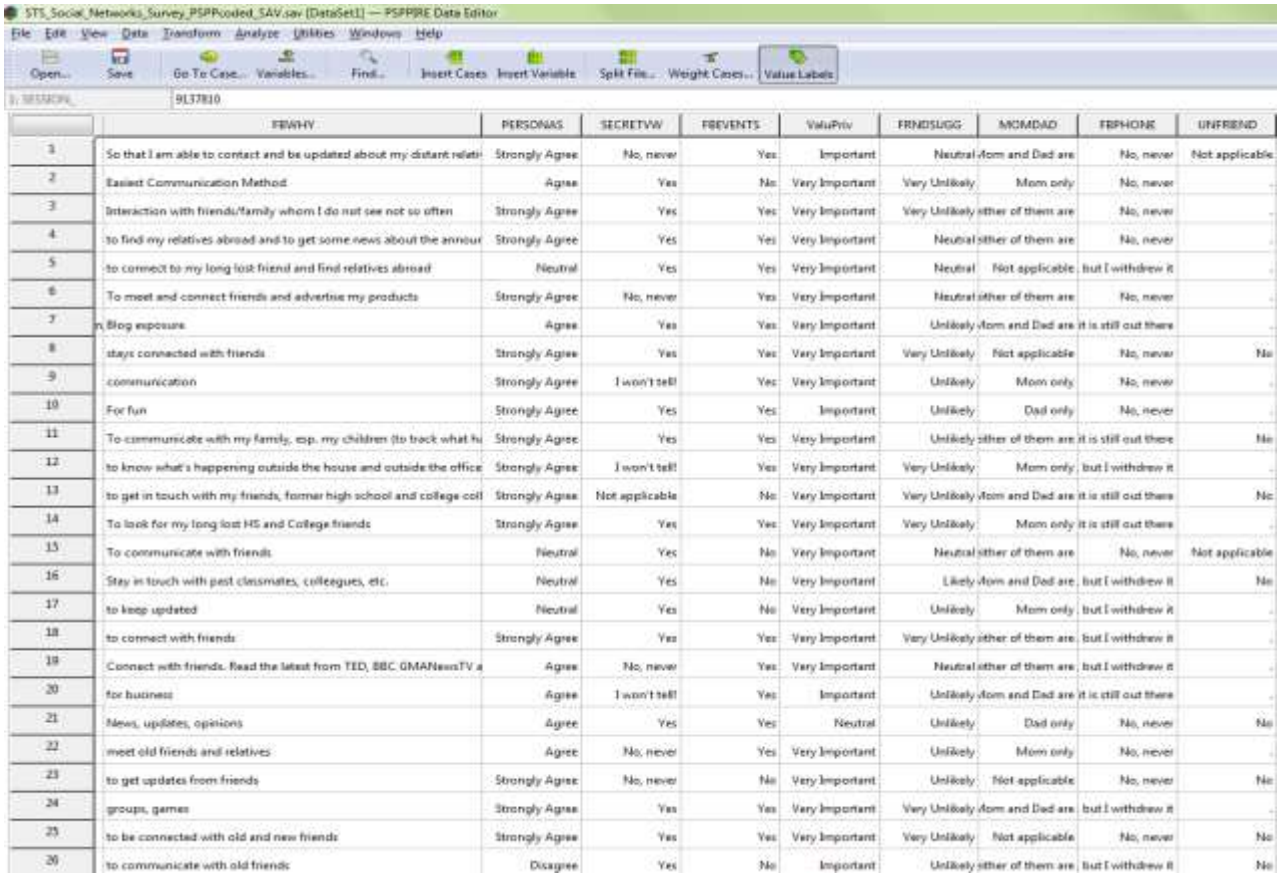
All of the questionnaire's questions were given variable names and each data value was assigned a variable value and variable label. This process is very similar to what is done in SPSS. The dialog boxes were almost identical to those of SPSS; *figure 3* shows a typical dialog box for assigning value labels to variables.

Recoding of variables was also done for some variables to facilitate data analyses. The *Kwiksurveys.com* website also provides the survey's summarized results with basic frequencies and graphs for answers to each question. This facilitates preliminary screening of variables for analyses, and also gives a quick check of data gathering accuracy. It also gives at a quick glance an organized list of answers to open-ended answers, something that SPSS is not that successful in displaying clearly in its output.

Figure 3 . Value labels dialog box for the EMPLSTAT (Employment Status) variable

Variable and data views in PSPP

PSPP appears and behaves much like SPSS in data handling, editing of variable names, variable types, variable values, and missing values. The interface is very much a look-alike of SPSS, and SPSS users will become familiar with PSPP very quickly. Data can also be viewed as numerical values or as text labels depending on what the user wants to do (*Figure 4*). At a glance, the spreadsheet can tell us how our data is organized, and a quick check of the number of respondents can be made by looking at the final row number.



	FBWHY	PERSONAS	SECRETVW	FBEVENTS	ValuPriv	FRNDSUGG	MOMDAD	FBPHONE	UNFRIEND
1	So that I am able to contact and be updated about my distant relati	Strongly Agree	No, never	Yes	Important	Neutral	Mom and Dad are	No, never	Not applicable
2	Easiest Communication Method	Agree	Yes	No	Very Important	Very Unlikely	Mom only	No, never	.
3	Interaction with friends/family whom I do not see not so often	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	either of them are	No, never	.
4	to find my relatives abroad and to get some news about the annour	Strongly Agree	Yes	Yes	Very Important	Neutral	either of them are	No, never	.
5	to connect to my long lost friend and find relatives abroad	Neutral	Yes	Yes	Very Important	Neutral	Not applicable	but I withdrew it	.
6	To meet and connect friends and advertise my products	Strongly Agree	No, never	Yes	Very Important	Neutral	either of them are	No, never	.
7	Blog exposure	Agree	Yes	Yes	Very Important	Unlikely	Mom and Dad are	it is still out there	.
8	stays connected with friends	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	Not applicable	No, never	No
9	communication	Strongly Agree	I won't tell!	Yes	Very Important	Unlikely	Mom only	No, never	.
10	For fun	Strongly Agree	Yes	Yes	Important	Unlikely	Dad only	No, never	.
11	To communicate with my family, esp. my children (to track what hi	Strongly Agree	Yes	Yes	Very Important	Unlikely	either of them are	it is still out there	No
12	to know what's happening outside the house and outside the office	Strongly Agree	I won't tell!	Yes	Very Important	Very Unlikely	Mom only	but I withdrew it	.
13	to get in touch with my friends, former high school and college coll	Strongly Agree	Not applicable	No	Very Important	Very Unlikely	Mom and Dad are	it is still out there	No
14	To look for my long lost HS and College friends	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	Mom only	it is still out there	.
15	To communicate with friends	Neutral	Yes	No	Very Important	Neutral	either of them are	No, never	Not applicable
16	Stay in touch with past classmates, colleagues, etc.	Neutral	Yes	No	Very Important	Likely	Mom and Dad are	but I withdrew it	No
17	to keep updated	Neutral	Yes	No	Very Important	Unlikely	Mom only	but I withdrew it	.
18	to connect with friends	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	either of them are	but I withdrew it	.
19	Connect with friends. Read the latest from TED, BBC, GMANewsTV a	Agree	No, never	Yes	Very Important	Neutral	either of them are	but I withdrew it	.
20	for business	Agree	I won't tell!	Yes	Important	Unlikely	Mom and Dad are	it is still out there	.
21	News, updates, opinions	Agree	Yes	Yes	Neutral	Unlikely	Dad only	No, never	No
22	meet old friends and relatives	Agree	No, never	Yes	Very Important	Unlikely	Mom only	No, never	.
23	to get updates from friends	Strongly Agree	No, never	No	Very Important	Unlikely	Not applicable	No, never	No
24	groups, games	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	Mom and Dad are	but I withdrew it	.
25	to be connected with old and new friends	Strongly Agree	Yes	Yes	Very Important	Very Unlikely	Not applicable	No, never	No
26	to communicate with old friends	Disagree	Yes	No	Important	Unlikely	either of them are	but I withdrew it	No

Figure 4. Value labels view of the PSPP data grid. The values of each variable (column) per survey respondent (rows) are shown in textual format. One can also view the numeric codes by clicking on the main menu options.

Data analyses using PSPP

The PSPP software (*Stover, 2010a*) is an open source statistical analysis software intended to be a "clone" of the commercial SPSS software by IBM (*IBM Corporation, 2011*) and is distributed by the GNU Software Foundation under standard open-source licensing agreements (GNU General Public License version 3.0 (GPLv3)) (*Free Software Foundation, 2011*).

PSPP (*version 0.7.8-g812c8f-blp-build20111213 (PSPP-Master-2011-12-13*) was downloaded from the URL:
http://sourceforge.net/projects/pspp4windows/files/2011-12-13/pspp-master-20111213-Setup.exe/download

It is a compilation for MS Windows built with MinGW and cross-compiled on openSUSE 11.3 (*Highlights of the current PSPP-for-Windows setup, 2011*). All of the installations, data import, and statistical analyses were performed on a PC platform running 64-bit Windows 7 on an Intel ® Core ™ i3-2100 CPU at 3.10 GHz with 4 GB of RAM. Tests for Independence of two variables were performed using the built-in Chi-square analysis module in PSPP. (Note: see further comments in the section below on the *limitations of PSPP*).

Results and discussion

Summarizing data for preliminary analyses

Data visualization for preliminary screening and understanding of the sample's data distribution is usually done using graphs and frequency tables. PSPP does both quite well, though it does not have the interactive graphs capability of SPSS where one can, in real time, move parts of the graph and see it change. The pie charts *(figure 4)* are pretty basic, with no capability for labeling with percentages. This may be a big disadvantage of the program, but it is easy to copy and paste the tabular data *(Table 1)* into MS Excel and generate better graphs from there.

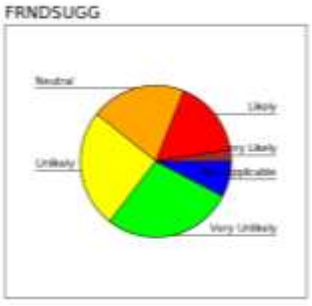


Figure 5. Pie chart of the variable FRNDSUGG (Likelihood of "friending" because of Facebook Friend Suggestion).

Table 1. Frequency table of the variable FRNDSUGG (Likelihood of "friending" because of Facebook Friend Suggestion). Percentages, Valid Percentages, and Missing Values are important measures in examining frequencies for each variable in the study.

FRNDSUGG

Value Label	*Value*	*Frequency*	*Percent*	*Valid Percent*	*Cum Percent*
Very Likely	1	2	2.53	2.53	2.53
Likely	2	13	16.46	16.46	18.99
Neutral	3	16	20.25	20.25	39.24
Unlikely	4	20	25.32	25.32	64.56
Very Unlikely	5	22	27.85	27.85	92.41
Not applicable	6	6	7.59	7.59	100.00
	Total	79	100.0	100.0	

FRNDSUGG

N	*Valid*	79
	Missing	0
Mean		3.82
Minimum		1.00
Maximum		6.0

Recoding data and variable values in PSPP

Data analysis of surveys is a highly-iterative process where the variables explored for relationships or causative effects are constantly cross-tabulated, recoded to remove missing values and low counts in specific categories of other variables paired with the main variable, and examined in terms of more detailed statistics (e.g., Pearson Chi-square values and significance levels). This constant hypothesis testing and re-testing are led by whatever results are obtained in the initial rounds of iterative analyses.

The ability to recode variable values into system missing values or into combined categories in crosstab analyses is crucial in iterative analyses. Many qualitative statistical tests rely on reasonably sufficient expected values in each given cell of a crosstab table, without which the analyses may be invalid because of low frequency counts in the cells of some categories of a given paired variable. *Figure 6* below shows a typical set of dialog boxes for recoding variables and variable values.

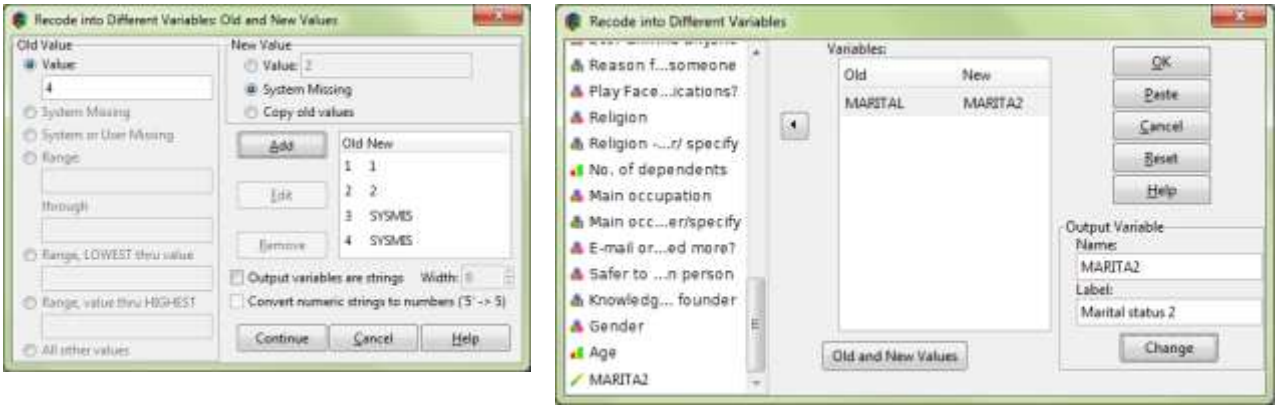


Figure 6. Dialog boxes for recoding two values (Separated/Divorced and Widowed) with very low counts for Crosstabulation analyses into system missing values for the variable MARITAL (Marital status).It has been recoded into a new Marital status variable (MARITA2)

Crosstabulations and data analyses in PSPP

Crosstabs is a useful technique in finding out if two categorical variables are independent of one another or not, and is used in conjunction with Pearson's Chi-square test. Sometimes we can be surprised at which variable pairs are actually independent or dependent on one another from our obtained samples.

An example (*figure 7*) from an online survey by the author (*Frio, 2011*) of students in STS (Science, Society and Technology) at UPOU showed an unexpected relationship between gender and whether or not the student played applications ("apps) on Facebook.

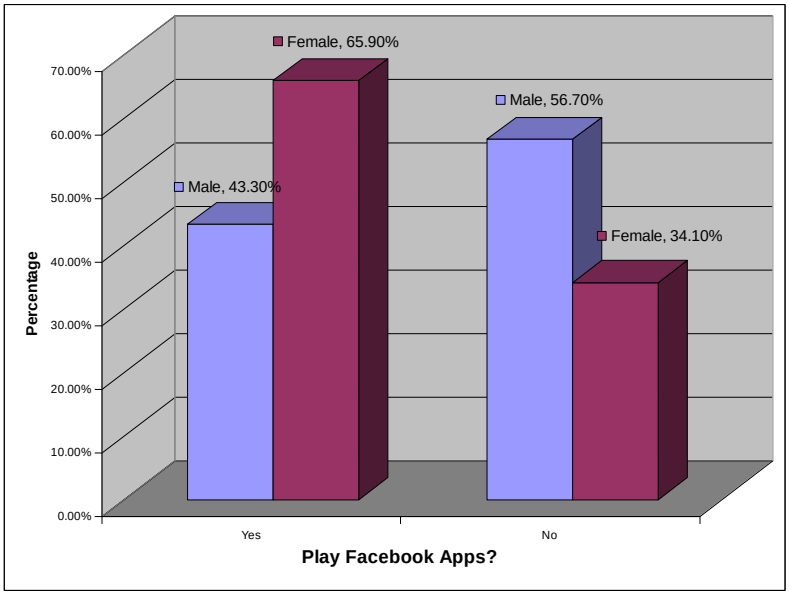


*Figure 7. Graph of the crosstab percentages of the variables **Gender** by **Play Facebook apps** generated in Microsoft Excel from the crosstabs data in Table 2.*

Figure 7 (above) and table 2 (below) show the crosstabs and Chi-square analyses. The Chi-square test shows a significant difference (p=0.05) between men and women in their playing of applications on Facebook, with more women (65.90%) than men (43.30%) saying that they played Facebook "apps" like Farmville. The results indicate that women tend to play more apps than men on Facebook! What these results mean is worthy of further inquiry – perhaps another survey centered only on Facebook apps would be interesting to do.

*Table 2. Crosstabulation and Chi-square results of **Gender** by **Play Facebook apps** variables. The Crosstabs percentages and Expected values were labeled outside of PSPP.*

	Play Facebook applications?		
Gender	**Yes**	**No**	**Total**
Male	13	17	30
Expected	17	13	0
Row %	43.30%	56.70%	100.00%
Column %	31.00%	53.10%	40.50%
Total %	17.60%	23.00%	40.50%
Female	29	15	44
Expected	25	19	0
Row %	65.90%	34.10%	100.00%
Column %	69.00%	46.90%	59.50%
Total %	39.20%	20.30%	59.50%
Total	42	32	74
Total %	56.80%	43.20%	100.00%

Chi-square tests.

Statistic	***Value***	***df***	***Asymp. Sig. (2-tailed)***
Pearson Chi-Square	3.7	1	0.05
Likelihood Ratio	3.71	1	0.05
Continuity Correction	2.84	1	0.09
Linear-by-Linear Association	3.65	1	0.06
N of Valid Cases	74		

The PSPP crosstabs output lacks labels which can throw one off as it is not clear which row and column values are which; the correctly-labeled tables are presented in this paper (*table 2*) with clear row and column percentages and expected values. These were re-calculated using formulas in MS Excel in order to make sure that the values labeled are indeed the correct ones. The expected values were calculated in principle as [row total x column total / main total], or more formally,

$$Exp = (R_i * C_j)/Total$$

where

Exp = expected value

R_i = Row total at i^{th} cell's row

C_j = Column total at j^{th} cell's column

Total = main total of all counts

Limitations of PSPP

In the MS Windows build *(version 0.7.8-g812c8f-blp-build20111213 (PSPP-Master-2011-12-13))*, various shortcuts and keyboard input options that are standard in normal Windows applications do not work in PSPP. Examples include the use of the left and right arrow keys, and the Home, Page Up and Page Down keys in the text fields when editing variable names in the Variable view mode of PSPP. In SPSS, variable values can be copied and pasted, which is a very useful and time-saving feature specially when recoding variables into new variables (this is important when we want clarity with which variables have already been recoded for analysis). This cannot be done PSPP and it is recommended to "fix" this with a new Windows build of the PSPP code.

The PSPP Output Viewer window also is not amenable to highlighting and copying of contents, which is usually done in SPSS when preparing MSWord write-ups and reports. The work-around is to export the output to html or PDF formats, which works acceptably well and enables copying and pasting crosstabs data directly into Microsoft Word.

Also in the PSPP build notes, the "bug" list in crosstabs analysis states that Pearson's R (but not Spearman) is "off a little" (*Stover, 2010b*). This statement may be incorrect, as Pearson's R is used for linear correlation analysis of two continuous variables; the Pearson chi-square (which falls under PSPP and SPSS' Crosstabs analysis) is a different measure used for testing for relationships between two categorical (i.e., non-continuous) variables (*Statsoft, 2011*). The degree to which this varies (or whether it really does deviate) from actual correct values is unclear at this time. This may be a valid consideration against using the program, but if one can have alternative means of calculating this statistic outside of PSPP, one can still use the program with sufficient confidence. It may be due to floating point limitations, but it is unlikely as modern day computers have no problems performing multiple floating point calculations. It is recommended that this calculation bug be examined further and corrected for whatever is causing it to be "off".

For the Crosstabs output, the labels are not presented in each cross-tabulation; The Expected counts, row, column, and total percentages are not labeled at all. These need to recomputed outside of PSPP (e.g., in MS Excel) in order to determine which of the numbers correspond to which count or percentage. The headings of the crosstabs output ([count, row %, column %, total %, expected]) do not correspond with the consecutive rows of the table.

The graphing capability of this build is useful only for basic overviews of the data, and are not really suitable as presentation graphics. However, once can always export the tabular data into graphing software (e.g., MS Excel) and use their functionality to generate more informative graphs.

Conclusions

It truly is remarkable what this duo of questionnaire design and statistical analysis software can do to facilitate online survey implementation and data analyses. The *Kwiksurveys.com* website facilitates free and quick questionnaire design and data gathering via implementation of online answering of the questionnaire. The PSPP program is quite a capable tool for data analysis, and will surely continue to improve in accuracy and user-friendliness in future builds of the code and platform ports. These two are completely free to use and have no time or machine locks, permitting anyone to do surveys and perform data analyses using one's own PC, Mac, or UNIX machine.

References

Free Software Foundation. 2011. GNU General Public License version 3.0 (GPLv3). Retrieved from: http://www.gnu.org/copyleft/gpl.html

Frio, E.J.L. 2011. Cybertechnology I - Social Networks survey (unpublished). Science, Technology and Society online class, UP Open University.

Highlights of the current PSPP-for-Windows setup. 2011. Retrieved from: http://pspp.awardspace.com/

IBM Corporation. 2011.IBM SPSS software for predictive analytics. Retrieved from: http://www-01.ibm.com/software/analytics/spss/

Kwiksurveys.com. 2011. Retrieved from: http://www.kwiksurveys.com/

PSPP build 0.7.8-g812c8f-blp-build20111213 program download. 2011. Retrieved from: https://sourceforge.net/projects/pspp4windows/files/2011-12-13/pspp-master-20111213-Setup.exe/download

StatSoft, Inc. (2011). Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/.

Stover, J. 2010a. GNU PSPP: A Free Clone of SPSS. Joint Statistical Meetings. Vancouver..

Stover, J. 2010b. GNU PSPP manual. GNU PSPP: A Free Clone of SPSS. Joint Statistical Meetings. Vancouver. 2010.