TESTS AS CONTENT SAMPLES: AN INDEX OF GOODNESS OF FIT IN TERMS OF INFORMATION

LUCIA BONCORI

Dipartimento di Psicologia
Università di Roma "La Sapienza"

INTRODUCTION

Content validity is usually referred to in connection with achievement tests. It mainly concerns the stimulus properties of a test, and it involves either general and complex problems related to test construction, hardly amenable to a single statistical index, or more specific problems of representativeness of the test as a content sample (Cronbach, 1971).

THE CONTINGENCY COEFFICIENT C AND RELATED INDEXES OF CONTENT VALIDITY

A proper statistical approach to the latter problems usually involves the most widely known non-parametric statistical methods-namely chi-square and contingency coefficient.

However, the contingency coefficient C is known to have several limitations (Siegel, 1956; Kerlinger, 1964, 1973). As to perspicuity of interpretation, it is rather inconvenient for *C* not to attain unity even when variables are perfectly correlated, to have a maximum value not easily determinable for individual contingency tables; not to have values distributed according to an equal interval scale; and not to be directly comparable to the most common measures of correlation, e.g. the Bravais-Pearson's *r* and the Spearman's *rho*. Other disadvantages are common to *C* and to most statistics based on nominal data — e.g. no sign to differentiate between positive and negative correlation (a meaningless notion if data are nominal), and lack od comparability among values yielded by contigency tables of different size.

In case *C* is calculated on a contingency table where an observed number of items falling in various categories is compared to expected frequencies for the same categories (Hoste, 1981), we have further problems as to the meaning of these statistics. A particular disadvantage is that it differs from Pearson's *r* in that it approaches its upper limit in case of minimum fit between observed and expected frequencies and its lower limit in case of perfect agree-

ment. This feature is not irrational if we read a contingency coefficient in terms of the underlying null hypothesis, i.e. assuming that agreement between variables depends on random factors, but it is as much disturbing as not to have an algebraic sign to show whether expected and observed frequencies are «rather unfit» or «rather fit». The use of the complementary quantity $C_{cv} = (1\text{-}C)$ as a coefficient of validity (Hoste, 1981) is effective in that C_{cv} approach its maximum value in case of perfect agreement, but leaves the «unfitness» end values rather indetermined, never reaching zero; and it is still distributed along an unequal interval scale, referred to a model so unfamiliar to psychometricians that most statistics textbooks don't even quote sources like Kelly (1923) where the standard error of C is discussed.

Statistics referring to information theory

A different approach to the problem of content validity quantification can be attempted using «uncertainty» or «information» statistics (Shannon & Weaver, 1949).

Information theory has originated statistics applicable to qualitative data and basically interpretable in terms of variance. For instance, given a discrete probability distribution based on four categories

A	B	C	D
P (A)	P (B)	P (C)	P (D)

we may quantify the amount of uncertainty or of lack of information over all possible cases drawn at random from this distribution. The extreme values will appear when a particular category has $P = 1$ (in which case there is no prior uncertainity and no further information to be gained drawing a case), and when all categories are equiprobable (in which case there is maximum prior amount of uncertainity and maximum information gain).

The «information contribution» or «surprise value» for each category in the distribution is defined to be

$$h = - \text{p (i)} \log_2 \text{p (i)} \qquad [1]$$

and the average amount of uncertainity or information for a series of categories is defined to be

$$H = \Sigma \text{p (i)} \log_2 \text{p (i)} \qquad [2]$$

Both h and H values refer to a unit called *bit* («binary unity digit») which

has a psychometric meaning in itself, being defined as the amount of information needed to decide between two equiprobable alternatives.

These statistics have some advantages over *C* and the Pearson's *chi square*. They may be expressed in the familiar terms of variance and proportions of variance accounted for, and they may be interpreted according to a sampling theory which is related to the likelihood ratio tests. It has also been shown (Attneave, 1959; Calonghi, 1978a) that

$$\chi^2 = -2 \log_e \lambda = 1.3863 \,.\, N \,.\, (H - H') \qquad [3]$$

where *lambda* is a likelihood ratio as described in Mood and Graybill (1963).

The equivalence between *lambda* and the Pearsons' chi square is good for large *n*. There is also reason to belive that *lambda* may be less affected by small sample size than are the Pearson's chi square tests, especially when degrees of freedom are more than one (Hays, 1963, 1973), which is nearly always the case with content validity problems.

Starting from these premises, we have worked out a statistical test useful for content validity problems.

Consider the following typical content validity problem.

	Content categories				
	A	B	C	D	
Expected frequencies (F_e)	25	25	25	25	
Expected proportions (P_e)	0.25	0.25	0.25	0.25	
Contribution to average information (h_e)	0.5	0.5	0.5	0.5	$H_e = 2$
Observed frequencies (F_o)	0	50	0	50	
Observed proportions (P_o)	0	0.5	0	0.5	
Contribution to average information (h_o)	0	0.5	0	0.5	$H_o = 1$
Difference in contribution to average information ($h_e - h_o = dh$)	0.5	0	0.5	0	$dh = 1$

The above data refer to a 100 items test in which we expect to have items evenly distributed over four content categories. In fact, we observe that our 100 items are evenly distributed over *two* out of the *four* expected categories. We may quantify the difference in variance or information contribution of each content category, to the test as a whole comparing expected and obser-

ved relative frequencies with reference to their h value. Adding up absolute differences over the four categories, we find a value $dh = 1$, roughly equivalent (see [3]) to a $\chi^2 = 69.315$ ($P \leq 0.01$ for d.f. = 3).

The value $dh = 1$ being expressed in bits (binary unity digits) can be read in absolute terms as «the difference in uncertainty or information between expected and observed proportions, i.e. the quantity of information needed to decide between two equiprobable alternatives, or the amount of uncertainty attached to two equiprobable alternatives». This definition is particularly fitting a case like ours, where observed frequencies are evenly spread over 1/2 of the expected categories. We also may reckon the average information values by row and finding $H_e = 2$ and $H_o = 1$. The difference $dh = 1$ may also be read in relation to H_e as $dh_{rel} = 1/2$, i.e. «the difference in relative information between expected and observed proportions is 1/2 the expected information» or «uncertainty about content equivalence of P_o and P_e is 1/2 the uncertainty attached to the expected distribution».

Computational problems may be easily solved reading h values corresponding to P from tables (Calonghi 1978; Senders 1958).

Statistics dh varies between a minimum value $dh = 0$ in case of no difference between expected and observed distribution, and a maximum value $dh = \log_2 K$, where K = number of categories. Its values are distributed according to a $\log_2$ scale, ad the null hyotesis can be tested with reference to *chi square*, as shown.

It may be read as the sum of differences in variance contribution of each content area. It is a sum of *absolute* differences - i.e. it does no matter whether a content area has a higher proportion of items in the expected or in the observed distribution. It has no algebraic sign to show whether the observed distribution is «rather fit» or «rather unfit» the expected one, because content areas are only nominal data and can be shifted from one location to the other along the contingency table withouth any alteration to their meaning. dh being an absolute measure of disagreement, its maximum values can be larger if yielded from a larger number of categories, just because the amount of disagreement is added up across the cells and may be larger if a larger number of opportunities for disagreement occurs. This feature is common to other contingency table statistics, and it is not so inconvenient if properly read. In fact, if we start with a four categories test as following:

	A_1	A_2	B_1	B_2
Expected proportions (P_e)	0.25	0.25	0.25	0.25
Observed proportions (P_o)	0	0.50	0	0.50

where there is an abvious lack of fit between P_e and P_o in categories A_1 and B_1, we may get the most perfect fit just reducing content areas to two:

	$A_1 + A_2$	$B_1 + B_2$
P_e	0.50	0.50
P_o	0.50	0.50

This does not mean that the latter test has more content validity than the prior one, but that four categories allow a more accurate analysis of disagreement than two.

Anyway, if comparisons among tables of different dimensions are involved, it may be useful to calculate a relative value of dh - dh_{rel} - dividing each dh value by the maximum dh value expecxted. dh_{rel} varies between zero in case of no difference between expected and observed proportions and one in case of perfect fit, and can be read as difference in «proportion of information» or «proportion of uncertainty».

Discussion of typical cases

A few typical cases will help to focus different properties of four statistics we may use in content validity problems - C; C_{cv}; and what we propose to call dh, i.e. «cumulative difference between paired values of h» or «absolute difference in information uncertainty between observed and expected proportions, measured in *bits*»; and dh_{rel}, i.e. «relative difference in information between observed and expected proportions» or proportion of maximum possible difference in information in fact observed. To simplify the matter, we will refer to two hypothetical tests, each one containing 100 items distributed in the first instance among subtests or content areas in the latter among 10 content areas, and we will discuss a few typical cases.

Case (i). Expected proportions (P_e) are equally distributed and observed proportions (P_o) are nearly the same (cumulative percentage difference below 5%)

4 categories

Content areas	A	B	C	D	
Expected proportions (P_e)	0.25	0.25	0.25	0.25	
values of h (h_e)	0.5	0.5	0.5	0.5	$H_e = 2$
Observed proportion (P_o)	0.25	0.25	0.26	0.24	
values of h (h_o)	0.5	0.5	0.5053	0.4941	$H_o = 1.999$

$C = 0.03 \quad C = 1 - 0.03 = 0.97 \quad \chi^2 = 0.08^{ns}$

$dh = (0.5 - 0.5) + (0.5 - 0.5) + (0.5 - 0.5053) + (0.5 - 0.4941) = 0.0006$

$\chi^2 = dh \cdot 1.3863 \cdot n = 0.083^{ns} \qquad dh_{rel} = 0.0006/2 = 0.0003$

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0.10	0.3322	
B	0.10	0.3322	0.10	0.3322	
C	0.10	0.3322	0.10	0.3322	
D	0.10	0.3322	0.10	0.3322	$C = 0.06$
E	0.10	0.3322	0.10	0.3322	$C_{cv} = 0.94$
F	0.10	0.3322	0.09	0.3127	$\chi^2 = 0.40^{ns}$
G	0.10	0.3322	0.11	0.3503	$dh = 0.08$
H	0.10	0.3322	0.09	0.3127	$\chi^2 = 10.42^{ns}$
I	0.10	0.3322	0.10	0.3322	$dh_{rel} = 0.08$ /
L	0.10	0.3322	0.11	0.3503	/ 3.322 = 0.02
	$H_e = 3.322$		$H_o = 3.319$		

In case (i) the meaning of C_{cv}, dh, and dh_{rel} is quite clear, in that $C_{cv} = 0.98$ and $C_{cv} = 0.94$ mean «content validity approaching maximum value»: and $dh = 0.006$, $dh = 0.08$, and $dh_{rel} = 0.003$, $dh_{rel} = 0.02$ mean «cumulative difference in information about content is nearly null». $C = 0.03$ and $C = 0.06$, both associated to a small value of *chi square* (0.08 and 0.40), are misleading if we are reading them as a measure of intensity of correlation. But their meaning becomes clear once read, as it should be, as a measure of non-systematic connection between *variations* in the two sets of attributes, because in case (i) one set shows some variation while the other does not.

Case (ii) Expected proportions (P_e) are equally distributed, having maximum dispersion across areas; observed proportions (P_o) have minimum dispersion

4 categories (cumulative percentage difference beyond 75%)

Content areas	P_e	h_e	P_o	h_o	
A	0.25	0.5	0	0	
B	0.25	0.5	0	0	$C = 0.87$
C	0.25	0.5	0	0	$C_{cv} = 0.13$
D	0.25	0.5	1.00	0	$\chi^2 = 300\cdots$
					$dh = 2$
					$\chi^2 = 238.63\cdots$
					$dh_{rel} = 2/2 = 1$
	$H_e = 2$		$H_o = 0$		

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0	0	
B	0.10	0.3322	0	0	
C	0.10	0.3322	0	0	
D	0.10	0.3322	0	0	
E	0.10	0.3322	0	0	$C = 0.95$
F	0.10	0.3322	0	0	$C_{cv} = 0.05$
G	0.10	0.3322	0	0	$\chi^2 = 900\cdots$
H	0.10	0.3322	0	0	$dh = 3.32$
I	0.10	0.3322	0	0	$\chi^2 = 460.25\cdots$
L	0.10	0.3322	1.00	0	$dh_{rel} = 3.322 / / 3.322 = 1$
	$H_e = 3.322$		$H_o = 0$		

$\cdots P \leqslant .001$

Case (ii) is the most extreme case for lack of correspondence we may in fact imagine for a content validity problem. Four categories are involved, but neither $C = 0.87$ nor $C_{cv} = 0.13$ seems to express this feature; the same statistics produce more expressive values in a ten categories problem: $C = 0.95$, $C_{cv} = 0.05$, as we should have expected. Statistics referring to information are quite clear in every case, telling us that the difference in information between expected and observed contents is equal to the maximum average information expected.

Case (iii) Both expected and observed proportions are evenly distributed, the number of cells with $P \leq 0$ in the observed distribution being 1/2 than in the expected distribution (cumulative percentage difference beyond 50%)

4 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.25	0.5	0	0	
B	0.25	0.5	0.5	0.5	$C = 0.71$
C	0.25	0.5	0	0	$C_{cv} = 0.29$
D	*0.25*	*0.5*	*0.5*	*0.5*	$\chi^2 = 100\cdots$
					$dh = 1$
					$\chi^2 = 69.322\cdots$
					$dh_{rel} = 1/2 =$
					$= 0.5$
	$H_e = 2$		$H_o = 1$		

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	.3322	0	0	
B	0.10	.3322	0.20	0.4644	$C = 0.71$
C	0.10	.3322	0	0	$C_{cv} = 0.29$
D	*0.10*	*.3322*	*0.20*	*0.4644*	$\chi^2 = 100\cdots$
E	0.10	.3322	0	0	$dh = 2.32$
F	0.10	.3322	0.20	0.4644	$\chi^2 = 321.899\cdots$
G	0.10	.3322	0	0	$dh_{rel} = 2.32 /$
H	0.10	.3322	0.20	0.4644	$/ 3.32 = 0.70$
I	0.10	.3322	0	0	
L	0.10	.3322	0.20	0.4644	
	$H_e = 3.322$		$H_o = 2.322$		

$\cdots P \leq .001$

Both examples in case (iii) refer to a difference in information between expected and observed distribution as large as the whole observed information value ($dh = H_o$), which can be regarded as a clear representation of the apparent lack of fit. *C* statistics are not as expressive, as they produce interme-

diate values ($C = 0.71$, $C_{cv} = 0.29$) just informing us that there is «some» agreement or disagreement.

It should also be noted that informational statistics appear to be more sensitive to the number of categories involved.

Case (iv) Expected proportions are evenly distributed, while observed proportions are distributed according to an arithmetic progression (cumulative percentage difference below 50%)

4 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.25	0.5	0.10	0.3322	
B	0.25	0.5	0.20	0.4644	$C = 0.41$
C	0.25	0.5	0.30	0.5211	$C_{cv} = 0.59$
D	*0.25*	*0.5*	*0.40*	*0.5288*	$\chi^2 = 20\cdots$
					$dh = 0.25$
					$\chi^2 = 35.115\cdots$
					$dh_{rel} = 0.25 /$
					$/ 2 = 0.12$
	$H_e = 2$		$H_o = 1.8465$		

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0.01	0.0664	
B	0.10	0.3322	0.03	0.1518	
C	0.10	0.3322	0.05	0.2161	
D	0.10	0.3322	0.07	0.2686	$C = 0.50$
E	0.10	0.3322	0.09	0.3127	$C_{cv} = 0.50$
F	*0.10*	*0.3322*	*0.11*	*0.3503*	$\chi^2 = 33\cdots$
G	0.10	0.3322	0.13	0.3826	$dh = 1.02$
H	0.10	0.3322	0.15	0.4105	$\chi^2 = 141.44\cdots$
I	0.10	0.3322	0.17	0.4346	$dh_{rel} = 1.02 /$
L	0.10	0.3322	0.19	0.4552	$/ 3.322 = 0.31$
	$H_e = 3.322$		$H_o = 3.0488$		

$\cdots P \leq 0.001$

Case (iv) is typical of an intermediate level of agreement. Inspectiong the two series of expected and observed proportions as wholes, we notice that expected proportions are evenly distributed, while observed proportions are distributed according to an arithmetic progression. Even taking into account that the way categories are ordered does not matter, it is sensible to regard a difference of this kind as a remarkable one as far as content validity is concerned. In fact, *chi square* values in case (iv) are always beyond $P = 0.001$. As to sensitiveness to the amount of disagreement between expected and observed proportions we notice that C and C_{cv} have values around 0.5, which conveys the message of an intermediate level of agreement/disagreement, with a slightly higher figure for agreement in the four categories case ($C = 0.41$, $C_{cv} = 0.59$). Informational statistics on the one hand indicate a similar average amount of information for expected and observed distribution (cfr. H's), on the other hand show that the difference is small in the four categories case ($dh = 0.25$, one quarter of a bit; $dh_{rel} = 12$ per cent of the maximum possible gap), more noticeable in the ten categories case ($dh = 1.02$, more than one bit; $dh_{rel} = 31$ per cent of maximum possible difference). In other words, C statistics seem to be more sensitive to differences in the ten categories case, less in the four categories case.

Case (v) Both expected and observed proportions are distributed according to arithmetic progression, following an inverse order (over 80% cumulative percentage difference)

4 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0.40	0.5288	
B	0.20	0.4644	0.30	0.5211	$C = 0.55$
C	0.30	0.5211	0.20	0.4644	$C_{cv} = 0.45$
D	0.40	0.5288	0.10	0.3322	$\chi^2 = 120.833\cdots$
					$dh = 0.51$
					$dh_{\text{rel}} = 0.25$
					$\chi^2 = 70.230\cdots$
	$H_e = 1.8265$		$H_o = 1.8465$		

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.01	0.0664	0.10	0.4552	
B	0.03	0.1518	0.17	0.4346	
C	0.05	0.2161	0.15	0.4105	$C = 0.90$
D	0.07	0.2686	0.13	0.3826	$C_{cv} = 0.10$
E	*0.09*	*0.3127*	*0.11*	*0.3503*	$\chi^2 = 453.30\cdots$
F	0.11	0.3503	0.09	0.3127	$dh = 2.035$
G	0.13	0.3826	0.07	0.2686	$dh_{rel} = 0.61$
H	0.15	0.4105	0.05	0.2161	$\chi^2 = 282.14\cdots$
I	0.17	0.4346	0.03	0.1518	
L	0.19	0.4552	0.01	0.0664	
	$H_e = 3.0488$		$H_o = 3.0488$		

$\cdots P \leq 0.001$

Case (v) compares distributions carrying the same average amount of information (cfr. *H*), with categories inversely ordered as to their proportional contribution. If we were dealing with ordinal data we would expect a negative correlation; with nominal data we only may acknowledge a partial lack of fit, in that each one of the ten pairs of values could be arranged in a different position. In the four categories case all the four statistics we have used seem to express a «partial lack of fit» quite clearly ($C = 0.55$, $C_{cv} = 0.45$, $dh = 0.51$, $dh_{rel} = 0.25$) *C* and C_{cv} are quite deceiving if we consider the ten categories case ($C = 0.90$, $C_{cv} = 0.10$), especially if compared to the four categories values. Both *dh* and dh_{rel} produce clear values-as in previous cases, even dh_{rel} happen to be sensitive to differences related to the number of categories involved.

Let us now discuss a few more cases approaching the area of rejection of the null hypothesis.

Case (vi) Expected proportions are evenly distributed, while observed proportions are not, showing approximately 1/3 frequencies equidistributed over 1/2 categories and 2/3 frequencies equidistributed across the remaining half (cumulative percentage difference below 33%).

4 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.25	0.5	0.17	0.4346	
B	0.25	0.5	0.17	0.4346	$C = 0.30$
C	0.25	0.5	0.33	0.5278	$C_{cv} = 0.70$
D	*0.25*	*0.5*	*0.33*	*0.5278*	$\chi^2 = 10.24\cdots$
					$dh = 0.19$
					$dh_{rel} = 0.09$
					$\chi^2 = 25.84\cdots$
	$H_e = 2$		$H_o = 1.9248$		

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0.07	0.2686	
B	0.10	0.3322	0.07	0.2686	
C	0.10	0.3322	0.07	0.2686	$C = 0.29$
D	0.10	0.3322	0.07	0.2686	$C_{cv} = 0.71$
E	*0.10*	*0.3322*	*0.07*	*0.2686*	$\chi^2 = 9^{ns}$
F	0.10	0.3322	0.13	0.3826	$dh = 0.57$
G	0.10	0.3322	0.13	0.3826	$dh_{rel} = 0.17$
H	0.10	0.3322	0.13	0.3826	$\chi^2 = 79.019\cdots$
I	0.10	0.3322	0.13	0.3826	
L	0.10	0.3322	0.13	0.3826	
	$H_e = 3.3322$		$H_o = 3.256$		

$^{ns}P > 0.005$
$\cdot P \leq 0.05$
$\cdots P \leq 0.001$

Case (vi) examples refer to partial lacks of fit involving approximately one third of frequencies, which can be regarded as noticeable in most content validity problems.

C and C_{cv} values seem to be sufficiently expressive, but the associated *chi square* values are in the four categories case beyond the $P = 0.05$ significance level, in the ten categories case below the same level. In other words, a difference involving one third frequencies should be regarded as «not significant» if spread across ten categories. The difference in significance between the four and the ten categories cases, is not surprising if we recall that a high

number of categories decreases the power of a statistical test, i.e. increase the probability of committing a Type II error - accepting the null hopothesis when in fact it is false.

On the other hand, *dh* values show little information difference between expected and observed proportions of contents - $dh = 0.19$ and $dh_{rel} = 0.09$ in the four categories case, $dh = 0.57$ and $dh_{rel} = 0.17$ in the ten categories cases. However, the amount of difference is enough to produce chi square's values beyond the $P = 0.001$ level in both cases. We may say that *dh* looks more sensitive than *C* to comparatively small differences and less biased toward a Type II error.

Let now examine a further example where differences are less noticeable than in case (vi).

Case (vii) Expected proportions are evenly distributed, while observed proportions are not, showing approximately 2/3 frequencies equidistributed over 3/4 categories and 1/3 frequencies equidistributed across the remaining 1/4 (cumulative percentage difference below 20%).

4 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.25	0.5	0.34	0.5292	
B	0.25	0.5	0.22	0.4806	$C = 0.19$
C	0.25	0.5	0.22	0.4806	$C_{cv} = 0.81$
D	0.25	0.5	0.22	0.4806	$\chi^2 = 3.64^{ns}$
					$dh = 0.08$
					$dh_{rel} = 0.04$
					$\chi^2 = 12.116^{\cdot\cdot}$

$H_e = 2$ $H_o = 1.971$

10 categories

Content areas	P_e	h_e	P_o	h_o	
A	0.10	0.3322	0.13	0.3826	
B	0.10	0.3322	0.13	0.3826	
C	0.10	0.3322	0.12	0.3671	$C = 0.14$
D	0.10	0.3322	0.09	0.3127	$C_{cv} = 0.86$
E	0.10	0.3322	0.09	0.3127	$\chi^2 = 1.91^{ns}$
F	0.10	0.3322	0.09	0.3127	$dh = 0.27$
G	0.10	0.3322	0.09	0.3127	$dh_{rel} = 0.08$

H	0.10	0.3322	0.09	0.3127	$\chi^2 = 37.735^{\cdots}$
I	0.10	0.3322	0.09	0.3127	
L	0.10	0.3322	0.09	0.3127	

$H_e = 3.3322$ $H_o = 3.3212$

$^{ns}P < 0.005$
$^{\cdot}P \leq 0.05$
$^{\cdots}P \leq 0.001$

Case (vii) examples refer to a cumulative difference involving from 15 to 18 per cent of the items, which can still be regarded as noticeable in some content validity problems.

Both C and C_{cv} values show a remarkable agreement, and Pearson's chi square values are below the $P = 0.05$ level of significance.

On the other hand *dh* and dh_{rel} show small values (*dh* = 0.08 and 0.27, dh_{rel} = 0.04 and 0.08), but match to *chi square* values significant beyond $P = 0.01$.

Generally speaking, we may say that *dh* and related *chi square* are more sensitive to differences and less biased toward Type II error, or in other words that they have more «power». This feature implies that *chi square* values can still get values beyond $P = 0.05$ when differences are comparatively small. For instance, in a case involving four categories and 100 items, an observed distribution showing proportions like

A	B	C	D
$P = 0.22$	$P = 0.24$	$P = 0.26$	$P = 0.28$

matched to an expected equidistribution, would still produce a chi square value beyond $P = 0.05$.

Conclusions

In summary, we may say that the four different statistics we have been using as content validity indexes produce values following approximately the same rank order, but differing as to sensitivity to number of categories and to comparatively small differences between expected and observed distribution (cf. tab. 1).

Table 1 - Summary of content validity indexes related to a few typical cases.

Case	Description	Index	Number of categories	
			Four	Ten
(ii)	Expected proportions are equally distributed, having minimum dispersion across areas; observed proportions have minimum dispersion (cumulative percentage difference beyond 75%)	C	0.87	0.95
		C_{cv}	0.13	0.005
		χ^2	300···	900···
		dh	2	3.32
		dh_{rel}	1	1
		χ^2	238.63···	460.25···
(iii)	Both expected and observed proportions are evenly distributed, the number of cells with P = 0 in the observed distribution being 1/2 the expected distribution (cumulative percentage difference beyond 50%)	C	0.71	0.71
		C_{cv}	0.29	0.29
		χ^2	100···	100···
		dh	1	2.32
		dh_{rel}	0.5	0.70
		χ^2	9.322···	321.899···
(iv)	Both expected and observed proportions are distributed according to an arithmetic progression, following an inverse order (over 80% cumulative percentage difference)	C	0.55	0.90
		C_{cv}	0.45	0.10
		χ^2	120.833···	453.30···
		dh	0.51	2.035
		dh_{rel}	0.25	0.61
		χ^2	'0.230···	282.14···
(v)	Expected proportions are evenly distributed, while observed proportions are distributed according to an arithmetic progression (cumulative percentage difference below 50%)	C	0.41	0.50
		C_{cv}	0.59	0.50
		χ^2	20···	33···
		dh	0.25	1.02
		dh_{rel}	0.12	0.31
		χ^2	Ç5.115···	141.44···
(vi)	Expected proportions are evenly distributed, while observed proportions are not, showing approximately 1/3 frequencies equidistributed over 1/2 categories and 2/3 frequencies equidistributed across the remaining half (cumulative percentage difference below 33%)	C	0.30	0.29
		C_{cv}	0.70	0.71
		χ^2	10.24·	9^{ns}
		dh	0.19	0.57
		dh_{rel}	0.09	0.17
		χ^2	25.84···	79.019···
(vii)	Expected proportions are evenly distributed, while observed proportions are not, showing approximately	C	0.19	0.41
		C_{cv}	0.81	0.86
		χ^2	3.64^{ns}	1.91^{ns}

Case	Description	Index	Number of categories	
			Four	Ten
	2/3 frequencies equidistributed over 3/4 categories and 1/3 frequencies equidistributed across the remaining 1/4 (cumulative percentage difference below 20)	*dh*	0.08	0.27
		dh_{rel}	0.04	0.08
		χ^2	£2.116···	37.735···
(i)	Expected proportions are equally distributed, and observed proportions are nearly the same (cumulative percentage difference below 5%)	*C*	0.03	0.06
		C_{cv}	0.97	0.94
		χ^2	0.08^{ns}	0.40^{ns}
		dh	0.0006	0.08
		dh_{rel}	0.0003	0.02
		χ^2	..083^{ns}	10.42^{ns}

$^{ns}P > 0.05$
$\cdot P \leqq 0.05$
$\cdot\cdot P \leqq 0.01$
$\cdot\cdot\cdot P \leqq 0.001$

When ten categories are involved, *C* and C_{cv} values vary in inverse order as to case (iii) and (v), due to different sensitivity to equidistribution and to the number of categories.

As to power, *dh* and related *chi square* appear to be more sensitive to small differences and less biased toward Type II error, i.e. less likely to induce to accept the null hypothesis when it is in fact false.

Both *dh* and dh_{rel} are more sensitive to the number of categories involved, showing larger values for disagreement attached to the ten categories cases.

As to ease of interpretation, *C* appears to be a *disagreement* index, rather than a «coefficient of correlation». A further source of ambiguity is the fact that its values range from 0.87 in case of maximum disagreement to 0.03 in case of maximum agreement for the four categories cases, from 0.95 to 0.06 in the ten categories cases - i.e. they never reach zero or unit, and vary their upper and lower end according to the number of categories involved. C_{cv} carries more or less the same amount of ambiguity as an agreement index, in that it varies between 0.97 in case of maximum agreement and 0.15 in case of maximum disagreement for the four categories cases, between 0.94 and 0.05 for the ten categories cases.

Statistics related to information theory have a clear, definite meaning in absolute terms of bits - binary units (*dh*), or in relative terms of «difference in information as a proportion of the maximum possible gap» (dh_{rel}). The first index varies between a maximum value of $\log_2 K$ (where K = number of categories) in cases of maximum disagreement, and zero in cases of maximum

agreement. The latter ranges from zero to one and is likely to be read as the proportion of the expected information actually observed.

Riassunto

L'articolo esamina criticamente le caratteristiche di alcune statistiche potenzialmente utili nelle verifiche di validità del contenuto: il chi quadro di Pearson, il coefficiente di contingenza C, il coefficiente di validità C_{cv} proposto da R. Hoste e due statistiche — *dh* e dh_{rel} — elaborate dall'A. con riferimento alla teoria dell'informazione. L'esame di casi tipici particolarmente rilevanti nell'esame della validità di contenuto mostra che le due statistiche d'informazione hanno significato meno ambiguo di quelle basate sul coefficiente di contingenza C, inducono meno ad errori di II tipo e sono più sensibili al numero di categorie implicato.

Resumé

L'emploi du coéfficient de contingence C pour mésurer la validité du contenu d'un test présente des inconvénients.

Des statistiques d'information envisagées par l'Auteur permettent une estimation plus claire et statistiquement «puissante» dans la majorité des situations.

Abstract

Content validity quantification is a complex problem, hardly amenable to a single statistical index. However, the particular problem of the goodness of fit of content samples to a set of expected frequencies can be properly quantified using non-parametric statistics for nominal data.

The paper discusses advantages and disadvantages of some statistics available for this purpose, and introduces a new index *dh* based on information statistics. The descriptive statistics *dh* is a measure of lack of fit. Its values range from a minimum $dh = 0$ in case of no difference between expected and observed frequencies to a maximum $dh = log_2 K$, where K = number of categories. It refers to a unit known as ***bit*** (***binary unity digit***), distributed along a logaritmic scale, and can be read as the difference in information between expected and observed proportions of items belonging to each content category.

A few typical cases are discussed, showing *dh* and the associated relative measure dh_{rel} as powerful indexes of content goodness of fit.

REFERENCES

— Attneave, F. (1959). *Application of information theory to psychology: a summary of basic concepts, methods and results*. New York: Holt.
— Calonghi, L. (1978). *Statistiche d'informazione e valutazione*. Roma: Bulzoni editore.
— Cronbach, L.J. (1971). Test validation. In *Educational Measurement* (2nd ed.) ed. R.L. Thorndike, Washington: American Council on Education), pp. 443-507.
— Hays, W.L. (1973). *Statistics for the Social Sciences* (2nd ed.). New York: Holt, Rinehart & Wiston.
— Hoste, R. (1981). How valid are school examinations? An exploration into content validity. *British Journal f Educational Psychology, 51*, 10-22.
— Kelley, T.L. (1923). *Statistical Methods*. New York: Macmillan.
— Kerlinger, F.N. (1964, 1973). *Foundations of Behavioral Research* (2nd ed.). New York: Holt, Rinehart & Winston.

— Mood, A.M., Graybill, F.A. (1963). *Introduction to the Theory of Statistics* (2nd ed.). New York: McGraw-Hill.
— Senders, V. (1958). *Measurement and statistics*. New York: Oxford University Press.
— Shannon, C.E., Weaver, W. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.
— Siegel, S. (1956). *Nonparametric Methods for the Behavioral Sciences*. New York: McGraw-Hill.

[Received October 18, 1983]