First published in *Educational and Psychological Measurement*, 1978, *38*, 81–88.

Domain Validity—Why Care?

Abstract

A strict equivalence presupposed by Kaiser and Michael to derive the coefficient of "domain validity" is defensible only as a biased approximation. But then, it is far from clear what psychometric significance this coefficient has in the first place.

In a recent issue of this journal, Kaiser and Michael (1975) derive the Tryon/Cronbach coefficient of "domain validity," or "generalizability" under charmingly weak assumptions. Their account is instructive, but not quite for the reason they propose.

Momentarily taking the received view of domain sampling at face value, consider two finite sets $\hat{\mathbf{X}}$ and $\mathbf{X}$ of test items, where $\hat{\mathbf{X}}$ is a subset (item sample) of $\mathbf{X}$ and the number n of items in $\mathbf{X}$ is very much larger than the number m of items in $\hat{\mathbf{X}}$. Define a subject's "observed score" to be the mean of his scores on the items in $\hat{\mathbf{X}}$ i.e. his score on the centroid $\hat{\mathbf{x}}$ of item-sample $\hat{\mathbf{X}}$, while his "domain score" is his mean score in $\mathbf{X}$, i.e. his score on the domain centroid $\mathbf{x}$. (I depart slightly from Kaiser and Michael here, since they define observed scores and domain scores as sums rather than averages. But this is merely a difference in scale unit except when n approaches infinity, in which case the domain sum is ill-defined.) As Kaiser and Michael point out, if $V_{\hat{x}}$ is the mean item variance in $\hat{\mathbf{X}}$, $C_{\hat{X}}$ is the mean covariance between (different) items in $\hat{\mathbf{X}}$, C_X is the mean covariance between items in $\mathbf{X}$ (or in $\mathbf{X} - \hat{\mathbf{X}}$, as preferred by Kaiser and Michael), and $C_{\hat{X}X}$ is the mean cross covariance between items in $\hat{\mathbf{X}}$ and items in $\mathbf{X}$ (or in $\mathbf{X} - \hat{\mathbf{X}}$), the squared linear correlation $R^2_{\hat{x}x}$ between the sample and domain centroids is

$$R^2_{\hat{x}x} \approx \frac{(mC_{\hat{X}X})^2}{[mV_{\hat{X}} + m(m-1)C_{\hat{X}}]C_X} \,, \tag{1}$$

where the approximation approaches an identity, given any positive lower bound on C_X, as n goes to infinity. Then if

$$C^2_{\hat{X}X} = C_{\hat{X}}C_X \qquad \text{(Kaiser-Michael assumption)}, \tag{2}$$

substituting (2) into (1) and dividing out m and C_X yields

$$\begin{aligned} \lim_{x\to\infty} R^2_{\hat{X}X} &= \frac{mC_{\hat{X}}}{V_{\hat{X}} + (m-1)C_{\hat{X}}} \\ &= \alpha_{\hat{X}} \end{aligned} \tag{3}$$

where $\alpha_{\hat{X}}$ is the classic Alpha coefficient for item set $\hat{\mathbf{X}}$.

Actually, this result can be established with even greater ease and generality than given by Kaiser and Michael. For any two finite, not-necessarily-disjoint sets $\mathbf{X}$ and $\mathbf{Y}$ of test items, let $\mathbf{x}$, V_x, C_X, and α_X ($\mathbf{y}$, V_y, C_Y, and α_Y) be respectively the centroid, centroid variance, mean between-item covariance, and Alpha coefficient for set $\mathbf{X}$ (set $\mathbf{Y}$), while C_{XY} is the mean cross covariance between items in $\mathbf{X}$ and in $\mathbf{Y}$. It is easily seen[1] that $\alpha_X = C_X/V_x$, whence $V_x = C_X/\alpha_X$. Similarly $V_y = C_Y/\alpha_Y$, while the covariance C_{xy} between centroids $\mathbf{x}$ and $\mathbf{y}$ is obviously equal to C_{XY}. Hence the squared correlation between the centroids is

$$R^2_{xy} = \frac{C^2_{xy}}{V_x V_y} = \alpha_X \alpha_Y \left(\frac{C^2_{XY}}{C_X C_Y} \right). \tag{4}$$

To get (3) from (4) and (2), we need only take $\mathbf{Y}$ to be a subset $\hat{\mathbf{X}}$ of $\mathbf{X}$, yielding

$$\begin{aligned} R^2_{\hat{x}x} &= \alpha_{\hat{X}} \alpha_X \left(\frac{C^2_{\hat{X}X}}{C_{\hat{X}} C_X} \right) \\ &= \alpha_{\hat{X}} \alpha_X \quad \text{if} \quad C^2_{\hat{X}X} = C_{\hat{X}} C_X, \end{aligned} \tag{5}$$

and observe that with the item homogeneity of domain $\mathbf{X}$ held parametrically constant at any positive value, α_X monotonically increases to unity as the number of items in $\mathbf{X}$ becomes arbitrarily large (cf. Rozeboom, 1966, Fig. 8.1).

But under that circumstances is (2) a reasonable assumption? The answer, quite simply, is in all likelihood *never*, not even as a statistical expectation. If the concept of "domain sampling" makes any sense at all, we must suppose that item set $\hat{\mathbf{X}}$ is obtained from domain $\mathbf{X}$ under circumstances that justify envisioning a probability distribution for the parameters of item samples so drawn from $\mathbf{X}$, including in particular sampling expectations $Exp[C_{\hat{X}}]$ and $Exp[C_{\hat{X}X}]$ for item-sampling covariance parameters $C_{\hat{X}}$ and $[C_{\hat{X}X}]$, respectively, and a sampling variance $Var[C_{\hat{X}X}]$ for $[C_{\hat{X}X}]$. Since C_X is the population parameter to which both $C_{\hat{X}}$ and $C_{\hat{X}X}$ converge as m appoaches n, both $Exp[C_{\hat{X}}]$ and $Exp[C_{\hat{X}X}]$ should approximately equal C_X; but in any case we have

$$Exp[C_{\hat{X}}] = C_X + a, \qquad Exp[C_{\hat{X}X}] = C_X + b$$

where a and b are sampling biases that will be essentially zero in any reasonable domain-sampling model. Then

$$Exp[C_{\hat{X}} C_X] = (C_X + a) C_X$$

whereas

$$Exp[C^2_{\hat{X}X}] = Exp[C_{\hat{X}X}]^2 + Var[C_{\hat{X}X}] = (C_X + b)^2 + Var[C_{\hat{X}X}];$$

[1] E.g. from equation (9.52) by definitions (7.75, 7.85) in Rozeboom (1966).

so the expected difference between the sides of (2) is

$$\begin{aligned} (6) \qquad Exp[C^2_{\hat{X}X} - C_{\hat{X}}C_X] &= C_X(2b-a)+b) + Var[C_{\hat{X}X}] \\ &= Var[C_{\hat{X}X}] \text{ if } a = b = 0 \end{aligned}$$

The quantity $Var[C_{\hat{X}X}]$ in (6) is a decreasing function of the number of items in $\hat{\mathbf{X}}$, but is greater than zero for all finite m unless all item covariances in $\mathbf{X}$ are identical. Hence apart from an extraordinary fluke of sampling biases, the lefthand side of (2) is expected to be somewhat larger than its righthand side.

I wish I could polish this point by exhibiting the sampling-expectation of $R^2_{\hat{x}x}$, or better, its expectation conditional upon the observed value of $\alpha_{\hat{x}}$ and perhaps other observable parameters of $\hat{\mathbf{X}}$. Unfortunately, the righthand side of (5) is too complex for the sampling behavior of $R^2_{\hat{x}x}$ to be readily discerned even in the most ideal sampling model for $\hat{\mathbf{X}}$. Just the same, the systematic tendency of $C^2_{\hat{X}X}$ to exceed $C_{\hat{X}}C_X$ makes it difficult to imagine that $\alpha_{\hat{X}}\alpha_X$ can ever be more than an imperfect, biased estimate of $R^2_{\hat{x}x}$.

To be sure, as an approximation to $R^2_{\hat{x}x}$, $\alpha_{\hat{X}}$ should be more than adequate in all applied circumstances. But do any such circumstances in fact exist? Despite the recent test-theoretic prominence of domain sampling notions (most importantly, in Cronbach, Gleser, Nanda, & Rajaratnam, 1972, and Lord & Novick, 1968), one may seriously question whether these really have any practical point, at least with their classical centroid focus.

In the first place, given an extant set $\hat{\mathbf{X}}$ of test items, in what way can we meaningfully conceive of an item domain $\mathbf{X}$ sampled by $\hat{\mathbf{X}}$? Is $\mathbf{X}$ some part of the finitely many stimuli which have been/will be *in fact* brought forth as a "test item" in the past/present/future history of psychometric practice? This would be an appropriate way to view $\mathbf{X}$ if its items were indexed e.g. by an array of physical tokens and $\hat{\mathbf{X}}$ were constructed by drawing a subset of these; but since item selection is virtually never remotely like that, I shall assume without argument that this is not what domain-sampling theorists have had in mind. Surely, domain $\mathbf{X}$ is envisioned not as some set of items actually constructed, but as a set which in some subjunctive sense *could* be generated in a determinate fashion G_X.[2] But if so, how do we characterize the parameters of $\mathbf{X}$ and its domain centroid?

The traditional approach, construing the domain parameters as derivative from the joint distribution of distinct items in $\mathbf{X}$, turns out under critical examination to

[2] That particular item domains are seldom if ever well-defined has been a frequent objection to domain-sampling theory. Strictly speaking, however, this criticism is misdirected. What domain-sampling theory needs to specify in any attempted practical application is not a set of items as such but a particular method of item *production.* (To be sure, that is never accomplished either; so with a little re-targeting, the standard complaint about domain indeterminacy remains as cogent as ever.)

be unworkable. Ontologically, it is questionable whether it makes sense to think of a stimulus that is never actually manifested in some setting as existing at all; but if all items that *could* be generated in fashion G_X *do* literally "exist" in some way that allows them to comprise an abstraction basis for the domain parameters, there will be an uncountable infinity of them unless G_X is defined to be so intrinsically sequential that not only each item in fact produced by G_X but also every one that *could* be has a distinct order index assigned to it by G_X. Now if $\mathbf{X}$ is uncountable, the mean between-item covariance and the domain centroid cannot be defined by limits at all; and this remains true even if $\mathbf{X}$ is a countably infinite sequence unless we postulate convergence properties for the sequence that are unwarranted unless inferred from a more fundamental characterization of G_X (see immediately below) and even then place no constraints on the item parameters of any finite segment of the sequence. In any case, returning to the first point, so far as we know there are no literally infinite sequences of G_X-generated items.

Alternatively, we can forego pretending that domain $\mathbf{X}$ is a set or sequence of real items by appealing, instead to an array of causal parameters that probabilistically determine the test-theoretic character of items generated in fashion G_X. Such parameters cannot be identified in terms of the (fictitious) items in $\mathbf{X}$; but there is no evident reason why there may not exist—really exist, even if not accessible to direct observation—a space of source-factors within which lie all test items that can be generated by G_X. The G_X may be thought to dispose a joint probability distribution for the factor loadings of item tuples produces in fashion G_X is a probability-theoretic "sample" from this distribution. (See Hunter, 1968, for an elegant introduction to this approach. If one wants, the sampling theory for the items' factor loadings can be developed with a model of item "facets". Also, factor models more general than the standard linear one are entirely admissible in principle albeit probably not useful to exploit at present.) A subject's score on the (m-wise) "domain centroid" may then be defined as the score statistically expected for that subject on the centroid of an item m-tuple generated in fashion G_X.[3] Given a reasonably well-behaved probability model for G_X—most ideally, with all items in $\hat{\mathbf{X}}$ drawn independently with the same marginal factor-loading probabilities—we can investigate the sampling behavior of $\hat{\mathbf{X}}$'s parameters in considerable technical depth.

Domain sampling does, then, appear to be a workable idea, even if more de-

[3] If G_X generates item m-tuple $\mathbf{X}$ with the same marginal probability distribution for the factor coefficients of each item, the domain centroid is simply the expectation for any one item generated in fashion G_X, and the domain parameters readily generalize to G_X-generated item tuples of any length. Otherwise, the m-wise domain centroid's factor structure is a nontrivial function of m; and generalization from an observed m-tuple of G_X-generated items to a tuple of length other than m—or even to length m ones as well, since one can argue that these are all different segments of the same sequence—becomes much more tenuously complicated.

manding of technical sophistication than the past lieteature would suggest.[4] But what practical point might there be in seeking to learn the domain centroid's correlation $R_{\hat{x}x}$ with the centroid of an extant G_X-generated item sample $\hat{\mathbf{X}}$? There seems to be essentially just two kinds of things that one might try to infer from $\hat{\mathbf{X}}$-data. (Cf. the G-study/D-study distinction in Cronbach et al., 1972.) One is the probability parameters of G_X-wise item production, and from there, if one cares, what to expect about the psychometric character of additional item tuples so produced. And secondly, one may have the standard psychometric desire to infer a subject's standing on one or more criterion variables from his scores on $\hat{\mathbf{X}}$. Regarding the first objective, if our item-sampling model is reasonably simple and the number m of items in $\hat{\mathbf{X}}$ is appreciable, quite a bit can be learned about this through analyzing the variance structure of $\hat{\mathbf{X}}$ in a suitably large subject-sample. But $R_{\hat{x}x}$, no matter how accurately estimated, tells virtually nothing about the domain parameters. $R_{\hat{x}x}$ is just a monotone function of m whose rate of approach to unity is modulated by domain parameters that can be estimated from the internal properties of $\hat{\mathbf{X}}$ without concern for $R_{\hat{x}x}$.

On the other hand, if one wishes to make criterion predictions for particular subjects, $R_{\hat{x}x}$ does tell how accurately a subject's domain-centroid score can be estimated from his sample-centroid score. But the domain centroid is just one dimension of what is generally a multidimensional space of item factors, all of which can be predicted from $\hat{\mathbf{X}}$ to one degree or another, so why should we care more about this one than about the others? To be sure, the domain centroid will almost always well-approximate the dimension that accounts for the maximum variance in the expected structure of a G_X-generated item m-tuple; but if the latter is what turns us on, we have more sophisticatedly accurate ways to get at that than through concern for $R_{\hat{x}x}$. The linear composite of an observed G_X-generated $\hat{\mathbf{X}}$ with maximum "generalizability" in $\mathbf{X}$ is not $\hat{\mathbf{X}}$'s centroid but its first principle axis. In any event, our purpose in generating test items is presumably to have something to predict *from*, not *to*. What subject attributes underlie responding to an array of test items is naturally a matter of psychonomic concern, and analysis of the variance structure within an extant item array can be invaluably diagnostic of that; but it is far from obvious that the psychonomic importance of an item factor is determined by how high a proportion of G_X-generated item variance it

[4] Complications of sampling models are less than ideally simple (cf. footnote 3), need for which has already surfaced in domain-sampling considerations of "stratified tests," seem basically amenable to standard probability-theoretic treatment. More enigmatic, however, is the not-heretofore-appreciated problem that arises if we are unwilling to assume—and there are good reasons for not wishing to assume this—that the factor space within which items are generated has finite dimensionality. For then we need to cope with infinite sums under conditions to which the traditional theory of limits does not cogently apply. (See Rozeboom, 1978, "The logic of unboundedly reactive systems," for a detailed study of startlingly large indeterminacies that remain in the mathematics of infinite sums and other infinite concatenations.)

accounts for.

I can think of just one psychometric application for which centroid-focused domain sampling theory has some possible value, namely, cases where we wish to interpret the score of a subject on an m-tuple of $\hat{\mathbf{X}}$ G_X-generated test items when we do *not* have normative data on the specific set $\mathbf{X}$ but *do* know (approximately, through analysis of normative data on other G_X-generated items) the item-generative parameters of G_X In this case, our subject's score on the sample centroid may be viewed as his observed score on the "generic" test $\mathbf{X}_g$ whose true-score component is the domain centroid and whose measurement-error component includes the discrepancy between sample centroid and domain centroid scores (cf. Lord & Novick, 1968). If, after Cronbach, we define the (m-wise) "coefficient of generalizability" $\alpha_{X(m)}$, of domain $\mathbf{X}$ to be the Alpha coefficient of a perfectly representative m-tuple $\hat{\mathbf{X}}$ of G_X-generated items—i.e., an $\hat{\mathbf{X}}$ for which $C_{\hat{X}} = Exp[C_{\hat{X}}]$ and $V_{\hat{X}} = Exp[V_{\hat{X}}]$, so that $R^2_{\hat{x}x} = \alpha_{X(m)}$—it can be shown that the reliability coefficient r_{X_g} of generic test $\mathbf{X}_g$, assuming independent item sampling albeit not necessarily the same marginal item probabilities, is

$$r_{X_g} = \alpha_{X(m)}\left(\frac{1}{1+e}\right), \quad e =_{\text{def}} \frac{\alpha_{X(m)} Var[\mu_{\hat{X}}]}{m\, Exp[C_{\hat{X}}]}\ , \tag{7}$$

in which $Exp[C_{\hat{X}}]$ is the expected average covariance within a G_X-generated m-tuple of test items and $Var[\mu_{\hat{X}}]$ the average sampling variance of item means for such tuples. (7) shows that the coefficient of generalization, or its estimate from an observed item m-tuple, is a somewhat biased approximation to the reliability of the corresponding m-wise generic test. However, the likely bias in this approximation due to $Var[\mu_{\hat{X}}]$ is too large to be ignored—nor needs be, since all terms in (7), not just $\alpha_{X(m)}$, estimated from sample data (albeit at least two item m-tuples are needed to estimate $Var[\mu_{\hat{X}}]$ if not all positions in the tuple have the same marginal item probabilities).

As a protest against assumption (2), the foregoing may well seem excessive. But my target is rather more serious than that. The history of test theory has repeatedly shown a proclivity to institutionalize theoretical fantasies in which some mathematical simplicity is embellished with little concern for what relevance, if any, it may have to the real world. I have no quarrel with such mathematical games in their own right-they are an inexpensive, joyful sport that at times can even be conceptually illuminating. But I do urge that this not be mistaken for serious analysis of foundational issues. As demonstrated by Cronbach et al. (1972), the notion of domain sampling can be given a breadth of vision that potentiates it for world-class standing as a theory of scientific observation. But that potential lies first of all in the sophisticated framework it affords for the conduct of searching inquiry into the nature and interpretation of measurements. In this note, I have tried to unstick domain-sampling presuppositional orthodoxies by arguing, overbriefly, (1) that the parameters of an item domain cannot cogently be defined as limiting values of observable item-sample properties; and (2) that the domain centroid's psychonomic and psychometric importance is at best problematic, while that of its correlation with the centroid of an observed item-sample is even more so. I submit these remarks as a challenge to re-think just what *are* the significant questions for domain-sampling theory to address.

References

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Hunter, J. E. (1968). Probabilistic foundations for coefficients of generalizability. *Psychometrika*, *33*, 1–18.

Kaiser, H. F., & Michael, W. B. (1975). Domain validity and generalizability. *Educational and Psychological Measurement*, *35*, 31–35.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.

Rozeboom, W. W. (1966). *Foundations of the theory of prediction.* Homewood, Illinois: The Dorsey Press.

Rozeboom, W. W. (1978). The logic of unboundedly reactive systems. *Synthese*, *39*, 435–530.