Instituto Superior de Economia e Gestão

VALIDADE E FIABILIDADE EM ESTUDOS COM DADOS RECOLHIDOS POR QUESTIONÁRIO: UMA DIGRESSÃO PELA LITERATURA

*Alberto Ferreira Pereira**

Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa

Resumo

Recorrendo a uma breve digressão pela literatura, este artigo aborda as questões da validade e da fiabilidade no contexto da utilização de questionários em estudos de gestão e sublinha a sua importância para a obtenção da indispensável credibilidade metodológica que deve caracterizar um estudo científico. Distingue os conceitos de fiabilidade entre itens e entre classificadores e descreve alguns métodos utilizados na sua medição. Este artigo não esgota o tema, porém, evidencia os cuidados metodológicos mínimos a serem tomados em consideração sempre que se utilizam questionários.

Introdução

Parte substancial do trabalho de investigação realizado com vista à obtenção de graus de pós graduação que temos orientado ou arguido, recorre a questionários para a recolha de dados. Os respondentes são diversos: trabalhadores de empresas cujo desempenho em relação a uma ou mais variáveis se pretende estudar; clientes de um dado serviço cujos factores descritivos de qualidade se pretendem determinar; proprietários de empresas ou gestores junto de quem se pretende coligir elementos que permitam perceber mecanismos de actuação e processos de decisão complexos como o estabelecimento de uma estratégia de negócio, ou ainda potenciais investidores cujo perfil de empreendedor se deseja perscrutar. Esta constatação é transversal a trabalhos em áreas tão diversas quanto a estratégia, o marketing ou as operações.

Alguns autores designam por *inquérito* o instrumento elaborado para recolha de dados. Preferimos a designação de *questionário*. Aquela designação tem uma forte conotação com práticas associadas ao uso de meios de coacção, legal ou outra. Esta, não só está desprovida daquela carga como tem associada a ideia de

* O autor agradece as sugestões de dois revisores anónimos que muito contribuíram para melhorar este artigo.

participação. Meredith, Raturi, Amoako-Gyampah e Kaplan (1989, pp. 300) referem explicitamente o uso de *questionários*.

Este estudo discorre acerca de algumas questões centrais relacionadas com a utilização de questionários, em particular a validade e a fiabilidade e sublinha a sua importância para a obtenção da indispensável credibilidade metodológica que deve caracterizar um estudo científico.

A investigação na área da estratégia de operações registou um progresso significativo no decurso da última década tanto em termos de quantidade como de qualidade dos artigos publicados (Boyer e Verma, 2000). Meredith, Raturi, Amoako-Gyampah e Kaplan (1989) descrevem as etapas mais importantes do percurso da investigação na área de operações (OM), suas vicissitudes e desafios. As duas etapas mais notáveis deste percurso são, primeiro, a brusca reorientação da área para uma abordagem analítica baseada em modelos quantitativos e, em segundo lugar, a constituição da OMA (Operations Management Association) e da OMA-United Kingdom e o início da publicação dos seus respectivos *journals*. A constituição de sociedades profissionais e de publicações especializadas produziram efeitos igualmente notáveis no desenvolvimento da área da gestão estratégica (Rumelt, Schendel e Teece, 1998, 27-32).

Para a área das operações, Meredith, Raturi, Amoako-Gyampah e Kaplan (1989) apresentam um quadro de referência genérico para a classificação de paradigmas que se baseia no quadro de referência que Mitroff e Mason (1984) desenvolveram para a política de gestão. As duas dimensões propostas por estes foram redefinidas por aqueles para melhor acomodar a área das operações tendo assim resultado as dimensões *racional/existencial* e *natural/artificial*. A primeira dimensão respeita à estrutura epistemológica do processo de investigação e a segunda diz respeito à fonte e tipo de informação utilizada na investigação. A percepção que o investigador tem da realidade é moldada pelos mecanismos utilizados para estudar o fenómeno de interesse. Podem, genericamente, classificar-se em três categorias: observação directa do objecto realidade; percepções das pessoas acerca do objecto realidade; e reconstrução artificial do objecto realidade. Os questionários, tal como as entrevistas ou muitas experiências laboratoriais, são instrumentos utilizados na segunda categoria para conduzir investigação através "dos olhos de outros". A preocupação, neste caso, é a percepção ou representação abstracta da realidade dos *indivíduos* expostos ao fenómeno.

Os questionários, tal como as entrevistas estruturadas, possibilitam posterior análise estatística. Possibilitam uma utilização mais eficiente do tempo quando comparadas com as entrevistas sobretudo quando se tem de lidar com a distância física dos respondentes uma vez que, se bem desenhado, o questionário pode ser enviado a um grande número de pessoas sem grandes dificuldades. Podem igualmente predeterminar-se amostras estratificadas para a avaliação de factores de interesse particular. Por esta razão, este método é extremamente popular na co-

munidade académica. Porém, apresenta desvantagens no facto de só uma fracção dos questionários ser devolvida e na possibilidade de que algumas "respostas" poderem ser de pouco valor. Pode deixar-se espaço para comentários livres mas não há a possibilidade de alterar as perguntas do questionário por forma a contemplar os comentários.

Validade e Fiabilidade

Existem duas questões críticas associadas à utilização de questionários. São, respectivamente, a validade e a fiabilidade. Ambas as questões estão intrinsecamente relacionadas entre si. A grande maioria dos estudos tende a enfatizar a fiabilidade devido provavelmente a preocupações de replicabilidade e de consistência. Aqueles dois conceitos não são, no entanto, a mesma coisa. Segundo Emory (1985, pp. 94-98) elevada fiabilidade pode bem estar associada com validade zero. Rungtusanatham (1998) sublinha igualmente a importância de que se devem revestir as questões de validade dos conceitos teóricos (*constructs*) e de fiabilidade na condução de investigação empírica em OM referindo mesmo (pp. 10) que se tornou norma indicar nos artigos publicados em *journals* ou em conferências de OM as avaliações de fiabilidade e de validade dos conceitos teóricos contidos nos instrumentos de medida.

Validade

Definir conceitos de extrema utilidade em estudos de gestão como são os de automatização (Ritzman e Safizadeh, 1999), práticas de gestão de qualidade (Ahmad e Schroeder, 2001), ou participação dos gestores de produção na formulação de estratégia (Tracey, Vonderembse e Lim, 1999) bem como medi-los, ilustram bem a natureza do problema da validade do conceito teórico. Ao contrário do que sucede com os bens tangíveis que podem ser medidos de forma objectiva através de indicadores como a durabilidade ou o número de defeitos, muitos conceitos em gestão são abstractos, intangíveis, heterogéneos e inseparáveis do contexto em que têm lugar, inviabilizando a sua observação e medição fora deste.

Jacoby (1978) citado por Parasuraman, Zeithaml e Berry (1988) refere a inadequabilidade de procedimentos de medida utilizados na disciplina de marketing dizendo que "muitas das nossas medidas são desenvolvidas por investigadores de forma ligeira sem uma reflexão sobre se estão relacionadas de maneira significativa a formulações conceptuais explícitas do fenómeno ou variável em questão. Em muitos casos os nossos conceitos não têm identidade quando destacados do instrumento ou dos procedimentos usados para medi-los".

Na mesma linha de preocupação, Corbett e Van Wassenhove (1993) citados por Noble (1995) afirmam que "a falta de definições de conceitos fundamentais geralmente aceites como os frequentemente utilizados de "prioridades competitivas" de custo, qualidade, fiabilidade, flexibilidade e inovação ... resulta numa falta de critérios de medida de operacionalização úteis para facilitar o trabalho empírico" (pp. 700). Embora os indicadores simples limitem a generalização por incapacidade de capturar a riqueza e complexidade das prioridades competitivas da produção, não existe uma via directa para combinar indicadores múltiplos que sustentem um conceito multidimensional, especialmente quando cada indicador mede uma subdimensão diferente do conceito e não o conceito no seu todo.

Na ausência de medidas objectivas, uma abordagem apropriada para medição é a utilização de um conjunto de indicadores descritivos do conceito. É frequente proceder-se ao desenvolvimento de conceitos a partir de duas ou mais variáveis mensuráveis. Neter, Wasserman e Kutner (1985) referem, a propósito da discussão da selecção de variáveis, que em muitas áreas como a das ciências sociais, comportamentais e na gestão, raramente existem modelos teóricos prontos a serem utilizados. Não existe um conjunto único de variáveis para descrever um dado conceito. Os conjuntos de variáveis utilizados para a descrição de um conceito podem diferir entre si não apenas pela natureza das variáveis como pelo número de variáveis utilizadas. Isto é, a natureza de cada uma das questões e o número de questões utilizadas reflectem a maior ou menor abrangência do domínio da definição teórica do conceito, a maior ou menor capacidade de captar as diferentes facetas do conceito. Em teoria, dizemos haver validade do conceito teórico se as variáveis seleccionadas constituem um subconjunto escolhido aleatoriamente do universo de variáveis representativas do domínio completo do conceito. Dito de outra forma, questionar a validade de conteúdo é equivalente a responder à questão: "É a substância desta medida [o conjunto das variáveis] representativa do conteúdo ou universo do conteúdo do conceito que se está a medir?" Para responder a esta questão assumimos a conveniência e a possibilidade de podermos especificar e seleccionar de forma aleatória uma amostra do universo de variáveis que constituem o domínio do conceito. Raramente, porém, isto é conseguido na prática. Por isso, as avaliações da validade de conteúdo se baseiam em "apelos à razão com respeito à adequabilidade com que o conteúdo foi amostrado e também à adequabilidade com que o conteúdo é reproduzido pelas variáveis" (Rungtusanatham, 1998).

A revisão da literatura constitui uma fonte de inestimável valor na identificação das questões cruciais para operacionalizar os conceitos teóricos de interesse para o estudo (Kettinger e Grover, 1997, pp.516; Mirani e Lederer, 1998; Tracey, Vonderembse e Lim, 1999). Uma revisão da literatura bem orientada e cuidada proporciona sempre a possibilidade de confrontar os conceitos teóricos que nos propomos estudar com os existentes na literatura. Kettinger e Grover (1997, pp.

524) referem, no estudo sobre o uso de comunicações através de computador em contexto inter-organizacional, que para desenhar o instrumento [questionário] cuidaram de adaptar medidas validadas já existentes para os conceitos teóricos e que nos casos em que tais medidas não estavam disponíveis, os itens [questões] foram construídos com base na literatura. Segars, Grover e Teng (1998, pp. 315) indicam que é "através de uma revisão intensiva da literatura" que se consegue a especificação do domínio do conceito teórico devendo o investigador ser rigoroso ao delinear o que se inclui e o que não se inclui na definição do conceito teórico sob investigação. Deste modo, os autores definem 37 itens para operacionalizar 7 características de planeamento não observáveis (conceitos teóricos): globalidade do planeamento; formalização; foco; fluxo de planeamento; participação; consistência e eficácia de planeamento. Germain e Dröge (1997, pp.623; 632-635) recorreram igualmente à revisão da literatura para definir 67 itens para medir as oito variáveis de interesse para o estudo empírico do impacto da amplitude de tarefas JIT *versus* integração de fluxos de trabalho JIT no design organizacional: amplitude de tarefas JIT; integração de fluxos de trabalho JIT; especialização; descentralização; controlo interno; controlo de benchmark; comités de integração e mecanismos de integração. Desta comparação entre o proposto e o existente na literatura emergirá um conjunto de questões que, através de aplicações e melhorias sucessivas, tenderá a ser universalmente aceite pela comunidade científica e, assim, estabelecer-se um conjunto padrão que tornará as comparações mais credíveis. Goodhue (1998) procede ao desenvolvimento conceptual de um instrumento para avaliar a globalidade dos sistemas de informação e de serviços nas organizações, apresenta um tratamento detalhado da validade usando uma amostra de 357 gestores em 10 organizações e procede à comparação do instrumento desenvolvido com os propostos por outros autores.

Este confronto constitui portanto uma via indispensável para que os itens de um questionário possuam uma base de apoio com fundamento e não sejam o resultado de mero capricho do investigador orientado pelo seu *feeling*. Finalmente, uma outra vantagem é a que permite sugerir melhorias no conjunto de questões que são tradicionalmente utilizadas para operacionalizar este ou aquele conceito teórico. O recurso ao contributo de especialistas de empresas ou académicos na formulação das questões para captar um dado conceito é também frequente (Dow, Samson e Ford, 1999; Tracey, Vonderembse e Lim, 1999).

No estudo que avalia em que medida diferenças culturais influenciam as percepções que gestores mexicanos e suecos têm da relação entre uso de sistemas de informação para executivos e diversas variáveis relacionadas com a tomada de decisão, Leidner, Carlsson, Elam e Corrales (1999) desenvolvem um questionário constituído por 20 questões para operacionalizar 5 conceitos teóricos.

No estudo sobre a existência de relações entre determinantes da gestão da qualidade hospitalar e o desempenho em relação à qualidade do serviço, Li (1997)

recorre à utilização de múltiplas questões sobre o mesmo conceito a fim de obter uma melhor aproximação deste. Por exemplo, para a obtenção de uma aproximação do conceito de "análise de informação para desenvolvimento contínuo", o autor utiliza 7 questões. Os restantes cinco conceitos aproximados, cada um deles através de conjuntos de diferentes questões são, respectivamente, liderança da gestão de topo; cooperação organizacional; liderança tecnológica; desenvolvimento da força de trabalho; e desempenho em relação à qualidade do serviço.

Uma vez identificadas as variáveis descritivas do conceito que se pretende avaliar sob a forma de um conjunto de questões extraídas da literatura, sugeridas pelo investigador, por académicos, consultores ou por gestores de empresas, executam-se em seguida algumas medidas com vista a racionalizar aquele conjunto. Uma primeira medida visa refinar a medição do conceito, isto é, verificar analiticamente se o conceito de interesse pode ser descrito por um número mais reduzido de variáveis (questões). Para determinar os itens a serem eliminados, a correlação entre o *score* do item e a soma dos *scores* de todos os outros itens descritivos do conceito é o critério mais frequentemente reportado na literatura (Parasuraman, Zeithaml e Berry, 1988; Tracey, Vonderembse e Lim, 1999). Os itens com correlação inferior a um valor estabelecido como referência são eliminados. Com os itens não eliminados determina-se o coeficiente alfa de Cronbach (1951) que se descreve adiante na discussão sobre fiabilidade. Em seguida, procede-se ao exame da dimensionalidade das questões remanescentes após a refinação recorrendo à análise factorial. O objectivo da análise factorial é o de, se possível, expressar as relações de covariância entre muitas variáveis em termos de poucas *quantidades* aleatórias não observáveis chamadas factores que lhes estão subjacentes (Johnson e Wichern, 1988, pp.378)

Para Flynn, Schroeder e Sakakibara (1995) a validade de um conceito teórico respeita à questão de saber em que medida todos os itens de uma escala medem o mesmo conceito teórico. Para o efeito recorrem à análise de factores para cada uma das escalas testando se todos os itens da escala têm pesos num factor comum. Todos os itens de todas as escalas contribuem para as respectivas escalas por excederem o peso mínimo aceitável de + 0,40.

Para Leidner, Carlsson, Elam e Corrales (1999) a validade de um conceito teórico refere-se à questão de saber se os conceitos são reais na forma como são medidas ou se constituem apenas artefactos metodológicos. Leidner *et al.* (1999) estabelecem três critérios de pesos para incluir os itens nos factores determinados pela análise factorial a que submeteram o questionário de 20 itens: ter um peso no factor de pelo menos 0,5 mas inferior a 0,3 noutro factor; ser conforme com afectações anteriores; e contribuir para a fiabilidade da variável. Após a aplicação dos critérios de selecção referidos, as 20 questões reduzem-se a 14, com pesos nos factores de que fazem parte oscilando entre 0,60 e 0,92.

A validade de um instrumento desenhado para medir um dado conceito teórico reflecte a medida em que o instrumento cobre o domínio da definição do conceito teórico, isto é, em que medida o instrumento captura as diferentes facetas do conceito teórico. Em teoria, um instrumento de medida desenhado para medir um conceito teórico específico tem validade se os itens do instrumento são um subconjunto escolhido aleatoriamente do universo de itens que representam a totalidade do domínio do conceito.

Fiabilidade

Indissociavelmente ligado à validade do conceito está a questão da fiabilidade. Cronbach (1951) no seu estudo sobre a estrutura interna dos testes refere, numa breve digressão histórica, que qualquer investigação baseada em medições deve ter em conta a precisão ou fiabilidade da medida. E, um pouco sarcasticamente, sublinha que mesmo aqueles investigadores que vêem a fiabilidade como uma sombra pálida da questão mais importante da validade, não podem evitar considerar a fiabilidade das suas medições. Um instrumento de medida diz-se fiável quando as medidas com ele realizadas estão isentas de erro e produzem resultados consistentes.

O questionário é, por excelência, o instrumento mais amplamente utilizado para medir conceitos de interesse para a investigação em diversas áreas da gestão. A fiabilidade do instrumento é uma questão central na condução de trabalhos de investigação que utilizam o questionário para a recolha de dados. Sempre que os dados são coligidos por questionário espera-se que na descrição metodológica do trabalho de investigação se encontrem referências explícitas à sua construção e à operacionalização dos conceitos teóricos relevantes. Referências à questão da fiabilidade são facilmente encontradas na literatura (Bates, Amundson, Schroeder, e Morris, 1995; Flynn, Schroeder e Sakakibara, 1995; Kikulis, Slack e Hinings, 1995; Noble, 1995; Kettinger e Grover, 1997; Williams e Wilson, 1997; Brandyberry, Rai e White, 1999; Frohlich e Dixon, 1999; Leidner, Carlsson, Elam e Corrales, 1999; Li, 1997; Narasimhan e Das, 1999; Wilson e Collier, 2000). Distinguem-se dois conceitos de fiabilidade: entre itens e entre classificadores.

Fiabilidade entre itens

A medida standard para a fiabilidade entre itens é o coeficiente alfa (α) proposto por Cronbach (1951):

$$\alpha = \left(\frac{n}{n-1}\right)\frac{\sum_i \sum_j C_{ij}}{V_t}$$

com $i, j = 1, 2, \ldots n; i \neq j$. Na fórmula, n representa o número de itens, C_{ij} a covariância de dois itens i e j e V_t a variância total. O segundo termo do produto é o rácio da covariância entre os itens pela variância total. Consideram-se como adequados para o coeficiente alfa, valores iguais ou superiores a 0,70.

O questionário é, portanto, um conjunto de questões convenientemente agrupadas para captar a essência dos conceitos de interesse. Um conceito não é necessariamente melhor captado por ser maior o número de questões elaboradas. Questões que inicialmente se consideram adequadas para a descrição de um conceito podem ser eliminadas se se verificar que o seu contributo explicativo é fraco.

No estudo empreendido para o desenvolvimento de um questionário, Parasuraman, Zeithaml e Berry (1988) partiram de um conjunto inicial de 97 questões elaboradas para descreverem 10 conceitos relevantes na área da qualidade de serviços. O conjunto inicial de questões tinha valores de alfa entre 0,55 e 0,78. Após sucessivas iterações, a descrição dos mesmos 10 conceitos pôde ser feita com apenas 54 questões e valores de alfa entre 0,72 e 0,83. Flynn, Schroeder e Sakakibara (1995), usam fiabilidade como medida de consistência interna medida pelo coeficiente alfa de Cronbach. A matriz de correlação entre os 87 itens que descrevem as 10 variáveis de interesse para estudar o impacto das práticas de gestão de qualidade no desempenho e na vantagem competitiva é utilizada para eliminar os itens que não contribuem de maneira significativa para o alfa e cujo conteúdo não é crítico. Os valores de alfa para todas as escalas excedem o mínimo aceitável de 0,60, alguns de forma substancial indicando assim que as escalas são internamente consistentes.

Leidner et all (1999) utilizam também o coeficiente alfa para avaliar a fiabilidade entre os itens seleccionados para a escala final desenhada para operacionalizar os conceitos de disponibilidade de informação; modelo mental; velocidade na tomada de decisão; grau de profundidade da análise na tomada de decisão; e envolvimento dos subordinados. No estudo sobre gestão da qualidade hospitalar, Li (1997) reporta valores entre 0,73 e 0,91 para o coeficiente alfa utilizado para medir a fiabilidade entre os itens descritivos dos 6 conceitos de interesse.

Ritzman e Safizadeh (1999) utilizaram igualmente questionários junto de 144 empresas para examinar como processos de produção se alinham com decisões de design com referência aos recursos humanos (trabalhadores directos e indirectos; job enlargment; job enrichment; formação) e a recursos de capital (equipamento; nível de automatização; flexibilidade do equipamento; utilização de capacidade; manutenção preventiva). Para os conceitos teóricos de

automatização, de *job enlargement* e *job enrichment* os autores reportam valores do coeficiente alfa de Cronbach de 0.94, 0.99 e 0.99 respectivamente.

Fiabilidade entre classificadores

Para a fiabilidade entre classificadores não existe consenso para identificação da melhor medida. Boyer e Verma (2000) sublinham a necessidade de distinguir entre acordo (*agreement*) e fiabilidade entre classificadores (*inter-rater reliability*). Para Kozlowski e Hattrup, citados por Boyer e Verma (2000, pp. 131), "fiabilidade refere-se a um índice de consistência; refere consistência proporcional de variância entre classificadores e é por natureza correlacional", enquanto que "acordo refere-se às trocas entre classificadores; aborda em que medida os classificadores produzem essencialmente as mesmas classificações".

Apesar do cuidado metodológico que se deve conferir à questão da fiabilidade, subsistem ainda algumas deficiências na forma como a investigação tem lidado com o problema como referem Boyer e Verma (2000, pp. 129) citando Speier e Swink (1995) e Malhotra e Grover (1998), nomeadamente o facto de muitos investigadores não utilizarem fontes de dados múltiplas dentro da organização objecto de estudo. Há diversas deficiências com a utilização de uma só fonte (*e.g.* um só respondente a um questionário) para representar uma organização, nomeadamente a possibilidade de enviesamento subjectivo devido à perspectiva única do indivíduo e ao acesso limitado à informação. Boyer e Verma (2000) sustentam que a investigação que utiliza múltiplos classificadores, embora mais difícil quando comparada com a utilização de um só classificador, proporciona um grau de rigor metodológico maior permitindo obter, consequentemente, mais confiança nos resultados. A recolha de dados por questionário dirigido apenas ao gestor local das empresas industriais é reconhecida por Dow, Samson e Ford (1999) como uma limitação do seu estudo sobre práticas de qualidade na Austrália e Nova Zelândia. Embora reconhecendo que os gestores locais são, nas circunstâncias em que o estudo foi realizado, os melhores respondentes disponíveis, não deixam de sublinhar que teria sido preferível ter respondentes múltiplos. Os respondentes múltiplos, além de aumentar de forma considerável a imposição sobre a empresa questionada, teria permitido a realização de testes sobre a fiabilidade dos respondentes. A limitação a um só respondente suscita aos autores preocupações de enviesamento sobretudo nas questões relacionadas com o comportamento da gestão de topo e desempenho da empresa que têm conotações pejorativas (pp. 24).

A despeito da inexistência de consenso referida, Rungtusanatham (1988) refere o coeficiente κ de Cohen (1970) como medida de fiabilidade entre classificadores. Na determinação do coeficiente κ de Cohen (1970) pede-se a J classifi-

cadores para, de forma independente, ordenar N itens independentes num conjunto C de escalas mutuamente exclusivas definidas à priori para conceitos teóricos diferentes. Os valores de κ variam entre -1 e $+1$. Valores de κ superiores a zero significam que a concordância observada entre os especialistas não é mera casualidade. Quando o coeficiente $\kappa = 1$, existe perfeita concordância entre os especialistas. Valores inferiores a zero significarão que "muito provavelmente se tratam de itens sem interesse prático adicional ..." (pp. 12).

$$K = \frac{F_a - F_c}{N - F_c}$$

onde F_a é o número de itens classificados na mesma categoria por todos os J classificadores para todas as categorias i com $i = \{1, ..., C\}$, ou seja,

$$F_a = \sum_{i=1}^{C} F_{i(a)}$$

onde $F_{i(a)}$ é o número de itens classificados na mesma categoria por todos os J classificadores e F_c é o número de itens para os quais se espera haver acordo em relação às suas classificações entre todos os J classificadores para todas as categorias i existentes com $i = \{1, ..., C\}$, isto é,

$$F_c = \sum_{i=1}^{C} F_{i(c)}$$

com

$$F_{i(c)} = N \bullet \left(\prod_{j=1}^{J} \frac{F_{ij}}{N} \right), \forall i$$

e F_{ij} denotando o número de itens classificados na categoria i pelo classificador j.

Considerações Comuns

Diversos autores em diversas áreas têm sugerido quadros de referência para reflectir sobre a validade de novos instrumentos (Campbell e Fiske, 1959; Cook e Campbell, 1979; Carmines e Zeller, 1979; Bagozzi, 1979, 1980; Straub, 1989). Tratam-se de contributos importantes para o gradual refinamento dos instrumentos de medida utilizados em investigação. Goodhue (1998) sumaria e contrasta os elementos que aqueles autores consideram dever tomar-se em conta para validar um conceito teórico. Goodhue (1998) elege o contributo de Bagozzi (1979,

1980) por abranger as questões suscitadas pelos demais e desenvolve com algum pormenor os seis elementos, ou áreas de preocupação, relevantes que este autor considera na avaliação de um conceito teórico: *(a)* significado teórico dos conceitos – o que é que se deseja medir?; *(b)* significado observacional dos conceitos – desenvolver questões para medir os conceitos definidos.; *(c)* consistência interna – são consistentes as medidas quando repetidas?; *(d)* validade discriminante – pode o instrumento distinguir conceitos teóricos distintos?; *(e)* validade convergente – é o instrumento concordante com outras medidas aceites na comunidade científica?; *(f)* validade nomológica – é maior a confiança nas novas medidas se os conceitos teóricos quando medidos por estas se "comportam" de forma previsível relativamente a outros conceitos teóricos já bastante bem percebidos.

O quadro de referência de Churchill (1979) para o desenvolvimento de medidas de variáveis de investigação complexas é um dos mais amplamente aceites (Segars, Grover e Teng,1998). Embora este quadro fosse inicialmente apresentado no contexto de desenvolvimento de conceitos teóricos de marketing, a sua natureza genérica tornou-o aplicável a uma grande variedade de estudos em gestão estratégica e em sistemas de informação. Como sublinha Churchill (1979) citado por Segars, Grover e Teng (1998, pp.315), "muitas variáveis de interesse são inerentemente complexas por natureza; não podem, portanto, ser medidas com precisão usando uma escala simples. As medidas simples têm, tipicamente, considerável singularidade e subsequentemente baixa correlação com o atributo que se deseja medir. Mais, os itens simples tendem a enquadrar os conceitos de forma muito limitada resultando assim num considerável erro de medida. As medidas com itens múltiplos superam estas dificuldades. A especificidade dos itens individualizados pode ser diluída, desenvolvem-se conceptualizações mais robustas de variáveis complexas e, portanto, reduz-se o erro de medida".

O *valor* de qualquer observação científica depende em parte da sua repetitibilidade ou replicabilidade. Uma observação que não pode ser replicada não é fiável e não deve ser admitida como evidência científica (Neale and Liebert, 1980). A necessidade de determinação da fiabilidade de um instrumento de medida e de procedimentos é extremamente importante em gestão e, de um modo geral, nas ciências sociais. Por exemplo, em algumas profissões, a progressão na carreira é determinada, entre outros critérios, pelo nível de desempenho aferido por instrumentos de medida que para o efeito são especificamente desenhados. Um dos problemas que imediatamente se coloca é o de saber em que medida aquele instrumento *mede* realmente *o que se quer aferir* ou se, pelo contrário, não medirá *outra coisa qualquer.* A fiabilidade é, assim, a condição segundo a qual a evidência e as medidas utilizadas são consistentes e estáveis (Remenyi et all, 1998), isto é, a fiabilidade implica que os investigadores obterão resultados semelhantes em diversas ocasiões sempre que na avaliação de um dado conceito descrito por um conjunto de variáveis utilizarem o mesmo instrumento de medida.

Consistência e estabilidade são pois dois atributos de extrema relevância quando os resultados da investigação se aplicam a outras situações que não apenas a estudada no contexto original em que a investigação foi conduzida.

Embora incidindo numa área limitada do conhecimento, a da gestão da qualidade, Dow, Samson e Ford (1999) referem-se ao debate mantido ao longo das duas últimas décadas com uma nota de desapontamento sustentada no facto de muitas das afirmações de profissionais e de académicos só em casos muito excepcionais serem verificadas por rigorosa investigação empírica de larga escala. Embora apenas parcialmente, os autores propõem-se superar este panorama examinando duas questões vitais: primeiro, a identificação das principais dimensões das práticas de gestão de qualidade e, segundo, as formas como as referidas práticas interagem para produzir resultados de superior qualidade. Fazem-no recorrendo à recolha de dados através de um questionário dirigido a cerca de 4000 empresas. É notável a descrição da metodologia, que representa parte substancial do trabalho (pp. 5-12), sobretudo no que se refere aos aspectos do desenvolvimento do questionário e aos aspectos de validade de conteúdo dos instrumentos de prática de qualidade e dos de resultados de qualidade.

Considerações Finais

A motivação para a realização desta digressão pela literatura decorre do facto de constatarmos que, em muitos estudos académicos, há pouco rigor metodológico na elaboração de questionários para abordar, quase sempre, temas de gestão de indiscutível interesse. Maior rigor permitiria que o nosso entendimento desses temas fosse mais profundo. A literatura contém inúmeras citações sobre a importância que a fiabilidade e validade ocupam na concepção de instrumentos de medida. Na medida em que o recurso ao questionário para recolha de dados continua a ter ampla utilização, este estudo contribui para sublinhar a importância do instrumento e dos cuidados metodológicos a considerar na sua preparação. Lateralmente, proporciona, sobretudo aos mais interessados, uma lista bibliográfica com cerca de cinco dezenas de inserções das quais três foram publicadas nos últimos dez anos.

Referências

Ahmad, S. e Schroeder, R. G. (2001), The Impact of Electronic Data Interchange on Delivery Performance, *Production and Operations Management*, 10, 1, 16-30.

Bagozzi, R. P. (1979), The role of measurement in theory construction and hypothesis testing: Toward a holistic model. In Ferrel, O. C., Brown, S. W., and Lamb, C. W., Jr. (eds.), *Conceptual and theoretical developments in marketing.* Chicago: American Marketing Association.

Bagozzi, R. P. (1980), Causal models in marketing. New York: John Wiley.

Bates, K. A., Amundson, S. D., Schroeder, R. G. e Morris, W. T. (1995), The Crucial Interrelationship Between Manufacturing Strategy and Organizational Culture, *Management Science*, 41, 10, 1565-1580.

Boyer, K. K. e Verma, R. (2000), Multiple Raters in Survey-Based Operations Management Research: A Review and Tutorial, *Production and Operations Management*, 9, 2, 128-140.

Brandyberry, A., Rai, A. e White, G. P. (1999), Intermediate Performance Impacts of Advanced Manufacturing Technology Systems: An Empirical Investigation, *Decision Sciences*, 30, 4, 993-1020.

Campbell, D. T. e Fiske, D. W. (1959), Convergent and discriminant validation by the multitrait-multimethod matrix, *Psychological Bulletin*, 56, 2, 81-105.

Carmines, E. G. e Zeller, R. A. (1979), Reliability and validity assessment. Newbury Park, CA: Sage Publications.

Churchill, G. A. (1979), A paradigm for developing better measures of marketing constructs, *Journal of Marketing Research*, 16, 3, 64-73.

Cook, T. D. e Campbell, D. T. (1979), Quasi-experimentation: Design and analysis issues for field settings. Boston: Houghton Mifflin.

Corbett, C. e Van Wassenhove, L. (1993), Trade-off? What trade-offs? Competence and competitiveness in manufacturing strategy, *California Management Review*, 25, 4, 107-122.

Cronbach, L. J. (1951), Coefficient Alfa and the Internal Structure of Tests, *Psychometrika*, 16, 3, 297-334.

Dow, D., Samson, D. e Ford, S. (1999), Exploding the Myth: Do All Quality Management Practices Contribute to Superior Quality Performance?, *Production and Operations Management*, 8, 1, 1-27.

Emory, W. C. (1985), Business Research Methods. 3rd ed. Homewood, IL: Richard D. Irwin.

Flynn, B. B., Schroeder, R. G. e Sakakibara, S. (1995), The Impact of Quality Management Practices on Performance and Competitive Advantage, *Decision Sciences*, 26, 5, 659-691.

Frohlich, M. e Dixon, J. R. (1999), Information Systems Adaptation and the Successful Implementation of Advanced Manufacturing Technologies, *Decision Sciences*, 30, 4, 921-957.

Germain, R. e Droge, C. (1997), An Empirical Study of the Impact of Just-in-Time Task Scope Versus Just-in-Time Workflow Integration on Organizational Design, *Decision Sciences*, 28, 3, 615-635.

Goodhue, D. L. 1998), Development and Measurement Validity of a Task-Technology Fit Instrument for User Evaluations of Information Systems, *Decision Sciences*, 29, 1, 105-138.

Jacoby, J. (1978), Consumer Research: A State of the Art Review, *Journal of Marketing*, 42, April, 87-96.

Johnson, R. A. e Wichern, D. W. (1988), Applied Multivariate Statistical Analysis. Englewood Cliffs, NJ: Prentice Hall.

Kettinger, W. J. e Grover, V. (1997), The Use of Computer-mediated Communication in an Interorganizational Context, *Decision Sciences*, 28, 3, 513-555.

Kikulis, L. M., Slack, T., e Hinings, C. R. (1995), Sector-Specific Patterns of Organizational Design Change, *Journal of Management Studies*, 32, 1, 67-100.

Leidner, D. E., Carlsson, S., Elam, J. e Corrales, M. (1999), Mexican and Swedish Managers' Perceptions of the Impact of EIS on Organizational Intelligence, Decision Making, and Structure, *Decision Sciences*, 30, 3, 633-658.

Li, L. X. (1997), Relationships Between Determinants of Hospital Quality Management and Service Quality Performance – a Path Analytic Model, *Omega, International Journal of Management Science,* 25, 5, 535-545.

Malhotra, M. e Grover, V. (1998), An Assessment of Survey Research in POM: From Constructs to Theory, *Journal of Operations Management*, 16, 4, 407-425.

Meredith, J. R., Raturi, A., Amiako-Gyampah, K. e Kaplan, B. (1989), Alternative Research Paradigms in Operations, *Journal of Operations Management*, 8, 4, 297-326.

Mirani, R. e Lederer, A. L. (1998), An Instrument for Assessing the Organizational Benefits of IS Projects, *Decision Sciences,* 29, 4, 803-838.

Mitroff, I. I. e Mason, R. O. (1984), Business Policy and Metaphysics: Some Philosophical Considerations, Manuscrito não publicado, University of Southern California.

Narasimhan, R. e Das, A. (1999), An Empirical Investigation of the Contribution of Strategic Sourcing to Manufacturing Flexibilities and Performance, *Decision Sciences,* 30, 3, 683-718.

Neale, J. M. e Liebert, R. M., (1980), Science and Behavior: An Introduction to Methods of Research, Prentice-Hall, Englewood Cliffs, NJ.

Neter, J., Wasserman, W. e Kutner, M. (1985), Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs. Homewood, IL: Richard D. Irwin.

Noble, M. A. (1995), Manufacturing Strategy: Testing the Cumulative Model in a Multiple Country Context, *Decision Sciences*, 26, 5, 693-721.

Parasuraman, A., Zeithaml, V. A. e Berry, L. L. (1988), SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality, *Journal of Retailing*, 64, 1, Spring 1988.

Remenyi, D, Williams, B., Money, A. e Swartz, E. (1998), Doing Research in Business and Management. London: SAGE Publications.

Ritzman, L. P. e Safizadeh, M. H. (1999), Linking Process Choice with Plant-Level Decisions About Capital and Human Resources, *Pròduction and Operations Management*, 8, 4, 374-392.

Rumelt, R. P., Schendel, D. E. e Teece, D. J. (1998), Questões Fundamentais de Estratégia. In *Questões Fundamentais de Estratégia*, Rumelt, R. P., Schendel, D. E. e Teece, D J., (eds.). Tradução de Pereira, A., Bertrand Editora, Venda Nova.

Rungtusanatham, M. (1998), Let's Not Overlook Content Validity, *Decision Line – A News Publication of the Decision Sciences Institute*, 29, 4, 10-13.

Segars, A., Grover, V. e Teng, J. (1998), Strategic Information Systems Planning: Planning System Dimensions, Internal Coalignment, and Implications for Planning Effectiveness, *Decision Sciences,* 29, 2, 303-345.

Speier, C. e Swink, M. (1995), Manufacturing Strategy Research: An Examination of Research Methods and Analytical Techniques, Indiana University Working Paper.

Straub, D. W. (1989), Validation instruments in MIS research, *MIS Quarterly*, 13, 2, 147-170.

Tracey, M., Vonderembse, M. e Lim, J-S (1999), Manufacturing technology and strategy formulation: keys to enhancing competitiveness and improving performance, *Journal of Operations Management*, 17, 411-428.

Williams, S. R. e Wilson, R. L. (1997), Group Support Systems, Power, and Influence in an Organization: A Field Study, *Decision Sciences*, 28, 4, 911-937.

Wilson, D. D. e Collier, D. A. (2000), An Empirical Investigation of the Malcolm Baldrige National Quality Award Causal Model, *Decision Sciences*, 31, 2, 361-390.

Abstract

This article discusses both validity and reliability when questionnaires are employed in management research to collect data. It stresses their relevance to ascertain methodological credibility. It compares both inter item and inter rater reliabilities and describes some methods to measure them. By briefly reviewing the literature, the article underlines the minimum requirements to be taken into consideration whenever the researcher selects the questionnaire as the instrument to gather data pertinent to the study.