# Eu quero acreditar! Da importância, formas de uso e limites dos testes de confiabilidade na Análise de Conteúdo

Revista de Sociologia e Política

DOI 10.1590/1678-987318266602

Rafael Sampaio e Diógenes Lycarião

RESUMO **Introdução**: A análise de conteúdo (AC) tem sido normativamente definida a partir de três princípios fundamentais: validade, replicabilidade e confiabilidade. Neste trabalho, identificamos que os estudos empíricos, no Brasil e no exterior, têm negligenciado esses princípios, em especial o último (confiabilidade). **Métodos**: Diante disso, oferecemos uma contribuição tanto operacional quanto crítica à AC. Em relação à contribuição operacional, o trabalho apresenta uma listagem detalhada de procedimentos sobre como se realizar um teste de confiabilidade em diferentes circunstâncias de pesquisa (com um ou mais pesquisadores). **Resultados**: Já a contribuição crítica é uma reflexão epistemológica acerca das vantagens e limites dos usos mais comuns desse tipo de teste, usos esses que, no limite, podem comprometer a própria confiabilidade científica dos resultados publicados. **Discussão**: De modo a evitar esse risco, propomos que pesquisas com AC podem reivindicar a *presunção de confiabilidade* quando (a) oferecem plenas condições de replicabilidade e (b) quando oferecem um teste de confiabilidade realizado por um ou mais codificadores, teste esse que seja aceitável como significativamente não-aleatório. Por fim, concluímos que, nos periódicos de alto impacto, predomina a importância de (b) em detrimento de (a), o que demonstra que, mesmo na elite da produção científica, ainda prevalece uma compreensão pouco exigente sobre a confiabilidade na AC.

PALAVRAS-CHAVE: análise de conteúdo; teste de confiabilidade; replicabilidade; validade; métodos quantitativos.

Recebido em 13 de Março de 2017. Aceito em 2 de Agosto de 2017.

## I. Introdução[1]

[1] Agradecemos aos comentários e sugestões dos pareceristas anônimos da *Revista de Sociologia e Política*.

Data de séculos relatos sobre técnicas de análise de textos, as quais foram realizadas sob o interesse de se catalogar e classificar textos e materiais sob os mais distintos propósitos (Bardin 2016; Krippendorff 2004; Neuendorf 2002; Riffe, Lacy & Fico 2014). Entretanto, a Análise de Conteúdo (AC), para fins científicos, surge para suprir uma necessidade de mensuração dos padrões das mensagens mediáticas, especialmente nos períodos das duas Guerras Mundiais do século XX, como foi o caso, a título de exemplo, do estudo de Harold Lasswell (1927) sobre as propagandas de guerra, por meio da análise de conteúdo.

Já ao longo da Segunda Guerra Mundial, o uso de mídias massivas (*e.g.* rádio e cinema) pelos Estados Unidos, pela Alemanha nazista e outros regimes totalitários emergiu como um fenômeno sobressalente, o que despertou interesse do governo norte-americano em verificar como seus adversários faziam uso destas mensagens. Harold Lasswell era o coordenador da "Divisão experimental para o estudo de comunicações em tempos de guerra", criada pelo Congresso americano para tal fim. Para além da avaliação das mensagens mediáticas adversárias, preocupava-o os possíveis efeitos das mensagens dos meios de comunicação de massa sobre as pessoas. O pesquisador estava interessado em verificar se o conteúdo das mensagens mediáticas apresentava um efeito direto no público. Para analisá-las, a técnica da análise de conteúdo quantitativa foi, então, largamente utilizada e aperfeiçoada[2].

Não obstante seu surgimento no campo da comunicação política, a análise de conteúdo[3] tornou-se uma técnica bastante difundida em toda a área de

[2] Sobre a trajetória da técnica, ver Bardin (2016), Jorge (2015), Krippendorff (2004), Neuendorf (2002).

humanidades, como, por exemplo, na Ciências Sociais (Pohlmann, Bär & Valarini 2014; Triviños 1987), Ciência da Informação (Lima & Manini 2017; Lima & Moraes 2017), Contabilidade (Alves 2011), Geografia (Paula 2015), História (Constantino 2002), Psicologia (Gondim & Bendassolli 2014), Serviço Social (Lara 2011), Turismo (Thomaz *et al.*, 2016), além, claro, de seu vasto uso nas áreas de Administração (Freitas 2011; Mozzato & Grzybovski 2011; Vergara 2011), Comunicação (Herscovitz 2007; Jorge 2015; Pessoni & Martinez 2015; Quadros, Assmann & Lopez 2014; Vimieiro & Maia, 2011), Educação (Franco 2008; Moro 1989; Oliveira *et al.*, 2003) e de Ciência Política (Alves, Figueiredo Filho & Henrique 2015; Carlomagno & Rocha 2016; Feres Jr. 2016; Figueiredo *et al.*, 1997; Panke & Cervi 2012). Também é visível sua utilização nas ciências da saúde, como a Enfermagem (Bellucci Júnior & Matsuda 2014; Campos 2004; Taquette & Minayo 2015).

Diante deste cenário, é seguro afirmar que a análise de conteúdo tem grande capilaridade na ciência brasileira, tendo sido aplicada em um considerável número de estudos, de diferentes áreas. Essa aplicação também se mostra acompanhada por debates sobre as benesses e os limites da técnica (Cavalcante, Calixto & Pinheiro 2014; Freitas 2011; Moro 1989; Mozzato & Grzybovski 2011; Gondim & Bendassolli 2014; Vergara 2011), sua diferenciação e separação das linhas de análise do discurso (Lima 2003; Lima & Moraes 2017; Rocha & Deusdará 2006) e, recentemente, o uso de *softwares* para sua realização (*e.g.* Alves, Figueiredo Filho & Henrique 2015; Lima & Manini 2017).

Buscando contribuir a esse debate e aperfeiçoar a aplicação da AC, esse trabalho parte da constatação de que a quase totalidade da pesquisa brasileira[4] e boa parte da internacional[5] parece não se preocupar com um componente desta metodologia que passou a ser central (ou mesmo imprescindível para sua validade científica) para a comunidade especializada nesta técnica: o teste de confiabilidade ou de concordância entre codificadores (Feng 2014; Hayes & Krippendorff 2007; Lombard, Snyder-Duch & Bracken 2002; Macnamara 2005; Matthes & Kohring 2008). Importante, entretanto, esclarecer que aqui trabalhamos com a AC quantitativa. Há discussões relevantes sobre análises de conteúdo qualitativas que não geram dados quanti, mas que não fazem parte do escopo do artigo (ver Krippendorf 2004, p.15).

O teste de confiabilidade entre codificadores (no original, *inter-coder reliability test*[6]) busca verificar se diferentes codificadores têm a mesma compreensão sobre as variáveis de análise e se a codificação pode ser replicada por outrem, gerando resultados similares. Em suma, enquanto há uma forte e vívida discussão acerca dos diferentes índices de confiabilidade, sobre modos de aumentar a confiabilidade da codificação e a respeito dos problemas que podem acontecer durante esta codificação, aparentemente boa parte da pesquisa disponível sequer faz algum teste para averiguar a confiabilidade de suas variáveis ou da análise em si.

Diante de tal lacuna, este texto possui dois objetivos[7]. Em primeiro lugar, apresentar uma breve discussão epistemológica sobre a análise de conteúdo[8] e sua fundação sob três princípios, a saber: validade, replicabilidade e confiabilidade, ressaltando-se a necessidade de uma compreensão mais exigente sobre a última. Uma compreensão que reconheça os limites epistemológicos que a simples execução de um teste de confiabilidade traz para a satisfação do princípio correspondente. Em segundo lugar, o trabalho visa expor, sucintamente, como realizar um teste de confiabilidade. Com o intuito de tornar o procedimento o mais acessível possível, explicitamos como fazê-lo de diferentes formas, indicando *softwares* e páginas na internet capazes de calculá-lo, assim como alternativas que possam abranger pesquisas individuais com menos recursos.

[3] Portanto, estamos nos referindo aqui à análise de conteúdo "clássica" ou, ainda, humana, baseada na codificação manual por pessoas, portanto boa parte de nossas inferências não se aplicam à AC computadorizada.

[4] Não é nossa intenção afirmar que não havia a realização de testes de confiabilidade anteriormente. Nosso objetivo é chamar a atenção para o fato da maior parte das pesquisas não se preocupar com tal etapa. Para discussões anteriores no Brasil, ver Belluci Júnior & Matsuda (2014) e Moro (1989).

[5] A nosso ver, a análise de conteúdo de periódicos de alto impacto (geralmente publicados no exterior) tende a respeitar mais a necessidade dos testes de confiabilidade. Porém, o estudo de Macnamara (2005) reporta que se trata de uma preocupação mais recente e ainda não universalizada na literatura internacional.

[6] Ou *inter-rater reliability test.*

[7] É importante ressaltar que este texto já parte do pressuposto que o leitor conhece e tem familiaridade com a técnica de análise de conteúdo, portanto não dedica especial preocupação à sua explanação, de suas etapas e afins.

[8] Em português, para além de Bardin (2016), há um conjunto de trabalhos que podem ajudar a compreender e a aplicar a técnica (Bauer 2007;

Carlomagno & Rocha 2016; Cavalcante, Calixto & Pinheiro 2014; Franco 2008; Herscovitz 2007; Jorge 2015; Oliveira 2008; Moraes 1999; Triviños 1987).

Por fim, concluímos com uma reflexão acerca do déficit metodológico e epistemológico na operacionalização do princípio de confiabilidade na AC, tanto na literatura internacional como nacional, sendo mais grave e evidente na segunda. Defendemos que esse déficit poderá ser reduzindo mediante (a) a oferta de plenas condições de replicabilidade e (b) a realização de um teste de confiabilidade por um ou mais codificadores cujos resultados sejam aceitáveis como, significativamente, não-aleatórios.

## II. A epistemologia da Análise de Conteúdo (AC)

A análise de conteúdo científica precisa se preocupar ativamente com três princípios epistemológicos fundamentais, a saber, validade, replicabilidade e confiabilidade (Krippendorff 2004; Macnamara 2005; Matthes & Kohring 2008; Neuendorf 2002; Riffe *et al.*, 2014)[9]. Enquanto a confiabilidade é o foco deste artigo, os outros dois princípios estão diretamente imbricados com ela e precisam ser devidamente considerados para que uma compreensão mais exigente sobre a própria confiabilidade possa aparecer.

### *II.1. Da validade e sua distinção em relação à confiabilidade*

[9] Pelos limites do artigo, esta discussão foi abreviada. Para além das discussões presentes na literatura especializada, recomendamos a discussão entre validade e fidedignidade realizada por Moro (1989).

A validade seria a "melhor aproximação possível à verdade ou a falsidade das observações, das descobertas, dos resultados interpretados" (Moro 1989, p.159), abrangendo a realidade observada, o dado empírico e a sua interpretação. Na pesquisa, então, a validade é a extensão na qual um processo de medição representa o conceito intencionado (Neuendorf 2002), ou ainda:

> "Dizer que uma categorização deve ser válida significa dizer que deve ser adequada ou pertinente. Esta adequação se refere aos objetivos da análise, à natureza do material que está sendo analisado e às questões que se pretende responder através da pesquisa. A validade ou pertinência exige que todas as categorias criadas sejam significativas e úteis em termos do trabalho proposto, sua problemática, seus objetivos e sua fundamentação teórica" (Moraes 1999, sp.).

A *validade* seria, desse modo, uma adequação epistemológica entre os propósitos da pesquisa e os instrumentos utilizados para identificar o fenômeno sob investigação. Se o pesquisador está, por exemplo, interessado em investigar o peso médio de uma população, ele precisará de instrumentos válidos para mensurá-la, a exemplo de uma balança. Entretanto, caso ele escolha fazer essa mensuração através do tamanho dos calçados dos indivíduos, isso, de maneira evidente, representaria um sério problema de validade à pesquisa. Isso independentemente do quão precisa for a mensuração do tamanho dos calçados, pois, *a priori*, este não seria considerado um instrumento válido para o propósito da pesquisa. Em comparação com a balança, o tamanho do calçado poderia ser um instrumento até mais preciso e, portanto, mais confiável do que a balança, especialmente se esta estiver desregulada. Entretanto, em nenhuma hipótese, o tamanho do calçado seria considerado um instrumento válido.

Esse exemplo hiperbólico almeja demonstrar que as dimensões de validade e confiabilidade são distintas, não podendo ser subsumidas. Enquanto a confiabilidade se refere à estabilidade e precisão do instrumento utilizado pela análise da pesquisa (ou seja, uma balança desregulada não é confiável), a validade corresponde ao julgamento sobre a pertinência epistemológica do instrumento mesmo.

A validade, na AC, entretanto, é, muitas vezes, tomada como um conceito bastante amplo por alguns autores, o que acaba por borrar essa distinção conceitual mais intuitiva com a confiabilidade. É justamente o que faz Neuendorf (2002) ao incluir, na definição de validade, a acurácia das categorias, no sentido de deixá-las livre de distorções e erros não aleatórios. Ainda segundo

Neuendorf, a validade pode ser distinta entre externa e interna. Enquanto validade externa se referiria à capacidade de os resultados da pesquisa serem generalizados, ou seja, de aferir se eles podem ser extrapolados para outros contextos; a validade interna corresponderia à conexão entre a definição conceitual das categorias e a sua operacionalização, estando, portanto, mais relacionada à medição em si[10]. Ou seja, enquanto a primeira (a externa) está mais próxima da noção de confiabilidade apresentada no parágrafo acima, a segunda, estaria mais fiel à definição de validade.

[10] Para mais sobre os diferentes componentes da validade, ver Moro (1989).

Feito esse alerta sobre as possíveis confusões conceituais que podem ocorrer entre validade e confiabilidade, nos cabe agora esclarecer algo imprescindível a esta que é a *replicabilidade*.

### *II.2. Da replicabilidade como condição de possibilidade da confiabilidade*

A replicabilidade é o parâmetro que permite aferir o nível com que uma pesquisa *pode* ser replicada por outros pesquisadores, em contextos diferentes. Isso implica que a replicabilidade não garante confiabilidade, mas é condição de possibilidade desta, pois só é possível que outros pesquisadores cheguem a resultados iguais ou similares se estes têm (a) a seu dispor uma descrição detalhada dos procedimentos utilizados para se reproduzir a análise e (b) acesso ao mesmo material codificado em condição de integralidade equivalente ou suficiente para uma nova codificação (fora do contexto da pesquisa).

Sob essa compreensão, para aumentar a capacidade de replicação de uma pesquisa, é importante que a exigência do item (a) seja viabilizada através da disponibilização do máximo de informações sobre os procedimentos utilizados na análise (Neuendorf 2002). Para tanto, Oliveira (2008, p.251) esclarece que:

> "(...) as unidades decompostas da mensagem, as categorias que servem para classificá-la, devem ser definidas com tal clareza e precisão que outros, a partir dos critérios indicados, possam fazer a mesma decomposição, operar a mesma classificação".

A AC tem tradicionalmente satisfeito essa exigência a partir da produção e publicação de um livro de códigos (*codebook*, em inglês). Tal livro, além de indicar os códigos alfanuméricos que correspondem a cada variável e categoria, devem discriminar, detalhadamente, inclusive com exemplos, como a codificação deve ser feita em cada opção disponível[11].

[11] Exemplos, nesse sentido, podem ser encontrados em Carlomagno e Rocha (2016) e Lycarião (2014).

Já em relação à exigência de se disponibilizar o material da pesquisa no mesmo nível de integralidade da pesquisa original, uma verdadeira revolução ocorreu com a internet. Isso porque, enquanto na realidade tecnológica anterior à era da comunicação digital havia dificuldades técnicas decisivas em se disponibilizar, com ampla acessibilidade, o material codificado, atualmente tais barreiras se mostram bem mais factíveis de serem superadas. Nesse caso, a emergência de serviços especializados em armazenar e tornar disponível, em modo *online*, uma grande quantidade de dados permite disponibilizar não apenas o texto, mas arquivos de áudio e vídeo com a mesma qualidade do material original.

Apenas munidos das aludidas exigências (a) e (b) é que outros pesquisadores poderão replicar os procedimentos analíticos de uma AC. Entretanto, ao replicar tais procedimentos, é possível que os resultados encontrados sejam significativamente distintos dos da pesquisa replicada. Isso seria um indício de que, apesar de ser replicável, essa pesquisa não teria se mostrado efetivamente confiável, pois os resultados encontrados foram significativamente distintos. Por isso é que a replicabilidade é uma condição de possibilidade da confiabilidade e não garantia dela.

Sendo assim, a confiabilidade de uma pesquisa é algo que se confirma uma vez que outros pesquisadores (fora do contexto da pesquisa original), cheguem a resultados iguais ou similares após terem utilizado os mesmos procedimentos e codificado o mesmo material. Com isso, demonstrar-se-ia, *em condições ideais*, um tipo de confiabilidade definida como sendo "o grau em que a descoberta, a observação, o resultado, são independentes de flutuação, de circunstâncias acidentais" (Moro 1989, p.166). Assim, dados confiáveis (*reliable*) seriam aqueles que permanecem constantes através de variações (*e.g.* eventos, instrumentos, codificadores) no processo de medição.

Em outras palavras, uma AC tenderá a ser pouco confiável se seus resultados não forem baseados em um conjunto de técnicas que visam diminuir o poder discricionário do codificador (estabelecimento de regras *ad hoc* e não registradas para se codificar parte do material), assim como a influência de fatores internos (*e.g.* cansaço na codificação) e externos (mudança de codificadores durante o processo)[12]. No limite, se diferentes pessoas chegam a diferentes resultados aplicando as mesmas variáveis às mesmas unidades de análise, isso tende a demonstrar fragilidades da pesquisa, a qual pode estar a utilizar categorias pouco inteligíveis, operacionalizáveis e/ou codificadores despreparados.

[12] "Se a meta é produzir interpretações menos arbitrárias e mais acuradas, há métodos para atingi-la, ou pelo menos para avaliar o grau de falta de acuidade, prontamente disponíveis na literatura acadêmica. Um deles, talvez o mais importante, é o teste de confiabilidade entre codificadores (*intercoder reliability test*)" (Feres Jr. 2016, p.318; grifos no original).

Sobre essa questão, deve estar claro que a AC "clássica" ou "humana" envolve uma série de decisões subjetivas. A ideia de confiabilidade não visa anular a subjetividade do codificador, mas sim padronizar as formas com que diferentes codificadores compreendem as mesmas categorias analíticas, aumentando a chance que a interpretação que estes codificadores fizeram do conteúdo analisado seja a mais próxima possível de uma interpretação mínima comum, de caráter, portanto, *intersubjetivo*. Como esclarecem Wozniak, Lück & Wessler (2015, p.471), "a análise de conteúdo emula como o usuário comum interpretaria um texto particular. A análise padronizada não está, assim, tão interessada no conjunto completo de interpretações possíveis, mas naquela interpretação dominante e amplamente disseminada". Para chegar justamente a essa interpretação, Krippendorf indica que ela depende:

> "[...] fortemente de uma leitura e uso consensual dos dados que representam, apontam ou invocam experiência com o fenômeno de interesse. [...]. Então, a confiabilidade é o grau no qual membros de uma certa comunidade confiam nas leituras, interpretações, respostas e usos de certos textos e dados" (Krippendorf 2004, p.212, tradução livre).

Para, então, aumentar a confiabilidade de uma pesquisa com AC, a comunidade especializada sugere que sejam realizados testes de confiabilidade entre codificadores. Como procedimento central, esse tipo de teste requer que as variáveis sejam verificadas por dois ou mais codificadores em separado, ou seja, tomando apenas como referência o livro de códigos. A premissa é que, quanto mais esses codificadores concordarem entre si, mais precisas seriam as categorias utilizadas na codificação. De outro modo (caso exista uma discordância significativa), a pesquisa estaria sob pena de ter suas variáveis e categorias análise consideradas imprecisas e, portanto, não fiáveis.

### *II.3. Da relevância dos testes de confiabilidade*

Enquanto os testes de confiabilidade existem desde as primeiras codificações de Lasswell e equipe (Kaplan & Goldsen 1982), é notável que sua maior exigência se tornou o padrão da literatura especializada apenas a partir da década de 1990 (Lombard, Snyder-Duch & Bracken 2002; Macnamara 2005). Oportuno destacar que parte relevante dos especialistas em AC considera que quando o uso desta metodologia não apresenta qualquer preocupação com a

confiabilidade, o uso da AC seria simplesmente "*meaningless*" (Neuendorf 2002, p.12).

Diante, então, da crescente exigência desses testes pelos periódicos mais qualificados e especializados, é possível observar um grande déficit metodológico da AC brasileira e internacional devido ao uso ainda raro desse tipo de teste. Talvez isso ocorra pelo fato de que, também, são raras as publicações não anglófonas que mostram de maneira pormenorizada como realizá-lo, sendo ainda mais raras as descrições sobre como realizar um tal teste em pesquisas individuais, ou seja, com apenas um codificador. De toda sorte, diante de um cenário de maior exigência de internacionalização da pesquisa, o que envolve, muitas vezes, parcerias com acadêmicos de diferentes países, faz-se necessário que a AC passe a considerar esta importante questão negligenciada até o momento.

Tendo isso em vista, pretendemos, nas próximas seções, ajudar a corrigir esse déficit a partir de uma sistematização detalhada de diferentes alternativas factíveis para o aumento da presunção de confiabilidade das pesquisas com AC.

## III. Como realizar um teste de confiabilidade

Enquanto há boas descrições em português acerca das fases da AC, que inclusive apresentam a necessidade do teste de confiabilidade (*e.g.* Bauer 2007; Herscovitz 2007; Vimieiro & Maia 2011)[13], não encontramos descrições mais extensas sobre os passos de um teste de confiabilidade entre codificadores[14]. Na subseção abaixo, iremos, então, apresentar um passo a passo desse tipo para pesquisas com dois ou mais codificadores.

### *III.1. Testes para pesquisas com dois ou mais codificadores*

Para realizar um teste de confiabilidade com dois ou mais pesquisadores, mostra-se oportuno adaptarmos o passo a passo desenvolvido por Neuendorf (2002). A esquematização adaptada se encontra abaixo:

1. Escrever um livro de códigos com variáveis e categorias válidas para o fenômeno de interesse da pesquisa;

2. Treinamento dos codificadores, com discussão;

3. Codificadores praticam em conjunto, engajando em discussões que buscam construir consensos;

4. Possíveis revisões do livro de código;

5. Treinamento dos codificadores nas revisões;

6. Codificadores praticam a codificação de modo independente em um número de unidades que represente a variedade da população estudada;

7. Codificadores discutem os resultados da prática de codificação;

8. Possíveis revisões do livro de códigos;

9. Treinamento dos codificadores nas revisões;

10. Codificadores fazem uma codificação piloto numa amostra para testar a confiabilidade;

11. Pesquisador verifica confiabilidade mediante uma operação matemática que pondere a chance aleatória de concordância;

12. Possíveis revisões ao livro de códigos;

[13] Bardin (2016), inclusive, afirma que "para um maior rigor, esses resultados [da AC] são submetidos a provas estatísticas, assim como a testes de validação" (Bardin 2016, p.131).

[14] Para além dos manuais especializados já elencados, o pesquisador Mattew Lombard mantém um site dedicado ao tema, que resume bem todas as questões envolvidas no teste: http://matthewlombard.com/reliability . Acesso em 16 de ago. 2017.

[14] Para além dos manuais especializados já elencados, o pesquisador Mattew Lombard mantém um site dedicado ao tema, que resume bem todas as questões envolvidas no teste: http://matthewlombard.com/reliability . Acesso em 16 de ago. 2017.

13. Treinamento dos codificadores nas revisões;

14. Codificação independente final (incluindo checagens de confiabilidade). Nessa checagem, o cálculo do nível de confiabilidade deve ponderar a chance aleatória de concordância;

15. Relato dos codificadores de suas experiências (Neuendorf 2002, p.134)[15].

[15] Na descrição original, não há números para etapas. Estes foram acrescentados por motivos exclusivamente didáticos.

No passo a passo acima, é oportuno destacar que toda modificação do livro de códigos deve ser treinada pelos codificadores e que este processo não se encerra até o livro de códigos estar concluído. Todas estas mudanças devem ser registradas diretamente no próprio livro de códigos, que passará a ter vários exemplos das codificações. Com isso, amplia-se o nível de detalhamento das instruções envolvidas no treinamento, além de facilitar a replicabilidade da pesquisa.

Voltando ao passo a passo, a maior fragilidade da pesquisa empírica com AC parece se manifestar a partir do passo 6 e, especialmente, depois do passo 10, quando ocorrem os primeiros pilotos e testes de confiabilidade. Deve-se esclarecer, também, que todos os treinamentos, pilotos (também chamados de "pré-testes"), assim como o teste propriamente dito, devem ocorrer, preferencialmente, em unidades de análise distintas. Isso visa deixar os codificadores aptos a codificarem a diversidade do material a ser analisado[16].

[16] Por exemplo, se uma pedagoga deseja estudar mensagens enviadas por alunos em um grupo de redes sociais, ela deveria sempre selecionar um conjunto diferente de mensagens para os testes piloto e de confiabilidade. De outra forma, os codificadores poderão se lembrar das mensagens e dos acordos realizados, o que significa que estarão treinando a memória e não as categorias em si.

Durante este treinamento inicial, espera-se que sejam criados consensos entre os codificadores, que a aplicação dos códigos e os exemplos destas aplicações sejam definidos e aceitos pelos codificadores em conjunto com o pesquisador. Assim, essas definições devem buscar a criação de dicas, pistas, consensos e critérios específicos para diferenciar uma categoria da outra. Todas essas dicas e esses consensos devem, contudo, estar devidamente registrados no livro de códigos.

Os testes de confiabilidade se dão a partir da comparação entre as codificações de dois ou mais codificadores sobre um mesmo excerto de material. Em outras palavras, nos testes de confiabilidade *todos os codificadores codificam exatamente o mesmo material, mas de forma independente*. Isso implica que eles não podem conversar ou trocar qualquer tipo de informação entre si, durante a codificação. Para isso, é recomendável que os codificadores realizem a codificação do teste em locais e/ou momentos diferentes.

Ademais, todos os testes de confiabilidade (com exceção dos pilotos) devem ocorrer em uma amostra que seja aleatória e representativa da população estudada[17]. Se, por exemplo, um estudo decide analisar os prontuários de um posto médico no período de dois anos (n=500), o teste de confiabilidade final deve ocorrer numa amostra escolhida de maneira aleatória, e que seja representativa do total[18]. Isto significa dizer que os prontuários escolhidos para o teste devem compreender todos os períodos dentro do recorte temporal.

[17] Há várias maneiras de se calcular como uma amostra é representativa do total e o intervalo de confiança desejado para o mesmo, ver Herscovitz (2007). Como uma regra simplificadora, recomendamos que seja analisado cerca 10% do total da população de textos, sendo 50 unidades de análise o mínimo para um teste de confiabilidade (cf. Riffe *et al.*, 2014).

[18] A exceção são os testes pilotos (ou pré-testes), quando o objetivo principal é afinar a compreensão dos codificadores, criar consensos e fazer as modificações ao livro de códigos. Neste caso, a

Incialmente, o resultado do teste de confiabilidade era calculado apenas pela concordância absoluta, ou seja, comparando-se as codificações de modo a obter a percentagem dos casos em que houve concordância. Entretanto, atualmente esse é considerado um procedimento bastante limitado. Isso porque há variáveis (especialmente as dicotômicas) em que há uma chance alta de concordância aleatória entre os codificadores (Hayes & Krippendorff 2007).

Essa situação pode ser ilustrada a partir de uma simplificação da variável denominada valência. Ela se encontra em diversos estudos da comunicação e da ciência política e visa identificar como matérias jornalísticas constroem a imagem pública de atores políticos (*e.g.* Feres Jr. 2016; Figueiredo *et al.*, 1997), ou seja, se a matéria apresenta uma imagem positiva ou negativa de um

amostra não precisa ser representativa do total, mas o ideal é que ainda seja escolhida de forma aleatória.

determinado ator político. Como há apenas duas possibilidades de codificação (ignoremos, para fins didáticos, a valência neutra), isso significa que há, em cada decisão, 50% de chance de os codificadores concordarem de maneira aleatória. Isso quer dizer que, mesmo sem ler o material e apenas preenchendo aleatoriamente uma das duas opções, os codificadores ainda terão uma alta chance de concordarem em cada decisão.

Diante disso, a literatura especializada aponta que a simples percentagem nominal da concordância entre os codificadores não é suficiente para garantir a confiabilidade da codificação (Hayes & Krippendorff 2007; Lombard *et al.*, 2002; Macnamara 2005; Matthes & Kohring 2008). Conforme essa literatura, é, então, necessário aplicar algum índice de teste de confiabilidade entre codificadores que pondere a chance aleatória de concordância[19]. Tais testes se dão a partir de equações matemáticas que incorporam, além da concordância absoluta entre os codificadores, a covariação, a particular raridade ou excessiva aparição de certas categorias, além, é claro, da chance de os codificadores concordarem aleatoriamente entre si[20].

[19] Não é nosso objetivo entrar nos meandros das diferentes fórmulas, uma vez que isto já foi realizado com propriedade anteriormente (Feng 2014; Hayes & Krippendorff 2007).

[20] Este é o motivo da maioria não recomendar o uso de técnicas estatísticas simples (alpha de Cronbach, r de Pearson ou qui-quadrado), uma vez que estes testes não medem a concordância entre os codificadores ou mesmo a chance da concordância aleatória, mas apenas a covariação entre os resultados encontrados.

Os valores dos testes de confiabilidade, geralmente, variam de "-1" a "1", em que "1" indica uma concordância perfeita; "0" uma falta de confiabilidade por serem pareamentos aleatórios; e abaixo de zero que há uma discordância não aleatória ocorrendo. Ou seja, quando temos um índice negativo, os codificadores estão codificando o mesmo material a partir de compreensão significativamente e sistematicamente distintas uns dos outros. Qualquer valor acima de 0,9 é, em geral, considerado muito confiável e acima de 0,8 suficientemente confiável. Já valores entre 0,667 e 0,8 são considerados suficientes para variáveis experimentais (em aperfeiçoamento) no caso do alpha de Krippendorff (ver abaixo), mas serão passíveis de diferentes interpretações, a depender do índice utilizado. Por sua vez, valores abaixo de 0,667 tendem a ser aceitos apenas para estudos em fase de teste (Neuendorf 2002).

### *III.2. Qual índice utilizar e como?*

Apesar de existirem mais de 20 índices diferentes para se realizar um teste de confiabilidade, três deles se destacam pelo seu uso mais disseminado, o kappa de Cohen, O pi de Scott, e o alpha de Krippendorff[21]. Desses, o alpha de Krippendorf, além de ser altamente exigente, mostra-se bastante prático e versátil, pois não tem restrição em termos de número de codificadores (alguns índices só funcionam para dois codificadores, como o kappa de Cohen[22]) e de natureza das variáveis, se são ordinais, categóricas ou contínuas[23] (Hayes & Krippendorff 2007). Entretanto, para pesquisas que trabalham com um número alto de categorias pouco presentes, a escolha de outro índice pode ser mais adequada (ver Wozniak, Lück & Wessler 2015).

A maior parte dos *softwares* de AC, como Nvivo[24], e *softwares* de análise estatística, como SPSS[25] e R[26], possui pacotes ou adaptações para a realização dos testes de confiabilidade, conforme os *links* disponíveis nas notas de rodapé. Há, também, *softwares* específicos para o cálculo, como o Simstate[27] e o *Program for Reliability Assessment with Multiple Coders* (PRAM)[28].

Apresentando uma curva de aprendizado mais baixa, há páginas *online* que calculam os índices de forma simples, sendo alimentados diretamente[29] ou por arquivos[30]. Um notório exemplo é a página desenvolvida pelo pesquisador Dean Freelon[31]. Ela faz o cálculo do pi de Scott, kappa de Cohen, kappa de Fleiss e alpha de Krippendorf, utilizando a ferramenta ReCal[32] (Freelon 2010). Basta considerar as linhas da planilha (.csv) como os casos (unidades de análise), as colunas como as categorias codificadas e colocar lado a lado as codificações de todos os codificadores. Vide um exemplo com dez casos e três

[21] Para mais sobre as diferenças entre os índices, ver Hayes e Krippendorf (2007) e Feng (2014).

[22] O kappa de Fleiss é uma adaptação do kappa de Cohen para mais de dois codificadores.

[23] Para mais sobre os tipos de variáveis, ver Krippendorf (2004).

[24] Ver mais no link: http://help-nv11.qsrinternational.com/desktop/procedures/run_a_coding_comparison_query.htm. Apenas em inglês. Acesso em 15 ago. 2017.

[25] Ver mais no link http://afhayes.com/spss-sas-and-mplus-macros-and-code.html para a instalação. Informações mais detalhadas disponíveis neste link: http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf). Apenas em inglês. Acesso em 15 ago. 2017.

codificadores no Quadro 1. Os números representam os valores atribuídos por cada codificador. Os negritos representam as discordâncias.

Ao utilizar a ferramenta ReCal, o site já retorna com o resultado completo, conforme o exemplo na Figura 1.

Em nosso exemplo, mesmo havendo apenas três discordâncias entre os codificadores, o resultado é de 0,707 no índice do alpha de Krippendorf, o que seria aceitável em praticamente qualquer avaliação, caso fosse uma amostra representativa do material a ser analisado. Note-se que o índice (seja qual for o escolhido) é calculado para cada variável. Deve-se destacar que o procedimento de um teste rigorosamente executado não deve fazer a média entre os valores obtidos em cada variável ou qualquer outra adaptação dos resultados. Sendo assim, é recomendável que seja exibido o valor do índice para cada variável do livro do códigos da pesquisa[33]. Se o menor valor for abaixo do índice aceitável, o teste falhou[34] e deve ser repetido, após nova rodada de treinamentos.

Segundo os manuais especializados (Krippendorf 2004; Neuendorf 2002; Riffe *et al.*, 2014), se uma variável não passou, o teste falhou e deve ser repetido para todas as variáveis. Particularmente, parece-nos aceitável que os testes sejam repetidos apenas para as variáveis que não passaram, em vez de todas, especialmente quando o livro de códigos for extenso. Todavia, esta prática deve ser utilizada com cuidado. Um tempo muito excessivo entre o treinamento e a efetiva codificação final tende a diminuir a memória dos codificadores sobre os consensos estabelecidos e diminuir a confiabilidade geral.

Caso a complexidade de uma variável seja elevada, o esperado é que sejam necessários vários testes para se alcançar a confiabilidade desejável. Nesta situação, mais testes pilotos e mais tempo de treinamento serão necessários em relação a variáveis que requerem alta densidade interpretativa. Por exemplo, identificar a *editoria* de uma matéria de jornal impresso requer menor densidade interpretativa do que identificar a *valência* de um ator político nessa mesma matéria. De toda forma, os novos treinamentos deveriam focar, particularmente, as divergências entre os codificadores e a compreensão das razões pelas quais eles discordaram durante as codificações.

Deve-se observar que os testes de confiabilidade tendem a falhar quando os codificadores alcançam próximo de 100% de concordância, uma vez que a maior parte dos índices espera alguma discordância entre sujeitos (cf. Hayes & Krippendorff 2007). Logo, os testes de confiabilidade entre codificadores são

Quadro 1 - Exemplo de teste de confiabilidade

| | **Codificador 1** | **Codificador 2** | **Codificador 3** |
|---|---|---|---|
| Caso 1 | 1 | 1 | 1 |
| Caso 2 | 2 | 2 | 2 |
| Caso 3 | 1 | 1 | **2** |
| Caso 4 | 3 | 3 | 3 |
| Caso 5 | 2 | 2 | **3** |
| Caso 6 | **3** | 2 | 2 |
| Caso 7 | 2 | 2 | 2 |
| Caso 8 | 3 | 3 | 3 |
| Caso 9 | 1 | 1 | 1 |
| Caso 10 | 2 | 2 | 2 |

Fonte: Os autores, a partir de Freelon (2010).

[26] No caso do R, como é um *software* aberto, há um enorme leque de opções de pacotes a serem instalados para isso. Por exemplo: http://www.cookbook-r.com/Statistical_analysis/Inter-rater_reliability/. Acesso em 15 ago. 2017.
[27] Disponível em: https://provalisresearch.com/downloads/trial-versions/. Acesso 15 ago. 2017.
[28] Disponível em: http://academic.csuohio.edu/kneuendorf/c63311/PRAMS/. Acesso 15 ago. 2017.
[29] No site http://justusrandolph.net/kappa/, o pesquisador preenche diretamente *online* as tabelas, evidenciando onde houve concordância entre os codificadores. Acesso 15 ago. 2017.
[30] Ver https://nlp-ml.io/jg/software/ira/. Acesso em 13 jan. 2017.
[31] Ver http://dfreelon.org/utils/recalfront/. Acesso 15 ago. 2017.
[32] Para mais explicações sobre o site, ver Freelon (2010).
[33] Caso a pesquisa apresente um único valor, infere-se que ele se refere ao menor valor obtido no teste de confiabilidade.
[34] Lima (2003, p.22) sugere, nesse caso, três alternativas para aperfeiçoar o livro de códigos e, assim, permitir um maior nível de concordância entre os codificadores: (a) proceder à fusão de categorias; (b) alterar os descritivos das categorias para torná-las mais claras; (c) inserir exemplos típicos mais adequados para ilustrar o conteúdo indicativo dessas categorias.

Figura 1 - Resultados da ferramenta ReCal

ReCal 0.1 Alpha for 3+ Coders
results for file “exemplo.csv”

| | |
|---|---|
| File size: | 70 bytes |
| N coders: | 3 |
| N cases | 10 |
| N decisions: | 30 |

Average Pairwise Percent Agreement

| **Average pairwise percent agr.** | Pairwise pct. agr. cols 1 & 3 | Pairwise pct. agr. cols 1 & 2 | Pairwise pct. agr. cols 2 & 3 |
|---|---|---|---|
| 80% | 70% | 90% | 80% |

Fleiss’ Kappa

| **Fleiss’ Kappa** | Observed Agreement | Expected Agreement |
|---|---|---|
| 0.697 | 0.8 | 0.34 |

Average Pairwise Cohen’s Kappa

| **Average pairwise CK** | Pairwise CK cols 1 & 3 | Pairwise CK cols 1 & 2 | Pairwise CK cols 2 & 3 |
|---|---|---|---|
| 80% | 70% | 90% | 80% |

Krippendorff’s Alpha (nominal)

| **Krippendorff’s Alpha** | N decisions | $\sum_c o_{cc}$*** | $\sum_c n_c(n_c - 1)$*** |
|---|---|---|---|
| 0.707 | 30 | 24 | 276 |

***These figures are drawn from Krippendorff (2007, case C.)

Fonte: Freelon (2018).

pensados para variáveis que exigem decisões dos codificadores. Se as variáveis são totalmente objetivas (os codificadores só discordam em caso de erro, por exemplo), elas não necessitam de teste de confiabilidade (*e.g.* preencher a data do material coletado).

Convém ressaltar que os testes de confiabilidade, conforme as etapas acima, podem ocorrer em diferentes momentos da pesquisa. Nas pesquisas especializadas mais criteriosas, repetidos testes de confiabilidade são realizados durante todo o processo de pesquisa, nomeadamente: (1) em testes pilotos; (2) no teste de confiabilidade oficial antes da codificação; (3) no meio da codificação e (4) após a codificação. No caso, o teste no ponto três visa conferir se os codificadores continuam concordando entre si durante a codificação e permitir que ajustes sejam realizados caso isso não esteja acontecendo. O teste, após a

codificação de todo o material, visa garantir a confiabilidade da codificação em si e não mais das variáveis[35].

Então, se pudéssemos resumir o passo a passo sugerido por Neuendorf (2002), sugerimos que um teste de confiabilidade seja composto das seguintes etapas:

1. Sucessivas codificações piloto em uma amostra heterogênea (mínimo de 10 unidades), levando a revisões do livro de códigos até que a confiabilidade se mostre aceitável (não aleatória);

2. Codificação independente final em amostra aleatória representativa (sugerimos 10% da amostra, sendo o mínimo de 50 unidades);

3. Pesquisador verifica resultado do índice de confiabilidade;

4. Se passar, relato dos codificadores de suas experiências. Em caso de falha, retornar ao passo 1 ou excluir a variável da análise do material completo;

5. Disponibilizar, na internet, o livro de códigos, a planilha de dados codificada e o material da pesquisa;

6. Apresentar um relato detalhado do processo no texto científico.

Após o teste de confiabilidade, se bem-sucedido, é importante que o pesquisador deixe claro como foi composta a amostra, como se deu o treinamento, qual o índice escolhido, a razão da sua escolha e o escore obtido para cada variável. Caso o resultado do índice do teste não seja superior a 0,75, explicar os motivos para a manutenção do resultado (*e.g.* categorias complexas, estudo inovador, codificadores inexperientes etc.).

Por último, o pesquisador precisa explicar como as divergências finais (ou seja, aquelas existentes no teste de confiabilidade) foram tratadas. Dito de outra forma, mesmo após o teste de confiabilidade ser bem-sucedido, existirão algumas discordâncias de codificação. Os textos especializados recomendam que essa divergência final seja discutida entre os codificadores e o pesquisador responsável (Lombard *et al.*, 2002; Neuendorf 2002). Afinal, a codificação do material em si será iniciada na sequência. Contudo, frisamos, o livro de códigos não pode ser mais alterado a essa altura[36].

### *III.3. Procedimentos para pesquisas individuais*

Enquanto os procedimentos acima correspondem a uma realidade mais profissionalizada de pesquisa, aquela que conta com ao menos dois codificadores, as indicações desta seção se voltam a organizar procedimentos para pesquisadores individuais e que não têm condições de contar com a colaboração de assistentes de pesquisa.

Esse é o caso de pesquisas produzidas em trabalhos de conclusão de curso (TCC), especializações, mestrado e doutorado[37]. Tais procedimentos respondem à demanda por confiabilidade não mediante a realização de um teste com outros codificadores, mas com o(a) próprio(a) pesquisador(a). Apesar de não recomendar seu uso, Krippendorf (2004, p.214) reconhece o método, classificando-o como "estabilidade" (tradução livre). Em suma, este método tende a indicar baixa presunção de confiabilidade, mas ao menos demonstra a estabilidade da classificação do próprio pesquisador. Visando aumentar a confiabilidade ao máximo, tornamos a exigência maior que aquela retratada por Krippendorf. Tais procedimentos se encontram explicitados e sua formulação se volta a um uso crítico e flexível por parte da comunidade da comunidade científica pertinente:

[35] Entretanto, boa parte da literatura reporta apenas o resultado do teste inicial, aquele realizado antes da codificação em si, o que é considerado por alguns autores como incorreto (*e.g.* Lombard *et al.*, 2002).

[36] Ou seja, não é possível mais alterar a descrição ou escopo das categorias ou variáveis. Por outro lado, pequenos refinamentos, como a melhoria de uma descrição ou o acréscimo de alguns exemplos, não tende a macular a confiabilidade já atingida.

[37] Contudo, no nível da pós-graduação, reforça-se a importância do uso de dois ou mais codificadores, como explicamos anteriormente.

1. Escreva um livro de códigos especificando, a partir de códigos alfanuméricos, além das variáveis e categorias, todas as regras de codificação para cada categoria[38];

2. Construa uma planilha de dados em que, ao lado de cada variável, conste uma outra coluna para se inserir o código da regra utilizada em cada codificação;

3. Quando não for possível identificar uma regra no livro de códigos que seja capaz de subsidiar a codificação, reformule o livro de códigos, incluindo a regra necessária à codificação;

4. Quando mais de uma regra for aplicável, reformule o livro, reduzindo a ambiguidade entre as regras ou criando uma adicional que "desempate" as demais;

5. Repetir o procedimento até que 10% do material da pesquisa (estabelecendo o mínimo de 50 unidades) tenha sido codificado[39];

6. Após finalizada a codificação desse material, realize, após o intervalo de ao menos uma semana, uma nova codificação das mesmas unidades, mas sem consultar a primeira codificação!

7. De posse da nova codificação, realize testes de confiabilidade contrastando a primeira com a segunda;

8. Caso os testes não tenham atingido níveis *ótimos* de confiabilidade, deve-se revisar os casos incongruentes e refinar o livro de códigos;

9. Repetir os procedimentos acima com outras unidades (evitando, portanto, codificar uma mesma unidade mais de duas vezes) até que se alcance níveis ótimos de confiabilidade;

10. Uma vez alcançado níveis ótimos de confiabilidade (acima de 0,9), codificar todas as unidades da pesquisa;

11. Reportar os níveis alcançados de confiabilidade para cada variável;

12. Disponibilizar, na internet, o livro de códigos, a planilha de dados codificada e, quando possível, o material da pesquisa.

No caso das instruções acima, deve-se esclarecer que a exigência em torno de um resultado acima de 0,9 no teste realizado é recomendável, pois se trata de apenas um único codificador. Se o próprio pesquisador não consegue replicar seus próprios resultados, isso significa que as regras e as distinções não estão claras ao mesmo. Portanto, em tal caso seria ainda mais improvável que outros pesquisadores venham a conseguir replicar os resultados.

Dito isso, concluímos essa seção reforçando a necessidade de, em qualquer pesquisa com AC quantitativa, realizar-se testes de confiabilidade. Isso porque eles são ferramentas cruciais para se garantir que uma pesquisa seja replicável. Uma vez que a replicabilidade é condição de possibilidade da confiabilidade, uma pesquisa sem um teste e procedimento similar ao aqui descrito deixa de garantir tal condição.

Ademais, a relevância e necessidade por testes de confiabilidade se tornou tão forte que a prática padrão dos periódicos de alto impacto é publicar apenas trabalhos que apresentam algum teste com alto nível de concordância. Entretanto, compreendemos que tal procedimento é apenas um dos que podem sustentar a presunção de confiabilidade, sendo preciso outras ações ainda mais garantidoras. A seção seguinte esclarece que ações seriam essas, assim como

[38] Tal procedimento também pode ser realizado em pesquisas com mais de um codificador. Trata-se de garantir, além de um nível maior de verificabilidade externa da codificação, um instrumento por meio do qual o pesquisador poderá ter clareza da precisão de suas variáveis e categorias.

[39] O ideal, no entanto, é utilizar material que não faça parte do banco de dados da pesquisa a ser reportada no trabalho final.

apresenta os argumentos que indicam os limites que (apenas) um índice de confiabilidade "alto" apresenta para a satisfação de seu princípio epistemológico correspondente.

## IV. Por uma compreensão mais exigente sobre a confiabilidade

Sob nossa perspectiva, a confiabilidade de uma pesquisa se expressa apenas de maneira preliminar quando se realiza um teste de confiabilidade com todos os responsáveis por realizar a codificação. Isso porque esse teste estará interessado no seguinte problema: ao dividir um *corpus* entre diferentes codificadores, qual é a garantia de que cada um deles possui a mesma interpretação dos códigos? Nesse caso, o teste de confiabilidade entre codificadores irá averiguar o quão a metodologia utilizada é confiável internamente, ou seja, entre os que estão envolvidos na codificação do material selecionado. Entretanto, essa confiabilidade interna pode advir não da precisão, clareza e validade das regras utilizadas para a codificação, mas de outras circunstâncias muito distintas.

Uma delas é a redução da complexidade das variáveis elaboradas para indicar o conceito de interesse da pesquisa. Isso porque, quanto maior for a complexidade conceitual e os pressupostos cognitivos necessários para se realizar uma AC, menores serão as chances de que as regras de codificação sejam aplicadas com a mesma eficiência por todos os codificadores.

Sendo assim, a manutenção de um alto nível de validade e confiabilidade para variáveis relativamente complexas tende a requerer uma formação extremamente especializada por parte do codificador, o que requer um intenso treinamento por um pesquisador mais experiente. Como isso implica custos elevados, tanto de ordem financeira como de tempo para a execução da pesquisa, tais custos podem ser drasticamente diminuídos mediante a redução da complexidade das variáveis. Com isso, estas passam a ser implementadas com mais concordância, mas podem já não apresentar a mesma fidelidade conceitual (validade) de outrora[40].

[40] Neuendorf (2001, pp.147-148), para além das indicações de Lima (2003), também sugere a divisão de variáveis em duas ou mais opções mais concretas (mais manifestas).

Outra circunstância comum que está por trás de altos índices de concordância entre codificadores é a geração de regras e decisões *ad hoc* (fora do livro de códigos inicial) após os primeiros testes de confiabilidade. Um exemplo, nesse sentido, é a pesquisa de Wessler e Rinke (2014) na qual, após uma dupla codificação (uma análise de todo o material e todas as variáveis por dois codificadores em separado), os casos discordantes foram resolvidos mediante um debate entre os codificadores. Isso é um procedimento extremamente raro e relativamente mais custoso do que a codificação distribuída. Entretanto, por mais que esse método represente, sem sombra de dúvidas, uma das operacionalizações mais exigentes do princípio de confiabilidade que existe na pesquisa atual com AC, o fato é que os debates travados entre os codificadores não são transparentes (como é o livro de códigos) nem replicáveis e, portanto, há grande chance de que as decisões tenham sido produzidas a partir de regras *ad hoc* geradas pelos codificadores durante o próprio debate.

Desse modo, mesmo que apresente um teste de confiabilidade satisfatório, a AC não se mostrará plenamente confiável caso ela não consiga responder à seguinte questão: se outra pessoa não envolvida na pesquisa codificasse o mesmo material usando o mesmo livro de códigos, ela *poderia* chegar a resultados suficientemente similares aos encontrados pela pesquisa original?

Para isso, o pressuposto é que qualquer indivíduo possa ter acesso ao livro de códigos e ao material utilizado pela pesquisa e, assim, consiga replicar a pesquisa fora de seu contexto institucional de origem. Argumentamos, então, que, para além da confiabilidade alcançada mediante testes de concordância entre participantes da pesquisa, os resultados da mesma só poderiam ser consi-

derados confiáveis quando forem replicadas por outros pesquisadores externos e tal replicação venha a apresentar resultados suficientemente similares aos da pesquisa replicada. Tais condições descreveriam uma abordagem ainda mais abrangente e exigente do princípio de confiabilidade da AC atualmente disponível na literatura especializada.

Ao definir esse conjunto de exigências como os elementos constituintes das condições ideais para a confiabilidade de uma AC, a premissa é que, apenas excepcionalmente, espera-se que uma replicação externa vá efetivamente gerar resultados iguais à pesquisa original. Tendo em vista que qualquer codificação humana é sujeita a discordâncias - não apenas referente à aplicação de uma regra de codificação num caso concreto, mas da própria validade da regra estipulada – a confiança de uma AC deve ser presumida mais em função da plena replicabilidade da pesquisa pela comunidade científica (como um todo) do que em função de uma concordância restrita a codificadores da pesquisa sem possibilidade de replicação externa.

Isso porque, sem essa possibilidade, acaba exigindo-se da comunidade o máximo de confiabilidade nos pesquisadores e não nos procedimentos utilizados para se chegar aos resultados. Além disso, em tal circunstância (sem replicabilidade), não é possível formular críticas, identificar casos questionáveis de perda extrema de validade em prol da confiabilidade e propor aperfeiçoamentos metodológicos para uma melhor acomodação entre suas dimensões epistemológicas. Portanto, a chance de avanço no conhecimento fica extremamente reduzida.

Isso significa que o conjunto de condições ideais aqui proposto funciona, na prática, como um parâmetro a partir do qual se pode avaliar o nível de confiabilidade científica de uma pesquisa. Esta se assentaria na disponibilidade de evidências, conceitos e justificativas passíveis de verificação, contestação e revisão por essa comunidade, e não na autoridade da(s) fonte(s) da pesquisa original.

A partir da definição operacional aqui proposta sobre a confiabilidade de uma AC, percebemos que praticamente todos os estudos que trabalham com essa técnica, tanto na literatura nacional quanto na internacional, não a atendem de maneira plena. Diante disso, em vez de decretar a falta de confiabilidade do imenso e valioso universo de pesquisas com AC que temos disponível, indicamos que estas pesquisas podem reivindicar a *presunção de confiabilidade*, quando (a) oferecem plenas condições de replicabilidade externa (mediante a acessibilidade do livro de códigos e do corpus da pesquisa) e (b) quando oferecem um teste de confiabilidade, realizado por um ou mais codificadores, que seja aceitável.

## V. Conclusões

Este artigo buscou fazer uma discussão epistemológica sobre os três componentes principais da análise de conteúdo: validade, replicabilidade e confiabilidade. Ao fazer isso, destacamos que este último princípio tem sido praticamente ignorado na literatura nacional e internacional, ao mesmo tempo que prevalece uma compreensão bastante limitada e pouco exigente sobre a mesma.

Com o objetivo de enfrentar esses problemas, propomos uma definição crítica e operacional mais exigente de confiabilidade na AC, e constatamos que, a partir de tal definição, não seria possível sequer encontrar registro de pesquisas que pudessem ser apontadas como comprovadamente confiáveis. Isso porque, ao verificar as práticas mais comuns nos periódicos de maior impacto e prestígio, a exigência por resultados aceitáveis de um teste de confiabilidade é tomada, frequentemente, como suficiente para que este princípio seja satisfeito,

não sendo, muitas vezes, disponibilizado o material analisado ou mesmo o livro de códigos.

Compreendemos que essa forma de proceder revela uma forma bastante limitada de satisfazer a exigência de confiabilidade. Tendo em vista que um teste de confiabilidade revela, por um lado, a capacidade da pesquisa ser replicada, com êxito, por aqueles que participaram da codificação, não há, por outro, qualquer garantia, nas aludidas circunstâncias (sem livro de códigos e material codificado amplamente acessíveis), de que os resultados possam ser replicados (e eventualmente revistos) pelo restante da comunidade científica. Isso implica que a satisfação do princípio de confiabilidade requer que todas essas fases sejam satisfeitas, algo que, para nosso melhor conhecimento, não é realidade da maior parte das pesquisas que utilizam a AC.

Todavia, no lugar de decretar a falta de confiabilidade do imenso e valioso universo de pesquisas com AC que temos disponível, indicamos que estas pesquisas podem reivindicar a *presunção de confiabilidade* quando (1) oferecem plenas condições de replicabilidade e (2) quando oferecem um teste de confiabilidade realizado por um ou mais codificadores que seja aceitável como não-aleatório. Para viabilizar a popularização desse procedimento no contexto da pesquisa empírica em língua portuguesa, procuramos, ao longo deste artigo, realizar uma descrição operacional (passo a passo) sobre (1) e (2), mas com maior foco no segundo ponto, devido à sua maior complexidade.

Rafael Cardoso Sampaio (cardososampaio@gmail.com) é Doutor em Comunicação e Cultura Contemporâneas pela UFBA e professor do Programa de Pós-Graduação em Ciência Política na Universidade Federal do Paraná. Vínculo Institucional. Programa de Pós-Graduação em Ciência Política, UFPR, Curitiba, PR, Brasil.

Diógenes Lycarião Barreto de Sousa (dramarc@gmail.com) é Doutor em Comunicação Social pela Universidade Federal de Minas Gerais e professor do Programa de Pós-Graduação em Comunicação Social na Universidade Federal do Ceará. Vínculo Institucional: Programa de Pós-Graduação em Comunicação Social, UFC, Fortaleza, CE, Brasil.

## Referências

Alves, D.; Figueiredo Filho, D. & Henrique, A., 2015. O poderoso NVivo: uma introdução a partir da análise de conteúdo. *Revista Política Hoje*, 24(2), pp.119-134.

Alves, M., 2011. Análise de conteúdo: sua aplicação nas publicações de contabilidade. *Revista Universo Contábil*, 7(3), pp.146-166.

Bardin, L., 2016. *Análise de conteúdo*. São Paulo: Edições 70 Brasil.

Bauer, M., 2007. Analise de conteúdo clássica: uma revisão. In M.W. Bauer & G. Gaskell, eds. *Pesquisa qualitativa com texto, imagem e som: um manual prático*. Petrópolis: Vozes.

Bellucci Júnior, J. & Matsuda, L., 2012. Construção e validação de instrumento para avaliação do Acolhimento com Classificação de Risco. *Revista BrasileiradeEnfermagem,* 65(5), pp.751-757. DOI: 10.1590/s0034-71672012000500006

Campos, C., 2004. Método de análise de conteúdo: ferramenta para a análise de dados qualitativos no campo da saúde. *Revista Brasileira de Enfermagem*, 57(5), pp.611-614. DOI: 10.1590/s0034-71672004000500019

Carlomagno, M. & Rocha, L., 2016. Como criar e classificar categorias para fazer análise de conteúdo: uma questão metodológica. *Revista Eletrônica de Ciência Política*, 7(1), pp.173-188. DOI: 10.5380/recp.v7i1.45771

Cavalcante, R.; Calixto, P. & Pinheiro, M., 2014. Análise de Conteúdo: considerações gerais, relações com a pergunta de pesquisa, possibilidades e limitações do método. *Informação & Sociedade,* 24(1), pp.13-18.

Constantino, N., 2002. Pesquisa histórica e análise de conteúdo: pertinência e possibilidades. *Estudos Ibero-Americanos*, 28(1), pp.183-194.

Feng, G., 2014. Intercoder Reliability Indices: Disuse, Misuse, And abuse. *Quality & Quantity*, 48(3), pp.1803-1815. DOI: 10.1007/s11135-013-9956-8

Feres Jr., J., 2016. Análise de valências, debate acadêmico e contenda política. *Revista Brasileira de Ciência Política*, 20, pp.313-322. DOI: 10.1590/0103-335220162009

Figueiredo, M.; Aldé, A.; Dias, H. & Jorge, V., 1997. Estratégias de persuasão eleitoral: uma proposta metodológica para o estudo da propaganda eleitoral. *Opinião Pública,* 4(3), pp.109-120.

Franco, M., 2008. *Análise de conteúdo*. Brasília: Liber Livro Editora.

Freelon, D., 2010. ReCal: Intercoder Reliability Calculation as a Web Service. *International Journal of Internet Science*, 5(1), pp.20-33.

_____, 2018. *ReCal: Reliability Calculation for the Masses*. Disponível em: http://dfreelon.org/utils/recalfront/. Acesso em: 5 jun. 2018.

Freitas, H., 2011. Réplica 1 – Análise de Conteúdo: faça perguntas às respostas obtidas com sua 'pergunta'! *Revista de Administração Contemporânea*, 15(4), pp.748-760. DOI: 10.1590/s1415-65552011000400011

Gondim, S. & Bendassolli, P., 2014. Uma crítica da utilização da análise de conteúdo qualitativa em psicologia. *Psicologia em Estudo*, 19(2), pp.191-199. DOI: 10.1590/1413-737220530002

Hayes, A. & Krippendorff, K., 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, 1(1), pp.77-89. DOI: 10.1080/19312450709336664

Herscovitz, H., 2007. Análise de conteúdo em jornalismo. In C. Lago & M. Benetti, eds. *Metodologia de pesquisa em jornalismo*. Petrópolis: Vozes.

Jorge, T., 2015, ed. *Notícia em fragmentos: Análise de conteúdo no jornalismo*. Florianópolis: Editora Insular.

Kaplan, A. & Goldsen, J., 1982 A confiabilidade das categorias de análise de conteúdo. In H. Lasswell & A. Kaplan, eds. *A linguagem da política*. Brasília: Editora da UnB.

Krippendorff, K., 2004. *Content Analysis: An Introduction to its Methodology*. London: Sage Publications.

Lara, R., 2011. *A produção de conhecimento no Serviço Social: o mundo do trabalho em debate*. São Paulo: Unesp.

Lasswell, H., 1927. The Theory of Political Propaganda. *The American Political Science Review*, 21(3), pp.627-631. DOI: 10.2307/1945515

Lima, J., 2013. Por uma análise de conteúdo mais fiável. *Revista Portuguesa de Pedagogia*, 47(1), pp.7-29.

Lima, J. & Manini, M.P., 2017. Metodologia para análise de conteúdo qualitativa integrada à técnica de mapas mentais com o uso dos softwares Nvivo e Freemind. *Informação & Informação*, 21(3), pp.63-100. DOI: 10.5433/1981-8920.2016v21n3p63

Lima, L. & Moraes, J.B.E., 2017. A legitimação dos elementos teórico-metodológicos da análise do discurso na ciência da informação brasileira: um aporte da análise de conteúdo. *Brazilian Journal of Information Science*, 11(2), pp.88-95.

Lima, M., 2008. Análise do discurso e/ou análise de conteúdo. *Psicologia em Revista*, 9(13), pp.76-88.

Lombard, M.; Snyder-Duch, J. & Bracken, C., 2002. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4), pp.587-604. DOI: 10.1093/hcr/28.4.587

Lycarião, D., 2014. *Mudanças climáticas sob o prisma da esfera pública: a mediação jornalística como fator de legitimação democrática no caso da COP-15*. Tese de Doutorado. Belo Horizonte: Universidade Federal de Minas Gerais.

Lück, J.; Wessler, H.; Wozniak, A. & Lycariao, D., 2016. Counterbalancing Global Media Frames with Nationally Colored Narratives: A Comparative Study of News Narratives and News Framing in the Climate Change Coverage of Five Countries. *Journalism*, Nov. DOI: 10.1177/1464884916680372

Macnamara, J., 2005. Media Content Analysis: Its Uses, Benefits and Best Practice Methodology. *Asia-Pacific Public Relations Journal*, 6(1), pp.1-34.

Matthes, J. & Kohring, M., 2008. The Content Analysis of Media Frames: Toward Improving Reliability and Validity. *Journal of Communication*, 58(2), pp.258-279. DOI: 10.1111/j.1460-2466.2008.00384.x

Moraes, R., 1999. Análise de conteúdo. *Revista Educação*, 22(37), pp.7-32.

Moro, M., 1989. O problema da validade na pesquisa sobre a alfabetização. *Educar em Revista*, 8, pp.157-181.

Mozzato, A. & Grzybovski, D., 2011. Tréplica-Análise de Conteúdo: ampliando e aprofundando a reflexão sobre a técnica de análise de dados qualitativos no campo da Administração. *Revista de Administração Contemporânea,* 15(4), pp.766-775. DOI: 10.1590/s1415-65552011000400013

Neuendorf, K., 2002. *The Content Analysis Guidebook.* London: Sage Publications.

Oliveira, D., 2008. Análise de conteúdo temático-categorial: uma proposta de sistematização. *Revista de Enfermagem da UERJ*, 16(4), pp.569-576.

Oliveira, E.; Ens, R.; Andrade, D. & Mussis, C., 2003. Análise de conteúdo e pesquisa na área da Educação. *Revista Diálogo Educacional*, 4(9), pp.1-17. DOI: 10.7213/rde.v4i9.6479

Panke, L. & Cervi, E., 2012. Análise da comunicação eleitoral–uma proposta metodológica para os estudos do HGPE. *Contemporânea*, 9(3), pp.390-404.

Paula, C., 2015. Análise dialógica de conteúdo e diálogos de saberes. *Boletim Gaúcho de Geografia*, 42(1), pp.44-63.

Pessoni, A. & Martinez, M., 2015. O uso da análise de conteúdo na Intercom: pesquisas feitas com o método (1996 a 2012). In T. Jorge, ed. *Notícia em fragmentos: análise de conteúdo no jornalismo*. Florianópolis: Editora Insular.

Pohlmann, M.; Bär, S. & Valarini, E., 2014. The Analysis of Collective Mindsets: Introducing a New Method of Institutional Analysis in Comparative Research. *Revista de Sociologia e Política*, 22(52), pp.7-25. DOI: 10.1590/1678-987314225202

Quadros, M.; Assmann, G. & Lopez, D., 2014. A análise de conteúdo nas pesquisas brasileiras em comunicação: aplicações e derivações do método. In E. Barichello & A. Rublescki, eds. *Pesquisa em comunicação: olhares e abordagens*. Santa Maria: Facos-UFSM.

Riffe, D.; Lacy, S. & Fico, F., 2014. *Analyzing Media Messages: Using Quantitative Content Analysis in Research*. London: Routledge.

Rocha, D. & Deusdará, B., 2006. Análise de conteúdo e análise do discurso: o linguístico e seu entorno. *Delta*, 22(1), pp.29-52. DOI: 10.1590/s0102-44502006000100002

Taquette, S.R. & Minayo, M.C.S., 2015. Características de estudos qualitativos conduzidos por médicos: revisão da literatura. *Ciência & Saúde Coletiva*, 20(8), pp.2423-2430. DOI: 10.1590/1413-81232015208.18912014

Thomaz, G.M.; Biz, A.A.; Bettoni, E.M. & Mendes Filho, L., 2016. Mineração de Conteúdo em Mídias Sociais: análise de conteúdos publicados por usuários sobre atrativos turísticos de Curitiba-PR. *Marketing & Tourism Review*, 1(2), pp.1-22. DOI: 10.29149/mtr.v1i2.3846

Triviños, A., 1987. *Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em Educação*. São Paulo: Atlas.

Vergara, S., 2011. Réplica 2 – Análise de conteúdo como técnica de análise de dados qualitativos no campo da administração: potencial e desafios. *Revista de Administração Contemporânea*, 15(4), pp.761-765. DOI: 10.1590/s1415-65552011000400010

Vimieiro, A. & Maia, R., 2011. Análise indireta de enquadramentos da mídia: uma alternativa metodológica para a identi ficação de frames culturais. *Revista Famecos*, 18(1), p.235-252. DOI: 10.15448/1980-3729.2011.1.8810

Wessler, H. & Rinke, E., 2014. Deliberative Performance of Television News in Three Types of Democracy: Insights from the U.S.; Germany, and Russia. *Journal of Communication*, 64(5), pp.827-851. DOI: 10.1111/jcom.12115

Wozniak, A.; Lück, J. & Wessler, H., 2015. Frames, Stories, and Images: The Advantages of a Multimodal Approach in Comparative Media Content Research on Climate Change. *Environmental Communication*, 9(4), pp.469-490.o DOI: 10.1080/17524032.2014.981559

## "I want to believe!" On the importance, uses and limits of inter-coder reliability tests in Content Analysis

ABSTRACT **Introduction**: Content analysis has been normatively organized by the following principles: validity, replicability, and reliability. This paper points out that empirical studies in Brazil and abroad have been ignoring such principles, especially the latter (reliability). From this vantage point, we offer a theoretical and practical contribution to content analysis. **Methods**: Regarding the practical contribution, this papers details check-lists of procedures for conducting intercoder reliability tests for different research conditions (*i.e.* for one or more coders). **Results**: Concerning the theoretical contribution, it departs from an epistemological critical assessment of the advantages and limitations embedded in the most common uses of this kind of test. On the limit, such uses might put the scientific reliability of the published results at risk. **Discussion**: In order to avoid such risk, we argue that empirical studies might claim the presumption of reliability when (a) they offer plain conditions for their replicability; and (b) when they offer a reliability test that might be regarded as significantly non-random. We conclude our paper by pointing out that, in the high ranked journals, prevails the importance of (b) under (a), which implies that even in the elite of scientific production prevails a low demanding comprehension of reliability.

KEYWORDS: content analysis; inter-coder reliability test; replicability; validation; quantitative research.